
Proceedings of the 10th Graz Brain-Computer Interface Conference 2026

Science or Application?

September 14-17, 2026
Graz University of Technology, Austria



Edited by
Gernot R. Müller-Putz, Kyriaki Kostoglou, Markus E. Oberndorfer, Selina C. Wriessnegger

Impressum

Editors Gernot R. Müller-Putz, Kyriaki Kostoglou, Markus E. Oberndorfer, Selina C. Wriessnegger

Cover picture Graz Tourismus – Markus Spenger

2026 Verlag der Technischen Universität Graz

Technikerstraße 4, 8010 Graz

verlag@tugraz.at

www.tugraz-verlag.at

ISBN 978-3-99161-093-9

DOI 10.3217/978-3-99161-093-9



This work is licensed under the Creative Commons

Attribution 4.0 International (CC BY 4.0) license.

<https://creativecommons.org/licenses/by/4.0/deed.en>

This CC license does not apply to the cover, third party material
(attributed to other sources) and content noted otherwise.

Welcome Note

Science or Application?

This year's conference theme, "Science or Application?", reflects a central question in the brain-computer interface (BCI) field. BCI research increasingly spans fundamental neuroscience, methodological development, and translation into real-world and clinical applications. The 10th Graz Brain-Computer Interface Conference (GBCIC2026) invites the community to explore how scientific discovery and application can complement and strengthen each other.

GBCIC2026 provides a platform for scientific exchange among researchers from the international BCI community. Contributions cover a broad range of topics, from basic research and signal processing to neural interfaces, rehabilitation, adaptive neurotechnologies, and clinical translation. Accepted papers have undergone peer review and will be published openly by Verlag der TU Graz.

As an endorsed event of the BCI Society, GBCIC2026 offers a diverse program including Satellite Events, talks, poster sessions, and keynote lectures. We are very pleased to welcome Marie-Constance Corsi, Michael Tangermann, Stanisa Raspopovic, Kai J. Miller, and Gernot R. Müller-Putz as keynote speakers. Their talks reflect the breadth of the field, from fundamental principles and interpretable BCIs to clinical translation and adaptive neurotechnologies.

The year 2026 also marks 35 years of Graz BCI research, making this edition a special milestone for our community.

We wish all participants an exciting and stimulating Graz BCI Conference 2026!



Gernot R. Müller-Putz
Conference Chair

Editorial Board

Prof. Dr. **Gernot Müller-Putz** is head of the Institute of Neural Engineering and its associated Laboratory of Brain-Computer Interfaces. He received his MSc in electrical and biomedical engineering in 2000, his PhD in electrical engineering in 2004 and his habilitation and “venia docendi” in medical informatics from Graz University of Technology in 2008. Since 2014 he is full professor for semantic data analysis. He has gained extensive experience in the field of biosignal analysis, brain-computer interface research, EEG-based neuroprosthesis control, communication with BCI in patients with disorders of consciousness, hybrid BCI systems, the human somatosensory system, and BCIs in assistive technology over the past 27 years. He has also managed several national projects (State of Styria) and international projects (Wings for Life, EU Projects) and recently coordinated the EU Horizon 2020 project MoreGrasp. Furthermore, he organized and hosted seven international Brain-Computer Interface Conferences over the last 25 years in Graz, the next one (10th GBCIC) in Sept. 2026. In 2014/15 he was Guest editor in chief of a special issue of the Proceedings of the IEEE “The Plurality of Human Brain-Computer Interfacing”. Since August 2019 he is Speciality Editor in Chief of Frontiers in Human Neuroscience: Brain-Computer Interfaces. He is Associate Editor of IEEE Reviews in Biomedical Engineering. He has authored more than 215+ peer reviewed publications and more than 210+ contributions to conferences which were cited more than 31700 times (h-index 83). In 2015, he was awarded with an ERC Consolidator Grant “Feel your Reach” from the European Research Council. In May 2017 he received the Ludwig-Guttman Award from the German Medical Spinal Cord Injury Association (DMGP). From 2018 he part of the Board of Directors of the International Brain-Computer Interface Society and served as Treasurer until June 2025. In May 2019 he received a research award from the State of Styria “Digitalisierung in der Wissenschaft”. He is founding member and scientific Co-Director of the NeurolS Society. Since 2024 he is board member of the Austrian Society for Biomedical Engineering. With October 2023 he was elected to be the Dean of the Faculty of Computer Science & Biomedical Engineering. In autumn 2025 he was elected Fellow of the International Academy of Medical and Biological Engineering.

Selina Christin Wriessneger is Associate professor and deputy head at the Institute of Neural Engineering (BCI-Lab), Graz University of Technology, Austria. From 2001 to 2005 she was PhD student at the Max-Planck-Institute for Human Cognitive and Brain Sciences Munich and received her PhD from the Ludwig-Maximilians University Munich, Germany. During that time, she spent one year in Rome as research assistant at IRCCS (Fondazione Santa Lucia), Laboratory for Human Psychophysiology. From 2005 to 2008 she was university assistant at the University of Graz, section neuropsychology. From 2009 until 2016 she was senior researcher at the Institute of Neural Engineering (BCI-Lab). In 2017 she was visiting professor at SISSA (Scuola Internazionale Superiore di Studi Avanzati), Trieste and in 2018 & 2019 she was guest professor in Padova. She received her Venia docendi in “Cognitive Neuroscience” from Graz University of Technology in March 2020. Her research interests are, neural correlates of covert actions, passive BCIs, VR-based neuroadaptive systems, affective computing, neuroscience of art and mental state detection.

Kyriaki Kostoglou received her diploma degree in Electrical and Computer Engineering from Aristotle University of Thessaloniki (AUTH), Greece and her M.Sc. degree in Computer Engineering from University of Cyprus (UCY), Cyprus. In 2017, she completed her Ph.D. studies and received her Ph.D. degree from the Department of Electrical and Computer Engineering, McGill University, Canada. The topic of her Ph.D. thesis was the identification of multiple-input time-varying systems and binary response systems for biomedical applications. She worked as a postdoc researcher in medical ultrasound imaging at the Institute of Signal Processing, Johannes Kepler University, Linz, Austria. Currently she is a postdoc at the Institute of Neural Engineering, Graz University of Technology, Graz, Austria. Her current research interests include system identification and signal processing for biomedical applications and brain computer interfaces.

Markus Erwin Oberndorfer is univ. assistant at the Institute of Neural Engineering, Graz University of Technology, Austria. Currently he is working towards his PhD degree in Biomedical Engineering.

Organizing Committee I

Conference Chair

Gernot R. Müller-Putz
Institute of Neural Engineering
Graz University of Technology
Austria
(BCI Society Member)

Conference Secretary

Irmgard B. Humenberger
Institute of Neural Engineering
Graz University of Technology
Austria

Industrial Sponsoring

Gernot R. Müller-Putz
Institute of Neural Engineering
Graz University of Technology
Austria
(BCI Society Member)

Papers and Proceedings

Kyriaki Kostoglou
Institute of Neural Engineering
Graz University of Technology
Austria

Irmgard B. Humenberger
Institute of Neural Engineering
Graz University of Technology
Austria

Markus E. Oberndorfer
Institute of Neural Engineering
Graz University of Technology
Austria

Poster Sessions

Patrick Suwandjeff
Institute of Neural Engineering
Graz University of Technology
Austria

Program Book

Kyriaki Kostoglou
Institute of Neural Engineering
Graz University of Technology
Austria

Michael Leitner
Institute of Neural Engineering
Graz University of Technology
Austria

Johanna Egger
Institute of Neural Engineering
Graz University of Technology
Austria

Talk & Poster Award

Selina C. Wriessnegger
Institute of Neural Engineering
Graz University of Technology
Austria

Volunteers

Selina C. Wriessnegger
Institute of Neural Engineering
Graz University of Technology
Austria

Organizing Committee II

Social Media & Graphics

Markus Crell
*Institute of Neural Engineering
Graz University of Technology
Austria*

Kathrin Sterk
*Institute of Neural Engineering
Graz University of Technology
Austria*

Workshops

Kyriaki Kostoglou
*Institute of Neural Engineering
Graz University of Technology
Austria*

Bernhard Golser
*Institute of Neural Engineering
Graz University of Technology
Austria*

Michael Wimmer
*Know-Center GmbH
Austria*

Exhibition & Infrastructure

Irmgard B. Humenberger
*Institute of Neural Engineering
Graz University of Technology
Austria*

Christoph Sabitzer
*Institute of Neural Engineering
Graz University of Technology
Austria*

Cortical Peaks Challenge

Markus E. Oberndorfer
*Institute of Neural Engineering
Graz University of Technology
Austria*

Bernhard Golser
*Institute of Neural Engineering
Graz University of Technology
Austria*

We are very grateful to all reviewers for their help, to make this conference a success!

Aarnoutse Erik Utrecht University

Berger Lisa
Brunner Clemens

University of Graz
University of Graz

Caraccia Valentina	Sapienza University of Rome
Cinel Caterina	University of Essex
Collinger Jennifer	University of Pittsburgh
Congedo Marco	French National Centre for Scientific Research
Corsi Marie-Constance	INRIA
Coyle Damien	University of Ulster
Crell Markus	Graz University of Technology

De Vico Fallani Fabrizio INRIA

Egger Johanna Graz University of Technology

Fidas Christos University of Patras

Golser Bernhard
Grosse-Wentrup Moritz

Graz University of Technology
University of Vienna

Jalilpour Shayan Paris Lodron University of Salzburg

Kanoh Shin'Ichiro	Shibaura Institute of Technology
Kleih Sonja	University of Bamberg
Kober Silvia	University of Graz
Kobler Reinmar	Meta Reality Labs

International Program Committee and Review Board II

Kostoglou Kyriaki
Krusienski Dean

Graz University of Technology
Virginia Commonwealth University

L

Leitner Michael
Lopes-Dias Catarina
Lotte Fabien

Graz University of Technology
University of Graz
INRIA

M

Martinez-Cagigal Victor
Mattia Donatella
Müller-Putz Gernot R.

University of Valladolid
Fondazione Santa Lucia, IRCCS
Graz University of Technology

O

Oberndorfer Markus E.
Ofner Patrick
Ortner Rupert

Graz University of Technology
University of Freiburg
g.tec medical engineering GmbH

P

Perdikis Serafeim
Pereira Joana
Pillette Léa
Prasad Girijesh

University of Essex
University of Freiburg
French National Centre for Scientific Research
University of Ulster

R

Ron-Angevin Ricardo
Roy Raphaëlle

University of Málaga
ISAE-SUPAERO

S

Savic Andrej
Sburlea Andreea
Scherer Reinhold
Schwarz Andreas
Srisrisawang Nitikorn
Stern Kathrin
Suwandjieff Patrick

University of Belgrade
University of Groningen
University of Essex
Austrian Research Promotion Agency
Thammasat University
Graz University of Technology
Graz University of Technology

T

Tangermann Michael
Thielen Jordy
Tonin Luca

Radboud University
Radboud University
University of Padova

International Program Committee and Review Board III

V

Valeriani Davide
Vansteensel Mariska
Vaughan Theresa
Vourvopoulos Athanasios

Technogym
University Medical Center Utrecht
National Center for Adaptive Neurotechnologies
ISR-Lisboa

W

Welter Marc
Wimmer Hannah
Wimmer Michael
Wriessnegger Selina Christin

INRIA
FH Joanneum
Know-Center GmbH
Graz University of Technology

List of Authors I

List of authors in alphabetical order with start pages of their respective contributions.

A

Adwane, Grace	340
Afankous, Mehdi	328
Aksenova, Tetiana.....	23, 169, 559
Al-Mfarej, Dalya	571
Alimardani, Maryam	346, 411
Almeida, Ana.....	142
Alvarez, Ana Maria Cebolla	181
Angrisani, Leopoldo	65
Aricò, Pietro	241
Arpaia, Pasquale.....	65

B

Badets, Franck.....	559
Bahadori-Nejad, Maryam	588
Baili, Diana.....	340
Baker, Jonathan Lee	547
Bartolomeo, Paolo	340
Bayen, Eleonore	340
Bechon, Loic	265, 447
Bekius, Annike	310
Berdyshev, Daniil A.	500
Berezutskaya, Julia.....	119
Bialonski, Stephan	35
Bidgoli, Seyed Javad	181
Bigioni, Alessandra	482, 523
Bobrov, Dmitry A.	488
Bobrov, Pavel D.	488
Boesso, Simone.....	482, 523
Bojorges-Valdez, Erik.....	494
Bonnet, Stéphane	169
Boon, Paul	364
Borghini, Gianluca.....	241
Botrel, Loïc.....	453
Bougrain, Laurent	181, 217, 316, 340

List of Authors II

Bour, Christel	29
Bourdil, Manon.....	265, 399
Bousfiha, Camile.....	340
Branco, Mariana Pedroso	107, 310
Bryan, Peter.....	101
Buchheit, Marina	588
Burkitt, Anthony N.	387

C

Capotorto, Rossella	241
Caracci, Valentina	482, 523
Carlson, Tom.....	101, 130
Carpio-Chicote, Ainara	352
Carrette, Evelien	364
Castillos, Kalou Cabrera	235
Chau, Tom	283
Cheron, Guy	181
Chevallier, Sylvain.....	271
Cho, Hanbin.....	588
Cho, Woosang	5
Cimarosto, Pietro	235, 293
Cimolato, Andrea	506
Cincotti, Febo.....	482, 523
Cisnal, Ana.....	83
Collins, Leslie.....	577
Corsi, Marie-Constance	47, 271, 316
Cortecchia, Tommaso	89, 247
Costecalde, Thomas	23
Couton, Marion	340
Cramer, Steven C.	588
Crell, Markus R.	163, 287
Cruz, Aniana	458
Cueva, Valérie Marissens	181

D

da Cruz, Janir	107
da Silva, Henrique Lefundes	217

List of Authors III

Daly, Ian.....	464
Damm, Laura.....	95
De Benedetto, Egidio.....	65
de Carvalho, Pierre-Baptiste Mathieu.....	316
De Jong, Ivo Pascal.....	77
de Ranitz, Alexander.....	435
de Surrel, Thibault.....	181
Deaconu, Tudor-Andrei.....	352
Dehais, Frederic.....	235, 271, 293
Demosthenous, Andreas.....	95
Di Flumeri, Gianluca.....	241
Di Marino, Lucrezia.....	65
Drapart, Saskia.....	517
Dreyer, Pauline.....	265
Du, Rui.....	441
Dumas, Cassandra.....	47
Dupont, Sophie.....	340
Duraccio, Luigi.....	65
Dussard, Claire.....	47
E	
Egger, Johanna.....	175
Eidel, Matthias.....	152
Engel, Veronika Isabella.....	41
Enkhbat, Amar.....	358
Enriquez-Geppert, Stefanie.....	259
Er, Sena.....	535
Ercole, Simone.....	241
Esposito, Antonio.....	65
Esteves, Daniela.....	29, 113
Etienne, Mathis.....	169
Evetović, Nina.....	229, 253
F	
Fabris, Federica.....	310
Faes, Axel.....	158
Fallani, Fabrizio de Vico.....	148, 298, 340

List of Authors IV

Faro, Alessio Lo	89
Fasching, Bernhard.....	506
Fernandes, Jean-Claude.....	142
Fernandes, Sofia.....	29
Fidas, Christos A.....	376
Figueiredo, Patricia	142
Fleck, Stéphanie	447
Fodor, Milan	124
Forin, Paolo	11
Fraile, Juan C.	83
Freudenburg, Zachary	107, 119, 310
Frey, Steffen.....	470
G	
Gabot, Camille	340
Gabuzov, Grigory G.	488
Galasso, Enza	65
García, Alejandro Antonio Torres.....	334, 423
Gargiulo, Ludovica	65
George, Nathalie.....	47
Germano, Daniele.....	241
Geukes, Simon	310
Gkogkidis, Constantin Alexis	547, 588
Gloria, Joeé Jesus Hernandez	512
Golser, Bernhard.....	199, 582
Gomel, Jules.....	235
Gómez, Enrique J.	352, 370
González-González, Daniela	494
Gracia, Carlos Alberto Reyes.....	334
Grayden, David B.....	387
Gueho, Lenaïg	217
Guger, Christoph.....	5
Gupta, Arnav	577
Gutiérrez, Jesús López	423
H	
Hardy, Emmanuel	559

List of Authors V

Hénaff, Patrick	217
Herron, Jeffrey	588
Hirshfield, Leanne	304
Hodúr, Adrián	229, 253
Hons, Manuel	71
Hornero, Roberto	59, 83, 136, 223, 417
I	
Ivanov, Leonid A.	488
Izac, Margaux	328
J	
Jakubík, Jozef	229
Jaoui, Ilan	559
Jeunet, Camille	328
Jiang, Dai	95
K	
Kalugina, Ekaterina	541
Kanatschnig, Thomas	41
Karakas, Serpil	23
Kasaei, Hamidreza	470
Keirse, Elina	158
Kelly, Merlin	101
Kier, Lexi	101
Kober, Silvia Erika	41, 71
Kohnen, Henriette	411
Kojima, Simon	193, 265
Komendziński, Tomasz	381
Kondur, Anna A.	488
Kontaxakis, George	352, 370
Kostoglou, Kyriaki	136, 199, 211, 287, 376
Kotov, Sergey V.	488
Krol, Laurens	152
Kübler, Andrea	53, 152, 453, 517
L	
La Rocca, Daria	148
Lakshminarayanan, Raghavendran	565

List of Authors VI

Le, Thien N.	506
Lécuyer, Anatole	205, 277
Lefeuvre, Théo	205
Leitner, Michael	187
Leung, Jason	283
Lotte, Fabien	181, 193, 265, 399, 447, 553
Lukyanov, Evgeny S.	488
Lupaşcu, Alberta Maria	259
M	
Macé, Marc	205, 277
Machado, Ana Matos	1
Maggi, Emilio	235
Mainsah, Boyla	577
Marcilliere, Quentin	340
Marcos-Martínez, Diego	136, 223
Martel, Félix	23, 169
Martín-Fernández, Ana	59, 83, 417
Martínez-Cagigal, Víctor	59, 83, 136, 223, 417
Mattia, Donatella	482, 523
Mavromoustakos-Blom, Paris	411
Meistelman, Claude	181
Menegatti, Emanuele	11, 89, 322, 346
Merino, Eva Calvo	158
Meurs, Alfred	364
Mihalovic, Miroljub	89
Mirsaeedi, Mani	158, 364
Mlinarič, Tjaša	364
Moccaldi, Nicola	65
Morales-Magallón, Fernando	494
Moreno-Calderón, Selene	83
Moritz, Chet	588
Mrachacz-Kersting, Natalie	512
Mu, Jing	387
Müller-Putz, Gernot R.	136, 163, 175, 199, 211, 287, 376, 387, 429, 582
Murali, Sainath	565

List of Authors VII

Musso, Mariacristina	435
N	
N'Kaoua, Bernard	328
Narębski, Stanisław	381
Nicol, Rozenn	217
Nikolopoulos, Emmanouil.....	370, 376
O	
Oberndorfer, Markus E.....	199, 582
Offenberg, Elena Charlotte.....	107
Ofner, Patrick	512
Oguz, Osman Cagri	470
Ojemann, Jeffrey G.....	588
Oliveira, Ines.....	142
Oropesa, Ignacio.....	370
Ortner, Rupert	5
P	
Pachoud, Mathieu	328
Palatella, Alessio.....	346
Pascual-Roa, Beatriz	223
Patel, Rishan	53, 101, 130
Pathak, Shreyansh.....	512
Paulmann, Silke	464
Pavaux, Marion	148
Peña, David Manuel Carmona	334
Percolla, Elisa	517
Pereira, Carla.....	142
Pérez-Hernández, Eduardo V.	476
Pérez-Turiel, Javier	83
Pérez-Velasco, Sergio	83
Petit, Jimmy	205, 277
Petrinotti, Serena	482, 523
Pfeiffer, Maria.....	453, 517
Pillette, Léa	205, 277
Pires, Gabriel	458
Plapous, Cyril.....	217

List of Authors VIII

Podmore, Joshua James	130, 464
Poljaka, Viktorija	77
Poli, Riccardo	464
Purpura, Keith P.	547
Q	
Qin, Yuzhen	535
Qureshi, Faiz Uddin	283
R	
Radman, Moein	464
Raheem, Soz	435
Ramsey, Nick	107, 119, 310
Raspopovic, Stanisa	506
Riccio, Angela	482, 523
Rimbert, Sébastien	181, 265, 447
Robert-Gonzalez, Adrià	588
Rojas, Maria Lorena Flórez	259
Ronca, Vincenzo	241
Rosipal, Roman	229, 253
Rošňáková, Zuzana	229, 253
Roy, Raphaëlle	265
Ruíz-Gálvez, C. Rubén	223
Rutkowski, Tomasz M.	381
Ryou, Jae Wook	547
S	
Saad, Joe	559
Santamaría-Vázquez, Eduardo	59, 136, 223, 417
Saraiva, Joana	29
Sarma, Devapratim	588
Sauzéron, Hélène	399
Savalle, Emile	205
Sburlea, Andreea Ioana	77, 470
Scharinger, Josef	5
Scheppink, Hanneke	35
Scherer, Reinhold	130
Schipper, Donja	535

List of Authors IX

Schuettler, Martin	547, 588
Schwarzgruber, Michael	5
Sebastián-Romagosa, Marc	5
Segas, Effie	328
Serles, Wolfgang	506
Settgast, Tomko	53, 152
Shevtsova, Yulia G.	488, 529
Shishkin, Sergei L.	488, 500, 529
Sieghartsleitner, Sebastian	5
Silveira, Sarita	152
Simonetto, Piero	322
Skhiri, Wafa	298
Slyunkova, Elena V.	488
Souriau, Rémi	23
Stephan, Emma	588
Sterk, Kathrin	211
Stober, Sebastian	405
Struber, Lucas	559
Sudeep, Aryan	565
Suwandjieff, Patrick	429
Svirin, Evgeniy P.	500, 529
T	
Tangermann, Michael	35, 124, 358, 393, 435, 535
Téllez, Ahirom Cortés	423
Theys, Tom	158
Thielen, Jordy	17, 35, 124, 271, 358, 393, 535
Thirugnanasambandam, Nivethida	565
Thomas, Alex	130
Thomas, Naomi	571
Tonin, Luca	11, 89, 247, 322, 346
Toppi, Jlenia	482, 523
Torres-García, Alejandro A.	476
Tortora, Stefano	11, 89, 247, 322, 346
Tresols, Juan Torre	235
Trombin, Edoardo	89

List of Authors X

Tung, James	571
U	
Ulloa, Miguel	547
V	
Valente, Madalena	142
Van Hoylandt, Anaïs	158
Van Hulle, Marc	158, 364
Van Roost, Dirk	364
Vansteensel, Mariska	107
Vasconcelos, Marta	113
Vasilyev, Anatoly N.	488, 529
Vázquez, Eduardo Santamaría	83
Velardi, Gian Maria	506
Velut, Sebastien	235, 271
Villaseñor-Pineda, Luis	476
Virgolini, Francesca	322
Volosyak, Ivan	35, 124
Volpato, Pietro	247
von der Linde, Lilian	512
Vourvopoulos, Athanasios	29, 113, 142
W	
Wagner, Johanna	152
Weaver, Kurt E.	588
Wei, Mingyi	441
Welter, Marc	553
Wimmer, Hannah S.	387
Wimmer, Michael	387
Wriessnegger, Selina	187
Wu, Ruoling	119
Wyman, Charlotte Genevier	304
Y	
Yang, Man	393
Yashin, Artem S.	488, 529
Yger, Florian	181
Yuce, Ayse Betul	405

List of Authors XI

Z

Zavanone, Chiara	340
Zhang, Hongxin	441
Zimmermann, Jonas	107

Table of Contents I

1.	Regulatory challenges of certification of implantable BCI medical devices in the EU and how to overcome them	1
	Ana Matos Machado	
	DOI: 10.3217/978-3-99161-093-9-001	
2.	Sensorimotor EEG Activity During Brain-Computer Interface Therapy Predicts Motor Function in Parkinson's Disease	5
	Sebastian Sieghartsleitner, Woosang Cho, Marc Sebastian-Romagosa, Rupert Ortner, Michael Schwarzgruber, Josef Scharinger and Christoph Guger	
	DOI: 10.3217/978-3-99161-093-9-002	
3.	Unsupervised Intent Gating: Improving Calibration and Online Reliability in Motor Imagery BCIs	11
	Paolo Forin, Stefano Tortora, Emanuele Menegatti and Luca Tonin	
	DOI: 10.3217/978-3-99161-093-9-003	
4.	Confidence-weighted cumulative rCCA with post hoc re-analysis: unsupervised adaptive learning for calibration-free c-VEP BCI	17
	Jordy Thielen	
	DOI: 10.3217/978-3-99161-093-9-004	
5.	Realtime Class Balancing for Efficient Decoder Adaptation in Asynchronous Brain-Computer Interface	23
	Rémi Souriau, Félix Martel, Thomas Costecalde, Serpil Karakas and Tetiana Aksenova	
	DOI: 10.3217/978-3-99161-093-9-005	
6.	Combining Virtual Reality Feedback and Transcranial Alternating Current Stimulation to Enhance Restorative Brain-Computer Interfaces	29
	Joana Saraiva, Daniela Esteves, Christel Bour, Sofia Fernandes and Athanasios Vourvopoulos	
	DOI: 10.3217/978-3-99161-093-9-006	
7.	Efficient Decoding of the Code-Modulated Motion Visual Evoked Potential (c-MVEP) for BCIs	35
	Hanneke Scheppink, Stephan Bialonski, Michael Tangermann, Jordy Thielen and Ivan Volosyak	
	DOI: 10.3217/978-3-99161-093-9-007	
8.	Advancing NIRS-Based Insights into Motor Imagery and Execution of Swallowing: Effects of Artifact Correction Methods and Hemodynamic Validation	41
	Silvia Erika Kober, Veronika Isabella Engele and Thomas Kanatschnig	
	DOI: 10.3217/978-3-99161-093-9-008	
9.	Inter-Expert Variability in Common Spatial Pattern Component Selection: Implication for Neurofeedback Applications	47
	Cassandra Dumas, Claire Dussard, Marie-Constance Corsi and Nathalie George	
	DOI: 10.3217/978-3-99161-093-9-009	

Table of Contents II

10. Resting-State EEG Predictors for BCI Performance: A Scoping Review.....	53
Tomko Settgast, Rishan Patel and Andrea Kübler	
DOI: 10.3217/978-3-99161-093-9-010	
11. Comparison of sequence, stimulus and decoding algorithm combinations in c-VEP-based BCIs	59
Ana Martín-Fernández, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez and Roberto Hornero	
DOI: 10.3217/978-3-99161-093-9-011	
12. Wearable brain-computer interfaces and their applications: advancements from the ARHeMlab research group	65
Antonio Esposito, Nicola Moccaldi, Ludovica Gargiulo, Luigi Duraccio, Egidio De Benedetto, Enza Galasso, Lucrezia Di Marino, Leopoldo Angrisani and Pasquale Arpaia	
DOI: 10.3217/978-3-99161-093-9-012	
13. Evaluating Baseline Correction Techniques for EEG Time-Frequency Feature Classification	71
Manuel Hons and Silvia Kober	
DOI: 10.3217/978-3-99161-093-9-013	
14. Riemannian Geometry-Preserving Variational Autoencoder for MI-BCI Data Augmentation	77
Viktorija Poljaka, Ivo Pascal De Jong and Andreea Ioana Sburlea	
DOI: 10.3217/978-3-99161-093-9-014	
15. A Feasibility Study of a Motor Imagery–Based Brain–Computer Interface for Post-Stroke Patients.....	83
Víctor Martínez-Cagigal, Selene Moreno-Calderón, Ana Císnal, Sergio Pérez-Velasco, Ana Martín-Fernández, Eduardo Santamaría Vázquez, Juan C. Fraile, Javier Pérez-Turiel and Roberto Hornero	
DOI: 10.3217/978-3-99161-093-9-015	
16. Noninvasive self-paced BCI for lower limb exoskeleton control: a multi-session investigation of user learning	89
Alessio Lo Faro, Tommaso Cortecchia, Miroljub Mihalovic, Edoardo Trombin, Emanuele Menegatti, Luca Tonin and Stefano Tortora	
DOI: 10.3217/978-3-99161-093-9-016	
17. The Effect of Single Session Implicit Motor Learning on Finger Motor Imagery for EEG BCI	95
Laura Damm, Dai Jiang and Andreas Demosthenous	
DOI: 10.3217/978-3-99161-093-9-017	
18. Diversity in BCI: Why is it important?.....	101
Merlin Kelly, Rishan Patel, Peter Bryan, Lexi Kier and Tom Carlson	
DOI: 10.3217/978-3-99161-093-9-018	
19. Decoding Sleep and Wakefulness from Chronically Implanted Intracortical Signals in a Communication BCI User	107

Table of Contents III

Elena Charlotte Offenberg, Zachary Freudenburg, Mariana Branco, Jonas Zimmermann, Janir da Cruz, Mariska Vansteensel and Nick Ramsey DOI: 10.3217/978-3-99161-093-9-019	
20. Proteus Effect in VR-Based BCI: How Avatar Age Influences Motor Imagery Behavior and EEG Daniela Esteves, Marta Vasconcelos and Athanasios Vourvopoulos DOI: 10.3217/978-3-99161-093-9-020	113
21. Deep learning based generalizable articulatory feature extraction for speech-BCI applications Ruoling Wu, Julia Berezutskaya, Zachary Freudenburg and Nick Ramsey DOI: 10.3217/978-3-99161-093-9-021	119
22. Optimizing Bit-wise Linear Decoding for Imperceptible Code-Modulated Visual Evoked Potentials (I-c-VEP) Milan Fodor, Michael Tangermann, Jordy Thielen and Ivan Volosyak DOI: 10.3217/978-3-99161-093-9-022	124
23. Ethical Considerations for User-Centric iBCI Studies: A Thematic Review Alex Thomas, Rishan Patel, Joshua James Podmore, Reinhold Scherer and Tom Carlson DOI: 10.3217/978-3-99161-093-9-023	130
24. EEG2EMG-Net: A Deep Learning Framework for Continuous EEG-to-EMG Estimation for Real-World Applications Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Kyriaki Kostoglou, Gernot Müller-Putz and Roberto Hornero DOI: 10.3217/978-3-99161-093-9-024	136
25. Neural and Clinical Variability in Immersive VR–BCI Training: A Case Study with Stroke Patients Madalena Valente, Ines Oliveira, Jean-Claude Fernandes, Carla Pereira, Ana Almeida, Patricia Figueiredo and Athanasios Vourvopoulos DOI: 10.3217/978-3-99161-093-9-025	142
26. Learning brain connectivity features for improving BCI performance Marion Pavaux, Fabrizio de Vico Fallani and Daria La Rocca DOI: 10.3217/978-3-99161-093-9-026	148
27. A Systematic Overview of EEG-Detectable Control-Related Mental States and Processes in HCI Matthias Eidel, Tomko Settgast, Sarita Silveira, Laurens Krol, Johanna Wagner and Andrea Kübler DOI: 10.3217/978-3-99161-093-9-027	152
28. Finger abduction trajectory prediction from high-density ECoG Axel Faes, Eva Calvo Merino, Mani Mirsaedi, Anaïs Van Hoylandt, Elina Keirse, Tom Theys and Marc Van Hulle DOI: 10.3217/978-3-99161-093-9-028	158

Table of Contents IV

29. EEG-Based Discrimination of Continuous and Discrete Hand Movements for Cursor Control	163
Markus R. Crell and Gernot R. Müller-Putz	
DOI: 10.3217/978-3-99161-093-9-029	
30. Riemannian approaches for ECoG-based Motor Imagery BCI decoding using spatial covariance matrices : a comparative study	169
Mathis Etienne, Felix Martel, Stéphane Bonnet and Tetiana Aksenova	
DOI: 10.3217/978-3-99161-093-9-030	
31. Why Simple Models Still Matter: Adaptive sLDA for Non-Stationary EEG Decoding	175
Johanna Egger and Gernot Müller-Putz	
DOI: 10.3217/978-3-99161-093-9-031	
32. Wrapped One-Class Riemannian EEG Classifier for BCI-Based Detection of Anesthetic States	181
Valérie Marissens Cueva, Thibault de Surrel, Laurent Bougrain, Seyed Javad Bidgoli, Guy Cheron, Ana Maria Cebolla Alvarez, Claude Meistelman, Fabien Lotte, Florian Yger and Sébastien Rimbart	
DOI: 10.3217/978-3-99161-093-9-032	
33. Filling the gap: sex-specific differences in resting interval dynamics	187
Michael Leitner and Selina Wriessnegger	
DOI: 10.3217/978-3-99161-093-9-033	
34. Is User-Independent MI-BCI performance influenced by Gender Proportion in Training Users?	193
Simon Kojima and Fabien Lotte	
DOI: 10.3217/978-3-99161-093-9-034	
35. Non-invasive measurement of somatosensory evoked spinal cord potentials	199
Bernhard Golser, Markus E. Oberndorfer, Kyriaki Kostoglou and Gernot R. Müller-Putz	
DOI: 10.3217/978-3-99161-093-9-035	
36. Towards thermal imagery-based brain-computer interfaces	205
Théo Lefeuvre, Emile Savalle, Jimmy Petit, Marc Macé, Anatole Lécuyer and Léa Pillette	
DOI: 10.3217/978-3-99161-093-9-036	
37. Movement-Onset-Aligned EEG for Generalizable Arm and Hand Movement Decoding: A Large-Scale Multi-Experiment Dataset	211
Kathrin Sterk, Kyriaki Kostoglou and Gernot R. Müller-Putz	
DOI: 10.3217/978-3-99161-093-9-037	
38. Benchmarking Machine Learning strategies to classify Steady-State Auditory Evoked Potentials	217
Henrique Lefundes da Silva, Lenaïg Gueho, Cyril Plapous, Laurent Bougrain, Patrick Hénaff and Rozenn Nicol	
DOI: 10.3217/978-3-99161-093-9-038	
39. Capacity-normalized Shannon graph complexity as an EEG biomarker for event-level	

Table of Contents V

cognitive load estimation in adaptive BCIs.....	223
Beatriz Pascual-Roa, Eduardo Santamaría-Vázquez, Diego Marcos-Martínez, Víctor Martínez-Cagigal, C. Rubén Ruíz-Gálvez and Roberto Hornero	
DOI: 10.3217/978-3-99161-093-9-039	
40. Interpretability of CNN-Based EEG Decoding During Mirror Box Training: A Pilot Study Using CWT Spectrograms and Explainable AI.....	229
Adrián Hodúr, Roman Rosipal, Nina Evetović, Zuzana Rošťáková and Jozef Jakubík	
DOI: 10.3217/978-3-99161-093-9-040	
41. Fast-Calibrated c-VEP BCI with Comfort-Optimized Stimulation for Robust Control on Screen and Immersive VR Using Semi-Dry EEG.....	235
Frederic Dehais, Emilio Maggi, Jules Gomel, Juan Torre Tresols, Pietro Cimarosto, Sebastien Velut and Kalou Cabrera Castillos	
DOI: 10.3217/978-3-99161-093-9-041	
42. Towards Reliable Out-of-Lab pBCIs: Riemannian Geometry Framework for Mitigating Covariate Shift in Real-World Passive Brain-Computer Interfaces.....	241
Daniele Germano, Vincenzo Ronca, Rossella Capotorto, Simone Ercole, Gianluca Borghini, Gianluca Di Flumeri and Pietro Aricò	
DOI: 10.3217/978-3-99161-093-9-042	
43. Decoding Spoken Syllables from EEG Data via Contrastive Embedding Alignment	247
Tommaso Cortecchia, Pietro Volpato, Stefano Tortora and Luca Tonin	
DOI: 10.3217/978-3-99161-093-9-043	
44. Mirror-Box Pretraining to Evoke Sensorimotor Activation to Facilitate Motor Imagery Detection in Brain-Computer Interface Stroke Rehabilitation	253
Nina Evetovic, Roman Rosipal, Adrián Hodúr and Zuzana Rošťáková	
DOI: 10.3217/978-3-99161-093-9-044	
45. Unprotected Minds: A Clinical Neuropsychology Perspective on the Ethics of Emerging Neurotechnologies	259
Alberta Maria Lupaşcu, Maria Lorena Flórez Rojas and Stefanie Enriquez-Geppert	
DOI: 10.3217/978-3-99161-093-9-045	
46. Circadian alignment modulates mental tasks BCI performance.....	265
Pauline Dreyer, Simon Kojima, Manon Bourdil, Loic Bechon, Sebastien Rimbart, Raphaëlle Roy and Fabien Lotte	
DOI: 10.3217/978-3-99161-093-9-046	
47. Introducing simulation-based inference to c-VEP BCI: A feasibility study	271
Sébastien Velut, Marie-Constance Corsi, Sylvain Chevallier, Frédéric Dehais and Jordy Thielen	
DOI: 10.3217/978-3-99161-093-9-047	
48. BCI-based Intelligent Tutoring Systems: A Literature Review and Meta-Analysis on Neuroadaptive ITS	277
Jimmy Petit, Marc Macé, Anatole Lécuyer and Léa Pillette	

Table of Contents VI

DOI: 10.3217/978-3-99161-093-9-048

49. A controlled evaluation of dry textile EEG electrodes using simulated P300 and SSVEP recordings with a gelatin-based skin phantom 283
Jason Leung, Faiz Uddin Qureshi and Tom Chau
DOI: 10.3217/978-3-99161-093-9-049
50. Movement-Related Interhemispheric Connectivity Dynamics and their Link to Asynchronous EEG-BCI Performance..... 287
Kyriaki Kostoglou, Markus R. Crell and Gernot R. Müller-Putz
DOI: 10.3217/978-3-99161-093-9-050
51. Tracking Vigilance Fluctuations with Aperiodic Semi-Transparent Visual Flickers..... 293
Pietro Cimarosto and Frederic Dehais
DOI: 10.3217/978-3-99161-093-9-051
52. Topological Higher-Order Features for Brain-Computer Interfaces 298
Wafa Skhiri and Fabrizio De Vico Fallani
DOI: 10.3217/978-3-99161-093-9-052
53. ERP-XTTN: Interpretable Cross-Subject Error-Related Potential Classification via Cross-Attention to Data-Driven ERP Prototypes..... 304
Charlotte Geneviev Wyman and Leanne Hirshfield
DOI: 10.3217/978-3-99161-093-9-053
54. TREAT: a toolbox for trial selection and alignment of unlabeled neural data..... 310
Simon Geukes, Annike Bekius, Federica Fabris, Zachary Freudenburg, Nick Ramsey and Mariana Pedroso Branco
DOI: 10.3217/978-3-99161-093-9-054
55. Selection and Validation of Spectral EEG Markers for Real-Time Monitoring of Attention and Cognitive Engagement in Learning Contexts 316
Pierre-Baptiste Mathieu de Carvalho, Marie-Constance Corsi and Laurent Bougrain
DOI: 10.3217/978-3-99161-093-9-055
56. Beyond Visual Feedback: Isolating Vestibular Error-Related Potentials During Wheelchair Operations 322
Francesca Virgolini, Piero Simonetto, Stefano Tortora, Emanuele Menegatti and Luca Tonin
DOI: 10.3217/978-3-99161-093-9-056
57. Expertise-Dependent Effects of Neurofeedback Design During Motor Imagery 328
Mathieu Pachoud, Mehdi Afankous, Effie Segas, Margaux Izac, Bernard N'Kaoua and Camille Jeunet
DOI: 10.3217/978-3-99161-093-9-057
58. Towards an Integrated Model of Onset Detection and Multiclass Recognition. 334
David Manuel Carmona Peña, Carlos Alberto Reyes Gracia and Alejandro Antonio Torres Garcia
DOI: 10.3217/978-3-99161-093-9-058
-

Table of Contents VII

59. Using brain network features in motor imagery-based BCI for stroke rehabilitation.....	340
Camile Bousfiha, Camille Gabot, Diana Baili, Marion Couton, Grace Adwane, Chiara Zavanone, Sophie Dupont, Quentin Marcilliere, Eleonore Bayen, Laurent Bougrain, Paolo Bartolomeo and Fabrizio de Vico Fallani	
DOI: 10.3217/978-3-99161-093-9-059	
60. MindRun: an Accessible Multi-Player BCI Game for Collaborative Motor Imagery Training	346
Alessio Palatella, Stefano Tortora, Emanuele Menegatti, Luca Tonin and Maryam Alimardani	
DOI: 10.3217/978-3-99161-093-9-060	
61. Benchmarking Feature-Level and Decision-Level Fusion Strategies for Hybrid EEG-fNIRS Cognitive State Detection	352
Ainara Carpio-Chicote, Tudor-Andrei Deaconu, Enrique J. Gómez and George Kontaxakis	
DOI: 10.3217/978-3-99161-093-9-061	
62. Scene-based Stimulation in the Context of a Visual BCI	358
Amar Enkhbat, Michael Tangermann and Jordy Thielen	
DOI: 10.3217/978-3-99161-093-9-062	
63. Connectivity Between Scalp EEG and Intracranial Recordings During 40 Hz Visual Stimulation	364
Mani Mirsaedi, Tjaša Mlinarič, Marc Van Hulle, Evelien Carrette, Dirk Van Roost, Paul Boon and Alfred Meurs	
DOI: 10.3217/978-3-99161-093-9-063	
64. Comparison of Lightweight EEG Pipelines for Silent Speech Interfaces	370
Emmanouil Nikolopoulos, George Kontaxakis, Enrique J. Gomez and Ignacio Oropesa	
DOI: 10.3217/978-3-99161-093-9-064	
65. EEG-Based Identification of Individuals With Spinal Cord Injury Using Motor-Related and Visual Evoked Potentials	376
Emmanouil Nikolopoulos, Kyriaki Kostoglou, Christos A. Fidas and Gernot Müller-Putz	
DOI: 10.3217/978-3-99161-093-9-065	
66. Deep Sleep Classification via EEG Signal Criticality: A Passive BCI Approach for Sleep-Improvement Neurofeedback	381
Stanisław Narębski, Tomasz Komendziński and Tomasz M. Rutkowski	
DOI: 10.3217/978-3-99161-093-9-066	
67. Execution and Omission Errors in Human–Robot Interaction: An EEG Study.....	387
Jing Mu, Michael Wimmer, Hannah S. Wimmer, Anthony N. Burkitt, Gernot R. Müller-Putz and David B. Grayden	
DOI: 10.3217/978-3-99161-093-9-067	
68. Investigating self-introspection in an auditory ERP-based BCI: A pilot experiment	393
Man Yang, Jordy Thielen and Michael Tangermann	

Table of Contents VIII

DOI: 10.3217/978-3-99161-093-9-068	
69. EEG markers of cognitive impairment in cerebral small vessel disease and neurofeedback: a narrative review.....	399
Manon Bourdil, Hélène Sauzéon and Fabien Lotte	
DOI: 10.3217/978-3-99161-093-9-069	
70. Benchmarking Positional Encoding Strategies for Transformer-Based EEG Foundation Models	405
Ayse Betul Yuce and Sebastian Stober	
DOI: 10.3217/978-3-99161-093-9-070	
71. Gamification Improves Neurofeedback Performance and Experience among Users with ADHD: Findings from NeuroWizard II	411
Henriette Kohnen, Paris Mavromoustakos-Blom and Maryam Alimardani	
DOI: 10.3217/978-3-99161-093-9-071	
72. Towards plug & play c-VEP-based BCIs: a pilot study on zero-calibration adaptive decoding using deep learning.....	417
Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Ana Martín-Fernández and Roberto Hornero	
DOI: 10.3217/978-3-99161-093-9-072	
73. Adaptive Classification of EEG Signals for Bimanual Motor Imagery using Riemannian Geometry	423
Ahram Cortés Téllez, Jesús López Gutiérrez and Alejandro Torres García	
DOI: 10.3217/978-3-99161-093-9-073	
74. Co-Adaptive MRCP-based BCIs: Modulation of Cortical Potentials by Enhanced Classifier Performance	429
Patrick Suwandjieff and Gernot R. Müller-Putz	
DOI: 10.3217/978-3-99161-093-9-074	
75. Brain-computer interface performance development over 30 hours of patient training with an auditory event-related potential protocol	435
Michael Tangermann, Alexander de Ranitz, Soz Raheem and Mariacristina Musso	
DOI: 10.3217/978-3-99161-093-9-075	
76. UniSwarm-SSVEP: Dynamic Content Mapping for Efficient SSVEP-Based Swarm Robot Control	441
Rui Du, Hongxin Zhang and Mingyi Wei	
DOI: 10.3217/978-3-99161-093-9-076	
77. Influence of Median Nerve Stimulation Intensity on ERD/ERS, BCI Performance and User Experience.....	447
Loïc Bechon, Fabien Lotte, Stéphanie Fleck and Sebastien Rimbart	
DOI: 10.3217/978-3-99161-093-9-077	
78. Mission Impossible? Inhibiting Frontal Midline Theta with Neurofeedback.....	453
Maria Pfeiffer, Loïc Botrel and Andrea Kübler	

Table of Contents IX

DOI: 10.3217/978-3-99161-093-9-078

79. Pre-Experiment Resting Baseline Improves Motor Imagery BCI Decoding in a Robotic Hand Rehabilitation Context..... 458
Aniana Cruz and Gabriel Pires
DOI: 10.3217/978-3-99161-093-9-079
80. Decoding Meaning: Cluster-Based Analysis of Semantic Processing Patterns in EEG..... 464
Joshua Podmore, Moein Radman, Silke Paulmann, Riccardo Poli and Ian Daly
DOI: 10.3217/978-3-99161-093-9-080
81. Investigating Human and Classifier-Level Uncertainties: An EEG Study Using Grasp Imitation Task 470
Osman Cagri Oguz, Hamidreza Kasaei, Steffen Frey and Andrea Ioana Sburlea
DOI: 10.3217/978-3-99161-093-9-081
82. A Comparative Study of EEG Representations and Channel Settings for EEG-to-Text Decoding during Reading Tasks 476
Eduardo V. Pérez-Hernández, Alejandro A. Torres-García and Luis Villaseñor-Pineda
DOI: 10.3217/978-3-99161-093-9-082
83. Evaluating the impact of electrode type and montage on classification performance in Assistive P300-based BCIs..... 482
Valentina Caracci, Angela Riccio, Serena Petrinotti, Alessandra Bigioni, Simone Boesso, Jlenia Toppi, Febo Cincotti and Donatella Mattia
DOI: 10.3217/978-3-99161-093-9-083
84. First Real-Time Use of Quasi-Movements for Exoskeleton Control: Comparison with MI-BCI in Healthy Participants and Pilot Application in Stroke Rehabilitation..... 488
Artem S. Yashin, Leonid A. Ivanov, Anna A. Kondur, Dmitry A. Bobrov, Elena V. Slyunkova, Grigory G. Gabuzov, Evgeny S. Lukyanov, Yulia G. Shevtsova, Anatoly N. Vasilyev, Pavel D. Bobrov, Sergey V. Kotov and Sergey L. Shishkin
DOI: 10.3217/978-3-99161-093-9-084
85. Volitional and Non-Volitional Motor Tasks Encode Distinct EEG Dynamics: Implications for Paradigm-Specific BCI Decoder Design 494
Fernando Morales-Magallón, Erik Bojorges-Valdez and Daniela Gonzáles-Gonzáles
DOI: 10.3217/978-3-99161-093-9-085
86. Goal-Directed Quasi-Movements Based on Natural Hand Actions: EEG Correlates and Relevance for Rehabilitation BCI..... 500
Evgeniy P. Svirin, Daniil A. Berdyshev and Sergei L. Shishkin
DOI: 10.3217/978-3-99161-093-9-086
87. HD-EMG-Based BCI for Control and Assessment in Post-Stroke Hand Paresis: A Proof of Concept..... 506
Thien N. Le, Gian Maria Velardi, Andrea Cimolato, Bernhard Fasching, Wolfgang Serles and Stanisa Raspopovic
DOI: 10.3217/978-3-99161-093-9-087
-

Table of Contents X

88.	Single case study of Spatiotemporal Dynamics of Cortical Preparation and Motor Execution Across Lower-Limb Movements of Varying Complexity	512
	Shreyansh Pathak, Lilian von der Linde, Joeé Jesus Hernandez Gloria, Patrick Ofner and Natalie Mrachacz-Kersting	
	DOI: 10.3217/978-3-99161-093-9-088	
89.	Combining Mindfulness and Implicit Neurofeedback to Upregulate Frontal Alpha Asymmetry	517
	Elisa Percolla, Saskia Drapart, Andrea Kübler and Maria Pfeiffer	
	DOI: 10.3217/978-3-99161-093-9-089	
90.	Beyond Accuracy: User Experience Matters in the Comparison of Two EEG Headsets for BCI.....	523
	Serena Petrinotti, Alessandra Bigioni, Valentina Caracci, Simone Boesso, Jlenia Toppi, Febo Cincotti, Donatella Mattia and Angela Riccio	
	DOI: 10.3217/978-3-99161-093-9-090	
91.	First real-time study of a BCI based on quasi-movements: Comparison to the MI BCI within a new hybrid eye–brain–computer interface paradigm.....	529
	Yulia G. Shevtsova, Artem S. Yashin, Evgeniy P. Svirin, Anatoly N. Vasilyev and Sergei L. Shishkin	
	DOI: 10.3217/978-3-99161-093-9-091	
92.	Contrast-dependent decoding performance in classic and grating c-VEP stimuli	535
	Sena Er, Donja Schipper, Yuzhen Qin, Jordy Thielen and Michael Tangermann	
	DOI: 10.3217/978-3-99161-093-9-092	
93.	Decoding Neuroethics: Quality as the Law for “Mind Reading”	541
	Ekaterina Kalugina	
	DOI: 10.3217/978-3-99161-093-9-093	
94.	Epidural micro-ECoG provides a robust interface to the cortex over 4 years in non-human primates	547
	Constantin Alexis Gkogkidis, Martin Schuettler, Miguel Ulloa, Jae Wook Ryou, Keith P. Purpura and Jonathan Lee Baker	
	DOI: 10.3217/978-3-99161-093-9-094	
95.	Hacking BCI Ethics	553
	Marc Welter and Fabien Lotte	
	DOI: 10.3217/978-3-99161-093-9-095	
96.	Low-Power and Compact Time-Domain Analog Feature Extraction for ECoG Motor Intent Decoding	559
	Ilhan Jaoui, Joe Saad, Emmanuel Hardy, Lucas Struber, Tetiana Aksenova and Franck Badets	
	DOI: 10.3217/978-3-99161-093-9-096	
97.	Beyond Hemispheric Laterality: Riemannian Metrics as Markers of Motor Imagery Performance in BCI Training	565
	Sainath Murali, Aryan Sudeep, Raghavendran Lakshminarayanan and Nivethida	

Table of Contents XI

Thirugnanasambandam

DOI: 10.3217/978-3-99161-093-9-097

98. Physiological-Based Change Point Detection During Sustained Attention: A multimodal EEG and Eye-Tracking Study 571
Naomi Thomas, Dalya Al-Mfarej and James Tung
DOI: 10.3217/978-3-99161-093-9-098
99. Evaluating the Impact of Generic Model Priors on User-Specific Calibration in P300 Brain–Computer Interfaces 577
Arnav Gupta, Boyla Mainsah and Leslie Collins
DOI: 10.3217/978-3-99161-093-9-099
100. Similar morphology, different timing: Cortico-spinal response delays following median and ulnar nerve stimulation 582
Markus E. Oberndorfer, Bernhard Golser and Gernot R. Müller-Putz
DOI: 10.3217/978-3-99161-093-9-100
101. Volitional neural control with the CorTec Brain Interchange by an individual with chronic stroke 588
Devapratim Sarma, Hanbin Cho, Adrià Robert-Gonzalez, Emma Stephan, Martin Schuettler, Marina Buchheit, Constantin Alexis Gkogkidis, Jeffrey Herron, Kurt E. Weaver, Steven C. Cramer, Chet Moritz, Jeffrey G. Ojemann and Maryam Bahadori-Nejad
DOI: 10.3217/978-3-99161-093-9-101

Regulatory challenges of certification of implantable BCI medical devices in the EU and how to overcome them

Ana Matos Machado¹

¹ TÜV SÜD Product Service GmbH, Munich, Germany

E-mail: ana.matos@tuvsud.com

ABSTRACT: Brain Computer Interface (BCI) technology has seen substantial growth over the past decade with increasing private investment into its research and development. With ongoing promising clinical studies, market entry of implantable devices is not far in the future and so regulatory discussions now start to become relevant. All medical devices in the EU fall under the scope of the Medical Device regulation 2017/745 (MDR) and potentially the AI Act 2024/1689. So how will certification work for this technology? What will be the expected challenges considering the current regulatory framework? With focus on the clinical assessment part of certification, this article will try to navigate these questions and provide clarity to the process.

INTRODUCTION

On May 2017 the Medical Device Regulation (MDR)[1] was published by the European Commission, replacing the Medical Device Directives that were previously in place. Originally planned to come into force by 2020, due to the COVID pandemic, this was delayed to May 2021. On December 2025, much earlier than initially expected, a proposal for revision of MDR [4] was published with the purpose to remove hurdles for marketing novel technologies and breakthrough devices, thus allowing more innovation in the European Union, however there are no set dates for its implementation until now.

In the EU, the medical devices' certification process has long been privatized, with the main stakeholders being Manufacturers and Notified Bodies (NB). The latter are companies that perform the conformity assessment procedures and are responsible for issuing the certificate, allowing the manufacturers to place medical devices on EU market. The NB operates in an independent and impartial manner, being forbidden to provide consultancy and being regularly audited by the competent authorities in the country they are based on. During the conformity assessment procedure different experts will review the technical documentation submitted by the manufacturer. Clinical assessment is one of the areas of review with its core being laid down in MDR [1] article 61 and Annex XIV.

The MDR also established the creation of the Medical Device Coordination Group (MDCG), which has

released several guidance papers to help all stakeholders navigate the requirements of the regulation. In December 2025 the MDCG guidance on Breakthrough Devices [3] was published, finally filling in a gap in the EU certification process, by establishing an official pathway and guidance for this kind of designation and conformity assessment.

The MDR [1] provides framework for creation of the Expert panels [currently under EMA] to provide scientific opinions in pre-set circumstances foreseen by the regulation. For the purposes of this paper, only their role with Breakthrough devices will be addressed below.

On July 2024, the EU published the AI Act [2]. It came into force in August that same year with a planned phased implementation until August 2027. Medical devices incorporating software are now required to comply with both MDR and AI Act provisions.

Implantable Brain Computer Interface (BCI) medical devices will be subject to conformity assessment under the MDR and AI Act. Given their high novelty degree and expected positive clinical impact on irreversibly debilitating diseases, they can also fit the breakthrough device criteria set in the EU. Until now, no specific guidance or special certification pathway for BCI medical devices exists in the European Union, so this article will look into the current regulatory framework.

MATERIALS AND METHODS

This paper will describe how the conformity assessment procedure needs to be done for class III implantable devices such as implantable BCI, considering the current requirements of MDR [1], AI Act [2] and MDCG for Breakthrough Devices [3].

Focus will be on the clinical assessment part of the procedure.

Potential regulatory challenges (and solutions) will be discussed, also taking into consideration the most recent proposed MDR changes [4].

RESULTS

BCI as active implantable neurological devices are class III [high risk] under Annex VIII of MDR [1]. This means that they are required to have a pre-market

clinical investigation under article 61, and although this article establishes some exceptions to this rule, for a high-risk device initially entering the European market [i.e. not CE marked before] the only one that could potentially apply would be Equivalence [Art 61(4) and 61(5)]. However, for the first few implantable BCIs applying for certification, this is not a viable option because of the lack of equivalent devices already on the market. Of note, devices only marketed outside of Europe would not qualify for equivalence in the case of a high-risk implantable medical device.

According to MDCG 2025-9 [3], in order to have their device be designated as a breakthrough device in the EU, manufacturers should request scientific advice from the EMA Expert panel established in accordance with MDR article 106 [1]. An opinion on the breakthrough status of a medical device can be requested from the Expert panel at any time of device development and once designated, this will remain valid even if, between the designation and the certification, similar devices enter the market. The proposed MDR revision [4] might introduce a small change to this pathway, that will be discussed below.

Being designated as a breakthrough device allows for special benefits such as targeted scientific input, prioritization by NB and a conformity assessment done under the considerations laid down in MDCG 2025-9, meaning, among other points, accepting a potential lower amount of pre-market clinical evidence [provided there is clear evidence of a positive benefit-risk ratio] and putting more emphasis on post-market follow-up. All this acknowledges the potential high clinical benefit of such devices, and potential treatment improvements in patients otherwise lacking sufficient medical treatment.

In November 2025 CORE-MD consortium (Coordinating Research and Evidence for Medical Devices) published their recommendations for clinical investigations of high-risk medical devices [5]. Despite not being legally binding, these overall align with MDCG 2025-9, as for breakthrough devices, a single arm observational study might suffice for certification.

So, although there may be limitations on available pre-market clinical evidence, such as but not limited to, a small cohort of patients or a short follow-up time considering the lifetime of the device, as long as a) the benefit-risk is positive and b) there is a solid post-market clinical follow-up (PMCF) plan to collect further evidence that will fill in the pre-market gaps, this would be sufficient to enter the European market.

Under Annex XIV of the MDR [1], manufacturers are required to create a Clinical Evaluation Plan (CEP) which, among other points, should establish performance and safety outcome parameters for their device. These outcome parameters are derived from State-of-the-art (SOTA) literature addressing similar devices or alternative treatment methods and need to have acceptance thresholds to allow an objective assessment of the subject device's clinical data against what is expected from SOTA. The premarket clinical

study endpoints should also for the most part align with these outcome parameters. Defining appropriate clinical outcome parameters and their thresholds is one of the challenges with BCI technology, that we'll also look into more detail below.

As previously mentioned, medical devices using artificial intelligence algorithms must comply with AI Act [2]. MDCG 2025-9 [3] also addresses breakthrough medical devices with AI, as it states manufacturers are expected to "ensure that the clinical evidence adequately demonstrates the performance, reliability, and generalizability of the medical device AI within its intended purpose and intended use environment". Furthermore, "this evaluation should take into account the dynamic characteristics of AI". Again, the proposed MDR revision [4] also introduces some changes to the interplay of MDR and AI Act, like shifting the MDR from Section A in the AI Act to Section B – which reduces the actual relevance of the AI Act but strengthens the requirement to follow the MDR.

DISCUSSION

Implantable BCI medical devices have a high degree of novelty with at the same time a high expected clinical benefit, which makes them suitable for Breakthrough Device designation in the EU. This will mean the acceptability of the higher uncertainty that comes with a high novelty degree, provided there is a clear clinical benefit and the risk profile is acceptable. Perhaps more than for other devices, for breakthrough technologies, post-market becomes extremely relevant. A solid PMCF Plan - MDCG 2020-7 [6] - accounting for device-specific clinical data collection from large(r) cohorts and with extended follow-up will be expected [e.g., device registries or activities following up on the majority of implanted patients]. Assessment of this post-market data will also be conducted at regular intervals by the Notified Bodies, the timeframe for which will ultimately depend on the type of clinical data gaps at initial certification.

The proposed MDR revision [4] introduces article 52(a) to specifically address breakthrough and orphan devices. On point 4 this article states "Upon a duly substantiated request by a manufacturer or a notified body, an expert panel referred to in Article 106 shall provide an opinion as to whether the criteria set out in paragraph 2 or 3 of this Article, as applicable, are fulfilled" thus not clearly stating the mandatory requirement of designation through the Expert panel [it would alternatively be a NB's decision]. As it is expected that this proposed MDR changes will undergo text revision before official publication, hopefully this will become clear by then. However, until the revised MDR is published, and this is not foreseen before the next one or two years, the current one and associated MDCG papers remain in place, meaning that current guidance of MDCG 2025-9 [3] shall be followed.

It is also important to notice that the proposed revision

also recognizes the panel's opinion as non-binding by stating in article 52(a) 6 "The notified body shall give due consideration to an opinion or advice provided by the expert panels in accordance with paragraph 4 or 5 and, where it does not follow such opinion or advice, it shall provide duly justified reasons. The notified body may ask the expert panel to clarify the opinion it has provided". The proposed revision also for the most part aligns with MDCG 2025-9 [3] specifically on the prioritization of these assessments by NB [art 52(a)6] and the acceptance of limited clinical data provided acceptable benefit-risk and solid PMCF Plan [art 52(a)7].

Early engagement in structured dialogues with NB [also encouraged by MDCG 2025-9] will be key to not only navigating the pathway for breakthrough designation but also, from an early stage, initiating discussions on available data, as, even without providing direct consultancy, NB can discuss available evidence.

What will be suitable outcome parameters for implantable BCIs is a matter of great discussion within the "BCI community". Because AI is a big component of these devices technical outcomes [e.g., speech rate as in words per minute] are inevitable, however, they need to be combined with more clinical ones as to actually translate the expected clinical benefit, these latter will most likely need to include at least some patient reported outcomes, especially whenever the BCI is addressing symptom improvement in debilitating conditions. Establishing acceptability thresholds for these parameters will be an additional challenge though. As previously mentioned, these should derive from SOTA, but what if there is nothing to compare them with? No similar devices and no possible direct comparison with alternative treatments? In such cases, even though alternative treatments might not be the ideal comparison, they are still the only option to provide approximate threshold values for the clinical parameters, especially for the safety ones, which most likely will address procedure complications. As more BCIs enter the market though, they will become points of comparison to each other.

Finally, regarding AI Act and proposed MDR revision the following is stated in the latter "To prevent those overlaps and to simplify the regulatory framework for artificial intelligence-enabled devices, the application of Regulation (EU) 2024/1689 to those devices should be limited to those provisions referred to in Article 2(2) of that Regulation" Should be noted that this, although somehow simplifying the process, does not contradict the statements in MDCG 2025-9.

CONCLUSION

Navigating what sometimes seems like rapidly changing and even contradictory requirements is not an easy task for any of the stakeholders involved in certification of medical devices.

Additionally, the first implantable BCIs will have the unique challenge of setting appropriate outcome parameters, a challenge not only from a clinical perspective but also from a regulatory one.

To mitigate these challenges and smooth the conformity assessment procedure, early engagement and discussions with NB are essential and should not be underestimated.

These are exciting times for neurological devices. Despite the challenges discussed, one thing that we can all agree is that fostering innovation and allowing patients access to groundbreaking technology is definitely possible with the EU regulatory framework, especially given the most recent published guidance and proposals.

REFERENCES

- [1] The European Parliament and The Council of the European Union. Regulation (Eu) 2017/745 of The European Parliament And of The Council of 5 April 2017 on medical devices. Official Journal of European Union. May 2017
- [2] The European Parliament and The Council of the European Union. Regulation (Eu) 2024/1689 of The European Parliament And of The Council of 13 June 2024 Laying Down Harmonised Rules On Artificial Intelligence. Official Journal of European Union. July 2024
- [3] Medical Device Coordination Group. MDCG 2025-9: Guidance on Breakthrough Devices (BtX) under Regulations 2017/745 & 2017/746. Official website of European Union. December 2025.
- [4] The European Parliament and The Council of the European Union. Proposal for a Regulation of The European Parliament And of The Council amending Regulations (EU) 2017/745 and (EU) 2017/746 as regards simplifying and reducing the burden of the rules on medical devices and in vitro diagnostic medical devices, and amending Regulation (EU) 2022/123 as regards the support of the European Medicines Agency for the expert panels on medical devices and Regulation (EU) 2024/1689 as regards the list of Union harmonisation legislation referred to in its Annex I. 2025/0404. December 2025.
- [5] Fraser A. et al. Recommended methodologies for clinical investigations of high-risk medical devices—Conclusions from the European Union CORE-MD Project. The Lancet Regional Health – Europe. 2025;58: 101460
- [6] Medical Device Coordination Group. MDCG 2020-7: Post-market clinical follow-up (PMCF) Plan Template A guide for manufacturers and notified

bodies. Official website of European Union. April
2020.

Sensorimotor EEG Activity During Brain-Computer Interface Therapy Predicts Motor Function in Parkinson's Disease

Sebastian Sieghartsleitner^{1,2}, Woosang Cho¹, Marc Sebastián-Romagosa¹, Rupert Ortner¹,
Michael Schwarzgruber¹, Josef Scharinger², Christoph Guger¹

¹ g.tec medical engineering GmbH, Schiedlberg, Austria

² Institute of Computational Perception, Johannes Kepler University, Linz Austria

E-mail: sieghartsleitner@gtec.at

ABSTRACT: Parkinson's disease (PD) is associated with motor impairment and altered sensorimotor oscillatory activity. Brain computer interface (BCI) based neurorehabilitation may provide both therapeutic benefits and a means to assess underlying neurophysiology. In this study, we investigated whether people with PD (pwPD) can effectively use a motor imagery (MI) based BCI and whether MI related neurophysiological markers allow prediction of motor function at the individual subject level.

Electroencephalography (EEG) data were analyzed from pwPD performing BCI therapy based on MI, functional electrical stimulation and visual feedback using a 3D avatar. PwPD achieved a mean BCI accuracy of 74%, which was not related to motor symptom severity as quantified by the Movement Disorders Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) Part III. Using cross validated regression models, beta band event related (de)synchronization (ERD/S) during MI of all four limbs reliably predicted MDS-UPDRS Part III on an individual subject level.

These findings demonstrate that pwPD can control an MI based BCI independent of motor impairment severity and extend existing literature beyond group level differences by identifying beta ERD/S as an individualized neurophysiological marker of motor function in PD.

INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disease and is characterized by a progressive loss of dopaminergic neurons in the substantia nigra, resulting in reduced dopamine levels [1]. PD is also one of the major causes of disability worldwide and has shown the fastest increase in prevalence and impairment among neurological diseases in recent years [2,3].

People with PD (pwPD) experience a range of motor symptoms, including rigidity, bradykinesia, postural instability, tremor, as well as speech and swallowing impairments. In addition, non-motor symptoms are common and include sensory and psychiatric disorders, cognitive impairment and dementia, depression, anxiety, autonomic dysfunction, gastrointestinal

disorders, and excessive daytime sleepiness [4].

Brain-computer interfaces (BCIs) have been shown to facilitate motor function improvements after stroke [5–9] and have also shown promise in multiple sclerosis [10]. As BCIs are based on recording brain activity, utilizing electroencephalography (EEG) for example, they provide a unique opportunity to track neurophysiological activity continuously throughout the course of therapy.

This is particularly relevant for PD, as numerous studies have linked motor function to changes in oscillatory activity over the sensorimotor cortex and the subthalamic nucleus (STN). However, most of this work has relied on group-level analyses, such as comparisons between pwPD and healthy controls [11,12], comparisons between on- and off-drug states [13–16], or correlations with motor function [13]. While informative, such analyses provide limited insight into predictive power at the individual patient level [17,18]. More recently, there has been a shift toward cross-validated and predictive approaches, with Kurbatskaya et al.[19] demonstrating good performance in differentiating pwPD from controls in a multi-site study, and Waldthaler et al.[20] predicting disease progression from baseline neurophysiological features.

With respect to PD, action observation as well as motor imagery (MI) have been shown to be beneficial for motor function [21]. In addition, also functional electrical stimulation (FES) has shown encouraging results in improving motor function [22]. These approaches target motor networks that are known to be affected in PD. However, the key neurophysiological correlate of MI, namely event-related desynchronization (ERD), is reduced in pwPD [23], raising the question to what extent MI-related brain activity can be reliably exploited for therapeutic and diagnostic purposes in this population.

Here, we show that pwPD can nevertheless use sensorimotor rhythm based BCIs relying on MI, proprioceptive feedback via FES, and visual feedback via a realistic 3D avatar. Furthermore, we extend the existing evidence beyond group-level analyses by demonstrating that ERD during MI constitutes an individualized neurophysiological marker of motor function in pwPD for the first time.

MATERIALS AND METHODS

Participants and Study Design

The study was approved by the Ethikkommission der Medizinischen Fakultät der JKU (1367/2023) and written informed consent was provided before participation.

Fifteen pwPD, participated in a BCI treatment comprising 24 therapy sessions, across 4 weeks. Throughout the entire BCI treatment, participants' medication (levodopa or dopamine agonists) remained unchanged and all sessions were conducted in the on-status. Participants' functional state was evaluated during pre- and post-assessments, which were performed at the same time of date within each patient (see Fig. 1A). The Pre1 assessment was performed one week before the first therapy session, whereas Pre2 was performed one to three days before the first therapy session. Two post-assessments were conducted, with Post1 performed one to three days after the last therapy session and Post2 performed two weeks after the last therapy session.

BCI System

Participants trained with a BCI system based on MI, FES and a realistic 3D avatar providing visual feedback (recoveriX, g.tec medical engineering GmbH, Austria). EEG was recorded using a wireless EEG system (g.Nautilus PRO, g.tec medical engineering GmbH, Austria), and stimulation was delivered using an FES stimulator (g.Estim FES, g.tec medical engineering GmbH, Austria). The EEG montage comprised 16 electrodes covering the sensorimotor cortex (FC3, FCZ, FC4, C5, C3, C1, CZ, C2, C4, C6, CP3, CP1, CPZ, CP2, CP4, and PZ) and reference and ground electrodes were placed on the right earlobe and AFZ, respectively. Throughout the BCI treatment, participants performed right hand versus left foot MI during odd sessions and right foot versus left hand MI during even sessions, with approximately six sessions per week. Fig. 1B illustrates the MI trial structure. Each trial started with an attention beep, followed by a visual and auditory cue indicating the side for MI. Both visual feedback and proprioceptive feedback via FES were provided 1.5 s after the cue. One BCI session consisted of three runs (i.e., blocks), each containing 40 MI trials per side, with the first run of each session used for calibration. Further details regarding the BCI system, including FES parameters and signal processing, can be found elsewhere [10,24].

EEG and Behavioral Data

As the study is still ongoing, the present analysis focuses on EEG data recorded during eight therapy sessions (i.e., four per MI paradigm). To enable across-participant comparisons, the EEG data were adjusted for participants' more and less affected side based on their MDS-UPDRS III in Pre2 [25]. Specifically, the MI tasks were redefined to reflect more and less affected hand and foot MI, assuming that the right side was the

more affected side. For participants whose left side was more affected, electrode positions were swapped pairwise between hemispheres (e.g., C3 ↔ C4).

With respect to the behavioral data, the Movement Disorders Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) Part III [26,27] was used, as it is the primary outcome measure of the study. Further, it represents the gold standard for assessing motor symptoms in PD and is recommended by the Movement Disorder Society. It evaluates rigidity, bradykinesia, postural stability, tremor, and other motor symptoms. Scores range from 0 to 132 points, with higher scores indicating greater motor impairment. For simplicity and readability, the MDS-UPDRS Part III is referred to as MDS-UPDRS III going forward and reflects the score obtained during the Pre2 assessment, as this assessment is closest to the early therapy sessions.

All processing steps were performed in MATLAB (The MathWorks Inc., United States), and all filters were implemented as zero-phase Butterworth digital filters.

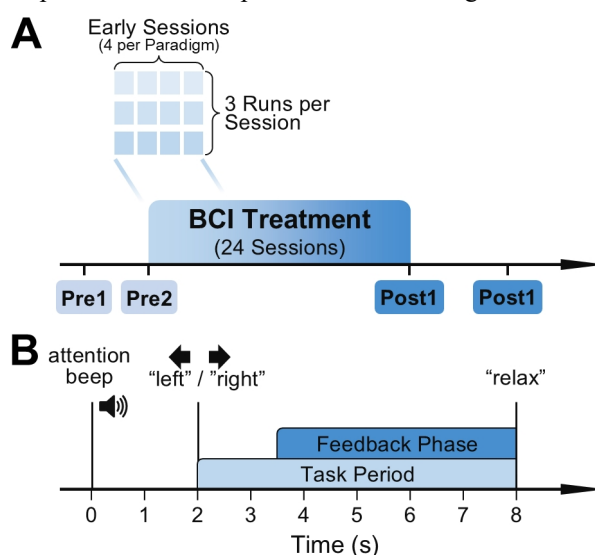


Figure 1: (A) Time course of BCI treatment with the first 4 sessions per paradigm being used for analysis. (B) Time course of a single motor imagery trial.

EEG Preprocessing

ERD/S features from the first eight therapy sessions were extracted for each subject, resulting in a total of 360 BCI runs. Raw EEG data were first notch-filtered at 50 Hz using a second-order notch filter. A trial rejection procedure was then applied to reduce the influence of potential artifacts. Specifically, the data were bandpass-filtered from 1 to 40 Hz using a fourth-order filter and epoched from 0 to 8 s (see Fig. 1B). For each MI trial, log-transformed power values were extracted and z-score normalized across trials. Trials were rejected if any channel exhibited an absolute z-score ≥ 3 .

Feature Extraction and Epoching

After notch-filtering and rejection of bad trials, the EEG data were common average referenced. Log-transformed power was extracted for the mu (8–12 Hz) and beta (13–30 Hz) frequency bands using fourth-order bandpass filters as they are related to motor function

and are modulated during MI, motor execution, action observation, and passive movement [28,29]. Power was estimated using 1 s windows with a 50 ms step size. After epoching from 0 to 8 s, the baseline period from 0 to 2 s was subtracted, resulting in ERD/S features reflecting average power changes in decibels (dB) for the time windows. Finally, ERD/S features were averaged across trials to obtain a robust estimate of subjects' ERD/S.

Predictive Analyses

To reduce dimensionality, multicollinearity, and allow for interpretable results, a subset of ERD/S features was selected. For spatial subsampling, nine regions of interest (ROI) were defined. The premotor cortex was represented by FC3, FCZ, and FC4, the primary motor cortex by C3, CZ, and C4, and the primary somatosensory cortex by CP3, CPZ, and CP4 [30–32]. Temporal subsampling was performed by selecting the following 1 s windows: 2–3 s, 3–4 s, 4–5 s, 5–6 s, 6–7 s, and 7–8 s. The final feature set comprised 432 features, corresponding to 9 ROIs $\times$ 6 time windows $\times$ 2 frequency bands $\times$ 4 MI tasks.

A cross-validated regression analysis was used to assess whether ERD/S patterns could predict subjects' motor function as quantified by MDS-UPDRS III. The model consisted of a support vector machine (SVM) regressor with a Gaussian kernel. Input features were standardized, the kernel scale was automatically determined using MATLAB's heuristic optimization, and the box constraint was fixed at 1.0 to reduce overfitting. Model performance was evaluated using 10 repetitions of a 5-fold cross-validation procedure due to the relatively small sample size. Pearson's r was used as the performance metric.

To examine the predictive contribution of individual MI tasks and frequency bands (mu, beta, and mu+beta), the following feature subsets were constructed: (1) all MI tasks, (2) more affected hand MI, (3) more affected foot MI, (4) less affected hand MI (5) less affected foot MI. Empirical P -values were computed for each of the 15 feature subsets (5 MI conditions $\times$ 3 frequency band conditions) to estimate the probability that observed model performances occurred by chance, according to Lee et al.[33]. A permutation approach was applied, generating 1,000 null models by randomly shuffling the target variable. For each permutation, the full 10 $\times$ 5 cross-validation procedure was repeated, resulting in 1,000 performance estimates under the null hypothesis of no association between features and the target variable. The resulting 15 P -values were corrected for multiple comparisons using the false discovery rate procedure [34]. Above-chance performance was defined as $P < 0.05$.

RESULTS

In the current sample, pwPD were able to control the BCI system with MI accuracies of 74.7% (SD = 7.4) for more affected hand versus less affected foot MI and

74.6% (SD = 10.2) for more affected foot versus less affected hand MI. These two accuracies were significantly correlated ($r = 0.712$, $P = 0.0029$), indicating that patients achieved similar MI performance irrespective of the MI paradigm. Importantly, MI accuracy during both MI tasks was independent of subjects' MDS-UPDRS III score in Pre2 ($P > 0.05$, respectively). Note that, MDS-UPDRS III scores did not differ significantly between the two pre-assessments ($P = 0.46$).

Fig. 2 shows the obtained model performances for the respective feature subsets and indicates which model performance corresponds to a corrected P -value of 0.05. All models that exclusively utilized beta ERD/S performed above chance level, whereas none of the models that exclusively utilized mu ERD/S performed above chance level. When combining mu and beta ERD/S, model performance slightly deteriorated on average and dropped below chance level when ERD/S during less affected hand MI was used. Overall, ERD/S predicted subjects' motor function similarly well across the different MI tasks.

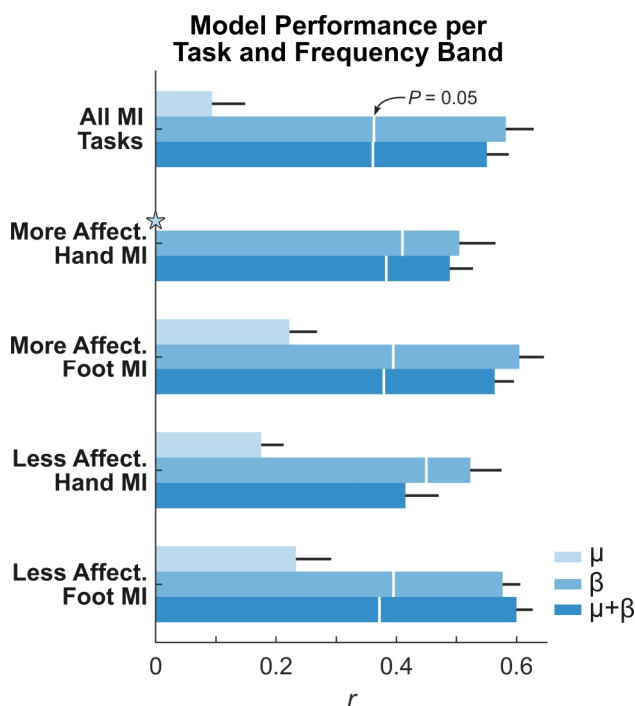


Figure 2: Model performance obtained for predicting MDS-UPDRS III score based on subsets of event-related (de)synchronization (ERD/S) features. Performance is quantified by Pearson's r and obtained in a 10 $\times$ 5 cross-validation. Bars and errorbars reflect the mean and standard error of the mean, respectively. Star indicates the model performance for mu ERD/S features during more affected hand motor imagery (MI): -0.26 (SEM = 0.05).

Investigating the relationship between beta ERD/S and MDS-UPDRS III during the respective MI tasks revealed statistically significant effects for both more and less affected hand MI, as well as strong and statistically significant correlations (see Fig. 3). ERD/S was extracted from the electrode location (C3, CZ, or

C4) expected to exhibit the strongest ERD during the respective MI tasks. For example, C3 is expected to show the strongest ERD during more affected hand MI, whereas CZ is expected to show the strongest ERD during both more and less affected foot MI due to its topographical position. Across all conditions, stronger ERD (i.e., more negative ERD/S values) was associated with better motor function, reflected by lower MDS-UPDRS III scores. Importantly, this relationship cannot be attributed to differences in task engagement, as classification accuracy did not differ between more and less affected patients ($P > 0.2$, respectively).

Finally, three subjects have so far completed the BCI treatment and the Post1 assessment (see Tab. 1). All subjects showed improvements in MDS-UPDRS III scores, with S2 and S3 exceeding the minimally clinically important difference [35]. These findings provide encouraging preliminary evidence for the study going forward.

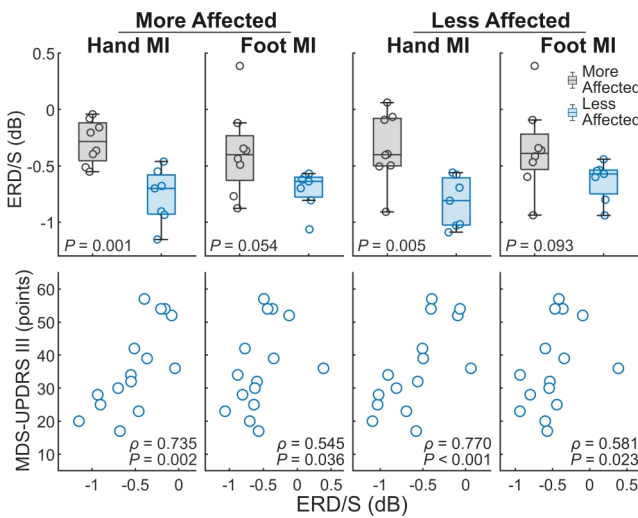


Figure 3: Relationship between beta event related (de)synchronization (ERD/S) and MDS-UPDRS III scores across motor imagery (MI) tasks. ERD/S is shown for the electrode expected to exhibit maximal ERD (e.g., Cz for foot MI and C3 for more affected hand MI). The upper panel shows subjects stratified by MDS-UPDRS III, and the lower panel shows correlations between ERD/S and MDS-UPDRS III.

Table 1: BCI accuracy for early (first 8) and late (last 8) therapy sessions and Pre2- and Post1 MDS-UPDRS III

Subject	Accuracy (%)		UPDRS III (points)	
	Early	Late	Pre2	Post1
S1	80.9	79.0	34	32
S2	73.2	73.0	54	40
S3	77.7	79.9	23	15

DISCUSSION

In this study, we investigated the relationship between ERD/S during BCI therapy and the motor functional state of pwPD. To this end, we utilized EEG data from the first eight BCI sessions and MDS-UPDRS III scores obtained during the Pre2 assessment.

The predictive analyses using cross-validated regression

models showed that beta ERD/S is a strong predictor of motor function in pwPD. Specifically, stronger ERD during MI was robustly associated with better motor function as quantified by MDS-UPDRS III. Similar observation have been reported previously, including movement-related ERD being: modulated by drug status (on versus off medication) [13–16], correlated with motor symptoms [13] and reduced in pwPD compared to healthy controls [11,12]. Building upon these findings, the present results demonstrate that beta ERD during MI even allows inference of motor function at the individual subject level, during on-drug status.

Beta ERD/S was as a markedly stronger predictor of motor function than mu ERD/S. This contrasts with findings in stroke patients, where both frequency bands were shown to contribute to functional outcomes [25]. This difference is consistent with PD motor symptoms being closely linked to increased beta synchronization at rest and with beta activity being strongly modulated by dopaminergic medication and STN-DBS, both of which enhance motor-related ERD and improve motor function [15,16,36,37].

Another notable observation was that the limb used for the MI task did not substantially influence the predictive power of beta ERD. Specifically, MI of both the more and less affected hand and foot provided sufficient information to allow robust prediction of patients' motor function. This suggests that beta ERD reflects a more global marker of motor impairment in pwPD rather than being strictly limb-specific.

The current findings are based on ERD/S patterns extracted during BCI therapy and therefore provide insight into subjects' motor functional state during and across the course of the BCI treatment. An advantage of this approach is that no additional or time-consuming recording sessions are required. At the same time, the findings should be interpreted in the context of MI combined with FES feedback [38]. Future studies in PD should therefore investigate the relationship between ERD/S and motor function across MI, motor execution, passive FES, and may compare its predictive performance to resting state power using a randomized cross-over design.

Finally, three patients have so far completed the full BCI therapy and the Post1 assessment, all of whom showed an improvement in motor function, with two exceeding the minimally clinically important difference. In addition, participants tolerated the BCI therapy well and no adverse events were observed. Taken together, these preliminary results are encouraging with respect to the feasibility and safety of BCI therapy for pwPD.

CONCLUSION

In the present study, we showed that pwPD can control an MI based BCI system independent of motor symptom severity. Furthermore, we demonstrated that beta ERD/S during BCI therapy based on MI, FES and visual feedback enables prediction of motor function in pwPD at the individual subject level. Beta ERD/S

predicted motor function irrespective of the MI task performed. Taken together, these findings extend previous group-level observations by identifying beta ERD/S as an individualized neurophysiological marker of motor impairment that can be assessed during ongoing BCI therapy.

REFERENCES

- [1] Elbaz A, Carcaillon L, Kab S, Moisan F. Epidemiology of Parkinson's disease. *Revue Neurologique* 2016;172:14–26
- [2] Xu T, Dong W, Liu J, Yin P, Wang Z, Zhang L, et al. Disease burden of Parkinson's disease in China and its provinces from 1990 to 2021: findings from the global burden of disease study 2021. *The Lancet Regional Health - Western Pacific* 2024;46:101078
- [3] Li M, Ye X, Huang Z, Ye L, Chen C. Global burden of Parkinson's disease from 1990 to 2021: a population-based study. *BMJ Open* 2025;15:e095610
- [4] Adler CH. Nonmotor complications in Parkinson's disease. *Mov Disord* 2005;20:S23–9
- [5] Cervera MA, Soekadar SR, Ushiba J, Millán JDR, Liu M, Birbaumer N, et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. *Ann Clin Transl Neurol* 2018;5:651–63
- [6] Bai Z, Fong KNK, Zhang JJ, Chan J, Ting KH. Immediate and long-term effects of BCI-based rehabilitation of the upper extremity after stroke: a systematic review and meta-analysis. *J NeuroEngineering Rehabil* 2020;17:57
- [7] Kruse A, Suica Z, Taeymans J, Schuster-Amft C. Effect of brain-computer interface training based on non-invasive electroencephalography using motor imagery on functional recovery after stroke - a systematic review and meta-analysis. *BMC Neurol* 2020;20:385
- [8] Ren C, Li X, Gao Q, Pan M, Wang J, Yang F, et al. The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: a systematic review and meta-analysis. *Front Hum Neurosci* 2024;18:1438095
- [9] Li D, Li R, Song Y, Qin W, Sun G, Liu Y, et al. Effects of brain-computer interface based training on post-stroke upper-limb rehabilitation: a meta-analysis. *J NeuroEngineering Rehabil* 2025;22:44
- [10] Sebastián-Romagosa M, Cho W, Ortner R, Sieghartsleitner S, Guger M, Von Oertzen TJ, et al. Toward Brain-Computer Interface motor rehabilitation for people with Multiple Sclerosis. *Front Med* 2026;12:1661972
- [11] Heinrichs-Graham E, Wilson TW, Santamaria PM, Heithoff SK, Torres-Russotto D, Hutter-Saunders JAL, et al. Neuromagnetic Evidence of Abnormal Movement-Related Beta Desynchronization in Parkinson's Disease. *Cerebral Cortex* 2014;24:2669–78
- [12] McLinden J, Deligani RJ, Abtahi MR, Akbar U, Mankodiya K, Shahriari Y. Disruptions of cortico-kinematic interactions in Parkinson's disease. *Behavioural Brain Research* 2021;404:113153
- [13] Brown P, Marsden CD. Bradykinesia and impairment of EEG desynchronization in Parkinson's disease. *Mov Disord* 1999;14:423–9
- [14] Devos D, Labyt E, Cassim F, Bourriez JL, Reyns N, Touzet G, et al. Subthalamic stimulation influences postmovement cortical somatosensory processing in Parkinson's disease. *Eur J of Neuroscience* 2003;18:1884–8
- [15] Doyle LMF, Kühn AA, Hariz M, Kupsch A, Schneider G, Brown P. Levodopa-induced modulation of subthalamic beta oscillations during self-paced movements in patients with Parkinson's disease. *Eur J of Neuroscience* 2005;21:1403–12
- [16] Oswal A, Litvak V, Brücke C, Huebl J, Schneider G-H, Kühn AA, et al. Cognitive Factors Modulate Activity within the Human Subthalamic Nucleus during Voluntary Movement in Parkinson's Disease. *J Neurosci* 2013;33:15815–26
- [17] Lo A, Chernoff H, Zheng T, Lo S-H. Why significant variables aren't automatically good predictors. *Proc Natl Acad Sci USA* 2015;112:13892–7
- [18] Gordillo D, Ramos Da Cruz J, Moreno D, Garobbio S, Herzog MH. Do we really measure what we think we are measuring? *iScience* 2023;26:106017
- [19] Kurbatskaya A, Jaramillo-Jimenez A, Ochoa-Gomez JF, Brønnick K, Fernandez-Quilez A. Machine Learning-Based Detection of Parkinson's Disease From Resting-State EEG: A Multi-Center Study. 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia: IEEE; 2023, p. 1–4
- [20] Waldthaler J, Comarovschii I, Lundqvist D. Magnetoencephalography-based prediction of longitudinal symptom progression in Parkinson's disease. *Npj Parkinsons Dis* 2026;12:29
- [21] Caligiore D, Mustile M, Spalletta G, Baldassarre G. Action observation and motor imagery for rehabilitation in Parkinson's disease: A systematic review and an integrative hypothesis. *Neuroscience & Biobehavioral Reviews* 2017;72:210–22
- [22] Popa L, Taylor P. Functional electrical stimulation may reduce bradykinesia in Parkinson's disease: A feasibility study. *Journal of Rehabilitation and Assistive Technologies Engineering* 2015;2:2055668315607836
- [23] Tremblay F, Léonard G, Tremblay L. Corticomotor facilitation associated with observation and imagery of hand actions is impaired in Parkinson's disease. *Exp Brain Res*

- 2008;185:249–57
- [24] Sebastián-Romagosa M, Cho W, Ortner R, Murovec N, Von Oertzen T, Kamada K, et al. Brain Computer Interface Treatment for Motor Rehabilitation of Upper Extremity of Stroke Patients—A Feasibility Study. *Front Neurosci* 2020;14:591435
- [25] Sieghartsleitner S, Sebastián-Romagosa M, Schreiner L, Grünwald J, Cho W, Ortner R, et al. Analysis of Cortical Excitability During Brain-Computer Interface Stroke Rehabilitation of Upper and Lower Extremity. 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE), St Albans, United Kingdom: IEEE; 2024, p. 1106–11
- [26] Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. *Movement Disorders* 2007;22:41–7
- [27] Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders* 2008;23:2129–70
- [28] Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology* 1999;110:1842–57
- [29] Cheyne DO. MEG studies of sensorimotor rhythms: A review. *Experimental Neurology* 2013;245:27–39
- [30] Homan RW, Herman J, Purdy P. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and Clinical Neurophysiology* 1987;66:376–82
- [31] Okamoto M, Dan H, Sakamoto K, Takeo K, Shimizu K, Kohno S, et al. Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10–20 system oriented for transcranial functional brain mapping. *NeuroImage* 2004;21:99–111
- [32] Koessler L, Maillard L, Benhadid A, Vignal JP, Felblinger J, Vespignani H, et al. Automated cortical projection of EEG sensors: Anatomical correlation via the international 10–10 system. *NeuroImage* 2009;46:64–72
- [33] Lee HS, Schreiner L, Jo S-H, Sieghartsleitner S, Jordan M, Pretl H, et al. Individual finger movement decoding using a novel ultra-high-density electroencephalography-based brain-computer interface system. *Front Neurosci* 2022;16:1009878
- [34] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995;57:289–300
- [35] Horváth K, Aschermann Z, Ács P, Deli G, Janszky J, Komoly S, et al. Minimal clinically important difference on the Motor Examination part of MDS-UPDRS. *Parkinsonism & Related Disorders* 2015;21:1421–6
- [36] Marsden JF, Limousin-Dowsey P, Ashby P, Pollak P, Brown P. Subthalamic nucleus, sensorimotor cortex and muscle interrelationships in Parkinson's disease. *Brain* 2001;124:378–88
- [37] Kuhn AA, Kempf F, Brucke C, Gaynor Doyle L, Martinez-Torres I, Pogosyan A, et al. High-Frequency Stimulation of the Subthalamic Nucleus Suppresses Oscillatory Activity in Patients with Parkinson's Disease in Parallel with Improvement in Motor Performance. *Journal of Neuroscience* 2008;28:6165–73
- [38] Reynolds C, Osuagwu BA, Vuckovic A. Influence of motor imagination on cortical activation during functional electrical stimulation. *Clinical Neurophysiology* 2015;126:1360–9

Unsupervised Intent Gating: Improving Calibration and Online Reliability in Motor Imagery BCIs

Paolo Forin¹, Stefano Tortora^{1,2}, Emanuele Menegatti^{1,2} and Luca Tonin^{1,2}

¹Department of Information Engineering, University of Padova, Padova, Italy

²Padova Neuroscience Center, University of Padova, Padova, Italy

E-mail: paolo.forin@phd.unipd.it

ABSTRACT: Motor Imagery (MI) Brain-Computer Interfaces (BCIs) are frequently hampered by fluctuations in user attention, which introduce noisy labels during standard calibration. To address this, we propose an unsupervised gating architecture using a Gaussian Mixture Model (GMM) to dynamically estimate Intentional Control (IC). By evaluating sensorimotor rhythm spatial focality, the GMM automatically filters out non-informative idling segments without manual pruning. Offline validation (N=5) demonstrated sharper neurophysiological topographies and higher class separability than standard methods. During real-time closed-loop control, the GMM probability dynamically weighted a continuous soft-increment integrator. While baseline sample-wise accuracy remained comparable (77.7% vs. 77.4%), the gating mechanism effectively converted neural uncertainty into safe timeouts, boosting effective active decision accuracy to 92.7% and maintaining a 76.0% safety margin against false positives during rest. This intent-gated paradigm offers a robust framework for safe, real-world assistive BCI technologies.

INTRODUCTION

While Motor Imagery (MI) remains a cornerstone of non-invasive Brain-Computer Interfacing (BCI), the transition from controlled calibration to reliable online control is frequently hampered by the inherent non-stationarity of Electroencephalographic (EEG) signals [1, 2]. This reliability gap is particularly critical in neurorobotics, where safety and precision are paramount for controlling wheelchairs or telepresence robots [3]. A predominant factor degrading classifier generalisation is high intra-subject variability, often driven by fluctuations in the user's psychophysiological state [4–6]. Standard calibration protocols typically employ a synchronous approach where the system assumes a stable state of "Intentional Control" (IC) throughout the entire duration of the performed task. Consequently, supervised learning pipelines assign a static class label to all time points within this interval [7, 8]. However, neurophysiological and behavioural studies suggest that sustained attention during mental imagery is rarely constant. Factors such as cognitive fatigue, "mind-wandering," or delayed task initiation can introduce significant periods of

"Intentional Non-Control" (INC) or idling states within the theoretically active trial [3, 9, 10]. From a machine learning perspective, this creates a "noisy labelling" problem: the classifier is forced to learn decision boundaries from a dataset where a subset of samples represents background noise. This ambiguity dilutes the separability of the feature space, ultimately reducing decoding accuracy [1, 11, 12]. To mitigate signal contamination, standard pipelines rely on artifact removal algorithms (e.g., Regression, ICA) to suppress EOG and EMG activity [13, 14]. While effective for physiological noise, these methods do not address "cognitive artifacts"—moments where the signal is clean, but the user's intent is absent [15]. Advanced geometric approaches, such as Riemannian Manifolds, have shown robustness to covariance shifts [16, 17], yet they typically remain supervised and susceptible to label noise. Another established strategy is shared control, where the intelligent system compensates for the BCI's uncertainty by handling low-level tasks (e.g., obstacle avoidance) [18]. In this scenario, there is a growing need for unsupervised or semi-supervised methods designed to explore the latent structure of the training data to isolate informative segments automatically [2, 19, 20]. In this study, we propose an unsupervised framework that integrates neurophysiological insights with machine learning to refine both BCI calibration and real-time continuous control. We hypothesise that effective motor imagery is characterised by high spatial focality compared to the resting state, so we apply a Gaussian Mixture Model (GMM) to a spatial focality feature. During the calibration phase, this probability-based gate automatically filters out non-informative idling segments from the training stream without requiring manual labels. By restricting the training of a Quadratic Discriminant Analysis (QDA) classifier to these GMM-verified samples, we achieve a sharper decision boundary, significantly boosting class separability (Fisher Score) and Signed R^2 while ensuring neurophysiological plausibility. Crucially, we extend this mechanism into the online closed-loop architecture, where the continuous GMM probability dynamically weights a soft-increment integrator. By doing so, the intent-gate actively translates user uncertainty into safe timeouts, drastically improving the effective decisional accuracy and providing a robust safety margin against false positives during resting states.

MATERIALS AND METHODS

Data acquisition and experimental paradigm:

EEG data were recorded from 5 healthy subjects (age 26.5 ± 1.5) using a g.USBamp system. Signals were acquired at a sampling rate of 512 Hz from 16 active electrodes positioned according to the international 10-20 system, concentrated over the sensorimotor area (including Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, Fp1, CP1, CPz, CP2, Fp2). Ground was placed at AFz and reference at the right earlobe. The experimental protocol was organised into a single recording session consisting of two distinct phases: calibration and online evaluation. The calibration phase was composed of three runs, designed to gather the necessary data for classifier training. Each subject performed 30 trials per Motor Imagery (MI) class (both feet and both hands), for a total of 60 calibration trials, where participants received always-correct visual feedback. Subsequently, the system's performance was assessed during the online evaluation phase, which included two runs. Each evaluation run consisted of 10 trials per MI class and 5 additional rest trials, totalling 25 trials per run, where real-time feedback reflected the user's control. Trials followed a standard cue-based synchronous paradigm (Fig. 1).

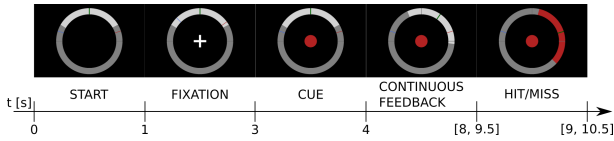


Figure 1: Trial organization: first the fixation (duration 2s), then a visual cue (duration 1s), which can be red/blue/yellow (both hands/both feet/rest), then the continuous feedback (the duration in the calibration is a random number between 4 to 5.5s, while in evaluation the max duration is 10s) and lastly the hit/miss (duration 1s).

Signal preprocessing:

Signal processing was performed using MATLAB (MathWorks, Natick, MA). First, a Common Average Reference (CAR) spatial filter was applied to enhance the signal-to-noise ratio. The signal was then filtered into two frequency bands of interest: μ (8–13 Hz) and β (18–24 Hz) using a 4th-order Butterworth filter. To extract the temporal envelope of the signal, we applied the Hilbert Transform, computing the instantaneous power. A moving average filter with a 1-second buffer was applied to smooth the power features. Before further analysis, an artefact rejection stage was implemented to remove high-amplitude noise. Trials contaminated by eye movements (EOG) or excessive muscle activity (EMG) were identified using threshold-based criteria on specific channels (Fp1, Fp2 for EOG) and rejected from the dataset.

Spatial focality feature extraction:

To distinguish between IC and INC states within a trial, we introduced a high-level feature quantifying the "spatial focality" or focality of the motor activation. For each time sample t , we computed the mean power in three Re-

gions of Interest (ROIs): Left (ROI_L : C3, C1, CP1), Right (ROI_R : C4, C2, CP2), and Centre (ROI_C : Cz, FCz, CPz). The focality feature $F(t)$ is defined as the log-ratio between the median power and the minimum power of the ROIs:

$$F(t) = \log\left(\frac{\text{median}(P_{ROI})}{\min(P_{ROI})}\right) \quad (1)$$

where $P_{ROI} = [P_{ROI_L}, P_{ROI_R}, P_{ROI_C}]$. This metric increases when a specific motor area exhibits strong desynchronization (ERD), indicating a focused high-control state. The choice of this specific feature is driven by the neurophysiological nature of Motor Imagery (MI). While the resting state is typically characterised by widespread synchronised activity (high power) across the sensorimotor cortex, MI induces a focal Event-Related Desynchronization (ERD) in the contralateral hemisphere. By computing the log-ratio between the median power (representing the background activity) and the minimum power (representing the active region), we obtain a metric with 2 characteristics: 1) Class-Agnostic: The min operator automatically identifies the active hemisphere without requiring prior knowledge of the task. 2) Robust to Non-Stationarity: The ratio normalises local activation against the current global power level, making the feature resilient to global drifts in EEG amplitude caused by fatigue or impedance changes. As a result, $F(t)$ yields low values close to zero during synchronised resting states (where $\text{median} \approx \min$) and high positive values during focal motor control (where $\text{median} > \min$).

Unsupervised GMM gating:

We classified the distribution of the spatial focality feature, $F(t)$, using a Gaussian Mixture Model (GMM) with $K = 2$ components, corresponding to two latent states: High Focus (IC) and Low Focus/Rest (INC). Before GMM fitting, the calibration features were normalised. The model was configured with full, unshared covariance matrices and a regularisation parameter of 10^{-5} . The cluster exhibiting the higher mean focality was automatically designated as the IC state. We then computed the posterior probability $P(IC | F(t))$ for each sample and applied a threshold-based gating policy. This allowed us to construct an optimised training set for the subsequent MI-QDA classifier:

$$\text{gate}(F(t)) = \begin{cases} 1 & \text{if } P(IC | F(t)) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

To construct an optimised training set, we applied a threshold-based policy where only samples satisfying $\text{gate}(F(t)) = 1$ were retained. The threshold 0.5 ensures that the MI-QDA classifier is trained on data with a higher probability of belonging to the IC cluster. This mechanism effectively prunes intra-trial idling periods. To validate the overall proposed method, we compared two strategies: 1) Standard Calibration: Using all time samples within the cue period. 2) GMM-Gated Calibration: Using only the samples selected by the GMM.

In addition, we divide the calibration data into train-test

(75%-25%) and for both strategies, we assessed the improvement in data quality, looking at: 1) Fisher Score: To quantify the discriminative power of the features for each channel and frequency band. 2) Signed R-squared: To measure the correlation between the spectral power and the class labels, providing a neurophysiological validation of the selected signal patterns. For the online evaluation, the GMM parameters were estimated using the entire calibration dataset and kept fixed during closed-loop control.

Online gated control architecture:

The calibration framework was translated into a real-time closed-loop system implemented within the ROS-Neuro middleware [21]. The online pipeline processes incoming EEG data streams in chunks of 32 samples (hence there is a downsample from 512 Hz to 16 Hz) and is composed of three hierarchical stages as follows:

1. Shared Signal Processing Pipeline

To ensure consistency between calibration and online control, the real-time system employs the exact same pre-processing chain used in the offline analysis. The raw signal is spatially filtered via CAR, band-pass filtered in the μ (8–13 Hz) and β (18–24 Hz) bands, and the instantaneous power envelope is extracted via Hilbert transform using a ring buffer of 1 second.

2. Parallel Probabilistic Estimation

At each time step t , the system computes two independent probabilistic metrics in parallel:

- Intent Detection (The Gate): The spatial focality feature $F(t)$ is extracted from the motor ROIs. The pre-trained GMM computes the posterior probability of IC $P(IC | F(t))$.
- Task Classification (The Direction): Simultaneously, the pre-trained QDA classifier computes the class posterior probabilities $P(QDA_t | x_t)$ (e.g., hands vs. feet) based on the band power features.

3. Reliability-Weighted "Soft" Integration

The core of the control logic is a dynamic accumulation framework defined as a "Soft-Increment" Integrator. Instead of using a raw classifier output, the system first computes a merged control probability $P_{merged}(t)$ by modulating the QDA output with the GMM confidence:

$$P_{merged} = P(IC) \cdot P(QDA) + (1 - P(IC)) \cdot 0.5 \quad (3)$$

This formulation ensures that as the user focus drops ($P(IC | F(t)) \approx 0$), the signal converges to the neutral state (0.5). Crucially, the control signal update rate is not constant. The integration velocity $v(t)$ is computed as the normalised deviation from neutrality:

$$v(t) = \min(1, |(P_{merged}(t) - 0.5)| \cdot 2 \cdot K_{gain}) \quad (4)$$

K_{gain} is a subject-specific gain empirically tuned to balance system responsiveness and stability. The final control command $C(t)$ is updated at each step:

$$C(t+1) = C(t) + \text{sign}(P_{merged}(t) - 0.5) \cdot \frac{|v(t)|}{b_s} \quad (5)$$

	c7	d7	i1	i2	m1
μ	0.87	0.92	0.82	0.84	0.93
β	0.89	0.78	0.85	0.96	0.86

Table 1: Pearson correlation coefficients (R) between Spatial Sparsity and GMM Confidence.

where, b_s is a fixed parameter set to 64, which defines the system's control horizon. Given the system's update rate of 16 Hz, this b_s establishes a nominal saturation time of 2 seconds. However, expecting the user to reach full saturation ($C(t)=1.0$) often leads to fatigue, as it requires maintaining perfect classification accuracy for the entire window. To improve responsiveness, we implemented an activation threshold (e.g., $Th_{act} = 0.8$). The system triggers the command as soon as $C(t) \geq Th_{act}$. This allows the user to execute tasks efficiently once sufficient cumulative evidence is gathered, without the need to drive the integrator to its asymptotic limit.

This architecture provides a dual advantage: the Gating mechanism ($P(IC | F(t))$) strictly pauses integration during resting states (reducing False Positives), while the Soft-Increment logic (K_{gain}) allows decoupling the system's responsiveness from the buffer length, enabling faster reactions for experienced users when confidence is high.

RESULTS

GMM behaviour and feature validation:

We first analysed the relationship between the extracted spatial focality features and the posterior probabilities $P(IC | F(t))$ generated by the GMM on the test set. This validation was performed separately for the μ (8–13 Hz) and β (18–24 Hz) frequency bands across all five subjects. Fig. 2a illustrates the trend of the focality feature as a function of the GMM confidence. In both frequency bands, a strictly monotonic positive relationship is observed across all subjects (reported in the Tab. 1). More in detail in the μ band (8–13 Hz), the Pearson correlation coefficients (R) indicate a strong linear dependency, ranging from a minimum of $R=0.82$ (Subject i1) to a maximum of $R=0.93$ (Subject m1). While in the β band (18–24 Hz), the correlations remain high, ranging from $R=0.78$ (Subject d7) to $R=0.96$ (Subject i2). Error bars indicate the standard error of the mean for each confidence bin. Fig 2b displays the probability density of the GMM confidence scores ($P(IC)$) grouped by subject. The distributions exhibit a bimodal behaviour with data accumulating at the extremes ($P(IC) \leq 0.1$ and $P(IC) \geq 0.9$). Notable inter-subject variability is observed in the proportion of high-confidence samples. For example, in Fig. 2b Subject i1 (yellow bars) shows a dominant peak in the low-confidence region, indicating a prevalence of resting or non-focal states during the calibration trials. Conversely, Subjects c7 (blue bars) and m1 (green bars) display a higher density of samples in the high-confidence region, suggesting a more stable generation of focal motor patterns.

Enhancement of calibration data quality:

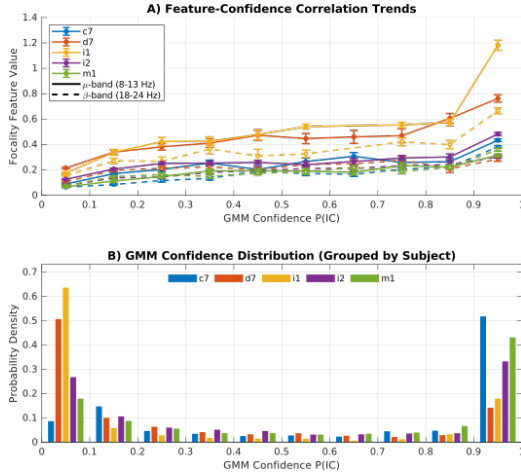


Figure 2: Validation of the unsupervised GMM gating model on the test set. (A) Correlation analysis between the extracted Spatial Sparsity feature and the GMM posterior probability of IC ($P(IC)$). Solid lines represent the μ band (8–13 Hz), while dashed lines represent the β band (18–24 Hz). Error bars denote the Standard Error of the Mean (SEM). (B) Grouped probability density of GMM confidence scores.

To quantify the benefit of the proposed unsupervised gating mechanism, we evaluated the discriminative power of the features extracted from the GMM-selected segments ($P(IC) \geq 0.5$) compared to the Standard calibration approach (using all data within the trial window). Furthermore, we analysed the rejected segments ($P(IC) < 0.5$) to assess the discarded data. All analyses report the grand average across the five subjects on the test set (25% of the calibration trials).

Quantitative analysis (Fisher score and R-square)

Fig. 3 presents the class separability and correlation metrics, analysed separately for the μ (8–13 Hz, left column) and β (18–24 Hz, right column) frequency bands. In both frequency bands, the GMM-selected data (red bars) yielded consistently higher mean separability compared to the standard approach (grey bars) over key sensorimotor channels. Channel by channel non-parametric statistics (Friedman and Wilcoxon) confirmed this enhancement: in the β band, the gating mechanism significantly increased the discriminative power over the primary sensorimotor channel C1 ($p < 0.05$, denoted by *). Furthermore, the GMM dataset demonstrated a marked increase in absolute R^2 coefficients, reflecting an improved Signal-to-Noise Ratio (SNR) while preserving the physiological directionality of task-related modulations (e.g., the negative correlation peak at Cz). Conversely, the rejected data (blue bars) exhibited lower discriminative power and higher cross-subject variability. Statistical comparisons proved that the GMM-isolated intent patterns were significantly superior to these rejected segments over specific channels (CP1 in μ band, FC4 in β band; $p < 0.05$, denoted by *), confirming that the mechanism safely discards non-informative idling periods.

Neurophysiological validation

The topographic distribution of cortical activity visually corroborates the quantitative metrics. Fig. 4 illustrates

the grand average difference maps (Class 1 minus Class 2) for the μ (left) and β (right) frequency bands. The GMM-selected data (top row) retain defined and highly contrasted activation patterns over the sensorimotor areas. In comparison, the standard full-window approach (middle row) yields spatially similar but noticeably attenuated topographies, demonstrating how the inclusion of idling or low-confidence segments dilutes the overall signal intensity. Interestingly, the rejected data (bottom row) still exhibit high-amplitude spectral differences; however, their spatial configuration is noticeably altered.

Online control performance:

The final validation of the proposed architecture was conducted through a real-time, closed-loop control session. To assess the impact of the unsupervised gating mechanism, we evaluated the performance at both the single-sample (frame-by-frame) and system (continuous trial) levels. The quantitative results across the five subjects are summarised in Tab. 2. Although real-time control was driven by QDA_{OUR} , comparison with offline-reconstructed QDA_{TRAD} performance showed comparable sample-wise decoding power ($77.4 \pm 12.7\%$ vs. $77.7 \pm 12.7\%$), indicating that GMM-based filtering preserves robust spatial patterns without degrading baseline model performance. This indicates that filtering the calibration dataset preserves the robust spatial patterns necessary for baseline classification without degrading the underlying model performance. The true advantage of the proposed architecture emerges at the system level, where the sample-wise probabilities are processed by the reliability-weighted integrator. The system achieved a grand average active trial accuracy of $77.5 \pm 12.6\%$ (Acc_{ACT}). Crucially, when excluding the runs that reach the timeouts to evaluate only the trials where the system reached a final decision ($Acc_{ACT_{noT}}$), the accuracy rise to $92.7 \pm 7.1\%$. This demonstrates that the GMM gate effectively translates user uncertainty or loss of focus into safe timeouts rather than executing erroneous commands. This high decisional reliability was achieved with a grand average timeout rate of only $16.5 \pm 5.4\%$, calculated as the ratio between the number of trials without a decision and the total number of trials. Finally, the system’s safety margin was evaluated during the Idle task. The rest trial accuracy (Acc_{REST}), defined as the ability to suppress false positives while the user is not actively intending to control the BCI, reached a grand average of $76.0 \pm 19.5\%$.

DISCUSSION AND CONCLUSION

The development of reliable BCIs for real-world applications requires systems capable of autonomously distinguishing IC from spontaneous neural fluctuations. This study introduced an unsupervised gating architecture driven by a Gaussian Mixture Model (GMM), and the results demonstrate that this approach is highly effective and fundamentally grounded in neurophysiology. The choice of using spatial focality as the primary

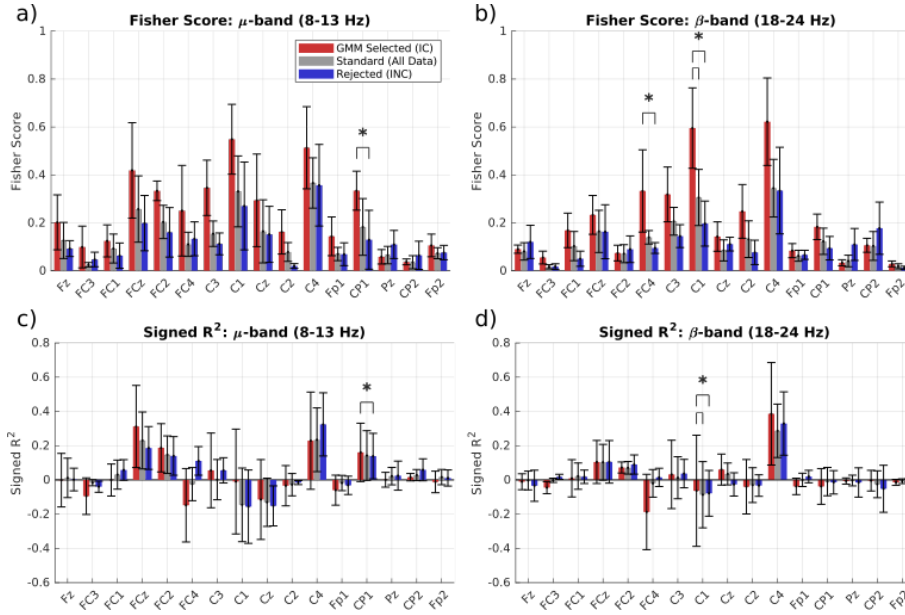


Figure 3: Quantitative assessment of calibration data quality across frequency bands. To account for inter-subject baseline variability, all metrics were subject-wise max-normalised prior to averaging. Top Row: Normalised Discriminative Power (Fisher Score) for the μ (left) and β (right) bands. The GMM-selected dataset (Red) demonstrates superior class separability compared to the full-window Standard method (Grey). Bottom Row: Normalised Class Correlation Direction (Signed R^2). The unsupervised gating amplifies the correlation magnitude (SNR) while preserving the physiological directionality of the signals. The Rejected data (Blue) consistently exhibits the lowest task-relevant modulation. Error bars represent the Standard Error of the Mean (SEM). Statistical significance between the GMM-selected dataset and either the Standard or Rejected approaches over key channels is denoted by asterisks (* $p < 0.05$, one-sided Wilcoxon signed-rank test).

	c7	d7	i1	i2	m1	mean
QDA_{OUR}	61.2 $\pm$ 4.0	93.5 $\pm$ 1.4	80.5 $\pm$ 9.5	83.3 $\pm$ 2.7	68.7 $\pm$ 0.2	77.4 $\pm$ 12.7
QDA_{TRAD}	63.6 $\pm$ 1.2	94.2 $\pm$ 1.0	76.2 $\pm$ 7.7	86.3 $\pm$ 3.3	68.1 $\pm$ 5.7	77.7 $\pm$ 12.7
Acc_{ACT}	62.5 $\pm$ 3.5	72.5 $\pm$ 3.5	85.0 $\pm$ 7.1	95.0 $\pm$ 0.0	72.5 $\pm$ 10.6	77.5 $\pm$ 12.6
$Acc_{ACT_{not}}$	86.2 $\pm$ 0.7	100.0 $\pm$ 0.0	91.8 $\pm$ 4.1	100.0 $\pm$ 0.0	85.4 $\pm$ 1.7	92.7 $\pm$ 7.1
Acc_{REST}	90 $\pm$ 14.1	100.0 $\pm$ 0.0	70 $\pm$ 14.1	70.0 $\pm$ 14.1	50.0 $\pm$ 14.1	76.0 $\pm$ 19.5
$Timeout_{RATE}$	27.5 $\pm$ 3.5	27.5 $\pm$ 3.5	7.5 $\pm$ 3.5	5.0 $\pm$ 0.0	15.0 $\pm$ 14.1	16.5 $\pm$ 5.4

Table 2: Online accuracy. Each row report the mean and the standard deviation of the two online runs.

feature to train the GMM proved to be highly logical and physiologically sound. Genuine Motor Imagery elicits localised Event-Related Desynchronization/Synchronisation (ERD/ERS) over the sensorimotor cortex. By evaluating spatial focality, the GMM captures this focalization, easily discriminating between the localised activity of Intentional Control (IC) and the diffuse, non-informative patterns characteristic of idling states (INC). This physiological rationale is strongly validated by the offline analysis (Fig. 2 and Fig. 3), where the GMM successfully isolated the most informative SMR segments, yielding sharper spatial topographies and higher class separability without requiring any manual data pruning (Fig. 4). During the online closed-loop session, the integration of the GMM probability as a dynamic weight for the control command demonstrated crucial benefits for both active and rest tasks. For active trials, the gating mechanism ensures that the system accumulates evidence only when the user is focused. By converting uncertain neural fluctuations into safe timeouts

rather than erroneous executions, the effective decision accuracy (excluding timeouts) reached 92.7%. Equally important is the system’s behaviour during the rest task. The GMM probability fundamentally alters the integrator’s dynamics: whenever the spatial focality drops (indicating a loss of focus or an intentional idle state), the GMM acts as a stabilising weight that actively pulls the control command back to a central, neutral resting state between the two classes. This mechanism successfully throttled the accumulator during non-intentional periods, maintaining an average safety margin of 76.0% against false positives.

Despite these systemic improvements, some limitations of the current study must be acknowledged. First, the online sample-wise accuracy of the Informed QDA did not show a massive improvement over the Standard QDA (77.7% vs. 77.4%). This indicates that while the GMM cleans the calibration data effectively, the primary benefit of the proposed architecture lies in the temporal gating of the integrator rather than a raw boost in frame-by-frame

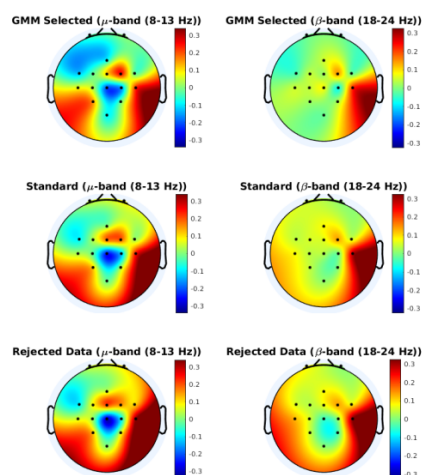


Figure 4: Neurophysiological Validation of the spatial activation patterns. Topographic maps illustrate the mean difference in spectral power between the two motor imagery classes for the μ band (8–13 Hz, left column) and the β band (18–24 Hz, right column). Top Row: The GMM-selected dataset. Middle Row: The Standard full-window. Bottom Row: The rejected dataset.

classification power. Second, the current implementation is restricted to a two-class paradigm. Finally, the validation was conducted on a limited cohort of five subjects with a small number of runs.

In conclusion, this paper presents a robust, unsupervised framework that significantly enhances the safety and reliability of Motor Imagery BCIs. By leveraging a physiologically meaningful GMM to dynamically weight the continuous control command, the system effectively manages both active control and resting states, minimizing dangerous false positive activations. While further validation on larger cohorts and multi-class extensions is necessary, this intent-gated architecture provides a highly practical and scalable solution for bridging the gap between laboratory research and safe, real-world assistive technologies.

REFERENCES

[1] Shenoy P, Krauledat M, Blankertz B, Rao RPN, Müller KR. Towards adaptive classification for BCI. *J. Neural Eng.*. 2006;3(1):R13–23.
 [2] Vidaurre C, Sannelli C, Müller KR, Blankertz B. Co-adaptive calibration to improve BCI efficiency. *J. Neural Eng.*. 2011;8(2):025009.
 [3] Tonin L, Millán JdR. Noninvasive brain-machine interfaces for robotic devices. *Annu. Rev. Control Robot. Auton. Syst.*. 2021;4(1):191–214.
 [4] Saha S, Baumert M. Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: A review. *Front. Comput. Neurosci.*. 2019;13:87.
 [5] Jeunet C, N’Kaoua B, Subramanian S, Hachet M, Lotte F. Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns. *PLoS One*. 2015;10(12):e0143962.

[6] Arvaneh M, Guan C, Ang KK, Quek C. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst.*. 2013;24(4):610–619.
 [7] Pfurtscheller G, Neuper C. Motor imagery and direct brain-computer communication. *Proc. IEEE Inst. Electr. Electron. Eng.*. 2001;89(7):1123–1134.
 [8] Lotte F et al. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *J. Neural Eng.*. 2018;15(3):031005.
 [9] Millán JDR et al. Combining brain-computer interfaces and assistive technologies: State-of-the-art and challenges. *Front. Neurosci.*. 2010;4.
 [10] Leeb R, Tonin L, Rohm M, Desideri L, Carlson T, Millán JdR. Towards independence: A BCI telepresence robot for people with severe motor disabilities. *Proc. IEEE Inst. Electr. Electron. Eng.*. 2015;103(6):969–982.
 [11] Lemm S, Blankertz B, Dickhaus T, Müller KR. Introduction to machine learning for brain imaging. *Neuroimage*. 2011;56(2):387–399.
 [12] Frénay B, Verleysen M. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.*. 2014;25(5):845–869.
 [13] Schlögl A, Keinrath C, Zimmermann D, Scherer R, Leeb R, Pfurtscheller G. A fully automated correction method of EOG artifacts in EEG recordings. *Clin. Neurophysiol.*. 2007;118(1):98–104.
 [14] Fatourechhi M, Bashashati A, Ward RK, Birch GE. EMG and EOG artifacts in brain computer interface systems: A survey. *Clin. Neurophysiol.*. 2007;118(3):480–494.
 [15] Lotte F, Larrue F, Mühl C. Flaws in current human training protocols for spontaneous Brain-Computer interfaces: Lessons learned from instructional design. *Front. Hum. Neurosci.*. 2013;7:568.
 [16] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain-computer interface classification by riemannian geometry. *IEEE Trans. Biomed. Eng.*. 2012;59(4):920–928.
 [17] Yger F, Berar M, Lotte F. Riemannian approaches in Brain-Computer interfaces: A review. *IEEE Trans. Neural Syst. Rehabil. Eng.*. 2017;25(10):1753–1762.
 [18] Tonin L, Leeb R, Tavella M, Perdakis S, Millán JdR. The role of shared-control in BCI-based telepresence. 2010 IEEE International Conference on Systems, Man and Cybernetics - SMC. 2010.
 [19] Kindermans PJ, Verstraeten D, Schrauwen B. A bayesian model for exploiting application constraints to enable unsupervised training of a p300-based BCI. *PLoS One*. 2012;7(4):e33758.
 [20] Verhoeven T, Hübner D, Tangermann M, Müller KR, Dambre J, Kindermans PJ. Improving zero-training brain-computer interfaces by mixing model estimators. *J. Neural Eng.*. 2017;14(3):036021.
 [21] Tonin L, Beraldo G, Tortora S, Menegatti E. ROS-neuro: An open-source platform for neurorobotics. *Front. Neurobot.*. 2022;16:886050.

Confidence-weighted cumulative rCCA with post hoc re-analysis: unsupervised adaptive learning for calibration-free c-VEP BCI

J. Thielen¹

¹Machine Learning & Neural Computing department, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

E-mail: jordy.thielen@donders.ru.nl

ABSTRACT: A brain–computer interface (BCI) typically requires a calibration period to train a decoding model, which is a tedious and time-consuming process. Previous studies have shown high performance of calibration-free decoding approaches, like adaptive reconvolution canonical correlation analysis (rCCA). rCCA can operate in an instantaneous mode, classifying trials independently without prior information, but can be improved by cumulatively learning from previously classified trials. Such cumulative strategy, however, relies on potentially unreliable pseudo-labels. In this study, two extensions to unsupervised rCCA are introduced. First, confidence-weighting ensures that model updates are driven primarily by high-confidence trials. Second, a post hoc re-analysis allows misclassifications to be corrected using a later, more accurate model. Four rCCA variants were evaluated on an existing code-modulated visual evoked potential (c-VEP) dataset, showing no significant effect of confidence-weighting, but a significant improvement of post hoc re-analysis. These findings pave the way for robust and fully calibration-free c-VEP BCI systems.

INTRODUCTION

Brain–computer interfaces (BCIs) based on code-modulated visual evoked potentials (c-VEPs) are among the fastest non-invasive systems using electroencephalography (EEG) [1]. In contrast to BCIs using the frequency-tagging stimulus protocol, which evoke narrow-band steady-state visual evoked potentials (SSVEPs), c-VEP BCIs use a noise-tagging protocol that evokes broadband neural responses [2]. This broadband nature enables high information transfer rates and robust decoding.

Beyond advances in stimulus design aimed at improving efficiency, usability and comfort [3, 4], substantial effort has been devoted to improving decoding for c-VEP BCIs. A limitation of many systems is their reliance on session-specific calibration data acquired during time-consuming passive training sessions [1]. Such calibration requirements hinder immediate use and reduce the practicality of BCIs in real-world applications.

To address this challenge, several strategies have been proposed that can work with small datasets [5]. For instance, the signal-to-noise ratio can be increased through optimized stimulation paradigms, like the noise-tagging

stimulus protocol [1]. Or, prior knowledge can be incorporated into the decoding via careful regularization, like the reconvolution approach [6]. Furthermore, unsupervised methods may eliminate explicit calibration altogether, commonly referred to as zero-training or calibration-free approaches, such as unsupervised adaptive reconvolution [7].

Within the c-VEP domain, multiple calibration-free decoding methods have been explored. Relying on a circularly shifted m-sequence for stimulation, Turi et al. proposed estimating relative temporal lags between unclassified trials combined with a language model [8]. Thielen et al. introduced an adaptive reconvolution embedded in a canonical correlation analysis (CCA), which decodes by directly optimizing model fit [7]. Zheng et al. proposed narrow-band random stimulus sequences combined with filterbank CCA [9]. Finally, Santamaria-Vazquez et al. as well as Behboodi et al. investigated cross-participant transfer learning and foundation models for calibration-free decoding, optionally with fine-tuning [10, 11].

The present study builds on the calibration-free reconvolution CCA (rCCA) framework as proposed by Thielen et al. [7]. Recent work demonstrated that rCCA can operate in two unsupervised modes: instantaneous and cumulative [12, 13]. In the instantaneous mode, each trial is decoded independently using only that trial’s data, making it computationally lightweight and robust to non-stationary data. While this mode is attractive for its simplicity and adaptability, its performance can be improved by allowing the model to learn from previously classified trials, yielding the cumulative variant [12, 13].

In earlier benchmarks, rCCA was evaluated not only on c-VEP data but also on a P300 event-related potential (ERP) dataset to assess its generalization capabilities beyond c-VEP [12, 13]. Additionally, rCCA was compared to unsupervised mean-difference maximization (UMM), a state-of-the-art method originating from the ERP domain [14]. Surprisingly, rCCA remained competitive with UMM even outside its original domain [12, 13].

This finding was notable because UMM incorporates several sophisticated mechanisms, like confidence-weighting. That is, in the cumulative mode, both rCCA and UMM rely on pseudo-labels, meaning that previously classified trials are used for learning, assuming that their predicted labels are correct. UMM explicitly accounts for

the uncertainty of pseudo-labels by applying confidence weighting, such that trials classified with higher confidence have a stronger influence on model adaptation, substantially improving its performance in a calibration-free scenario [14]. In contrast, rCCA treats all pseudo-labeled trials equally, but still performed on par with UMM on both c-VEP [12] and ERP [13] data.

Finally, when using a cumulative adaptive classifier, it is implicitly assumed that the model improves as more data are incorporated. Under this assumption, early unreliable pseudo-labels may be corrected by reanalyzing previously classified trials using a later, more accurate model. Similar post hoc re-analysis strategies have been shown to be beneficial in other unsupervised classification settings, including approaches based on expectation maximization [15] and learning from label proportions [16].

Motivated by these observations, the present study provides a first attempt to extend vanilla rCCA with a confidence-weighting and a post hoc re-analysis mechanism. The performance of these two extensions is systematically compared against the vanilla instantaneous and cumulative variants, with the aim to further improve calibration-free rCCA. All methods are evaluated on one existing open-access c-VEP dataset.

MATERIALS AND METHODS

Dataset: An open-access c-VEP dataset [17] was used as recorded by Thielen et al. [7]. From this dataset, only the offline part was used, in which 30 participants performed a copy-spelling task. EEG was recorded from eight electrodes positioned according to the 10–10 system (Fz, T7, T8, POz, O1, Oz, O2, Iz), using a Biosemi ActiveTwo amplifier at a sampling rate of 512 Hz.

Participants interacted with a 4×5 matrix speller displayed on a 12.9-inch iPad Pro with 60 Hz refresh rate and 1920×1080 resolution. Each key was modulated by a unique pseudo-random binary sequence derived from optimized modulated Gold codes [6], presented at full contrast (white for 1, black for 0). These sequences contained two flash durations (16.67 ms and 33.33 ms) and had a cycle length of 126 bits (2.1 s per cycle at 60 Hz).

Each participant completed five runs, each containing one trial per key in randomized order. Trials began with a 1 s green cue highlighting the target key, followed by 31.5 s of simultaneous flashing of all keys (i.e., 15 cycles), during which participants fixated on the target key. No feedback was provided. In total, each participant contributed 100 trials, containing 5 repetitions per key.

The EEG data were preprocessed as follows. First, a band-pass filter between 6–30 Hz was applied, as found optimal in previous work [12]. Second, the data were segmented into trials from -0.5 s to 32.0 s relative to stimulus onset. Third, the epochs were downsampled to 120 Hz. Finally, the initial and final 0.5 s of each trial were removed to mitigate filtering artefacts while maintaining accurate marker alignment, resulting in 31.5 s trials aligned with stimulation onset.

Algorithm 1 Instantaneous rCCA

Require: stimulus structure matrices $\{\mathbf{M}_i\}_{i=1}^N$

- 1: **for** trial $j = 1$ to K **do**
- 2: **for** stimulus $i = 1$ to N **do**
- 3: $\mathbf{w}_i, \mathbf{r}_i \leftarrow \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 4: $\rho_i \leftarrow \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 5: **end for**
- 6: $\hat{y}_j \leftarrow \arg \max_i \rho_i$
- 7: **end for**

rCCA classification: The current j -th trial is represented by $\mathbf{X}_j \in \mathbb{R}^{C \times T}$, where $C = 8$ denotes the number of EEG channels and T the number of EEG temporal samples. The objective of reconvolution canonical correlation analysis (rCCA) is to identify the attended target key $\hat{y}_j$ for trial j by maximizing the Pearson’s correlation:

$$\hat{y}_j = \arg \max_i \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i) \quad (1)$$

Here, $\mathbf{w}_i \in \mathbb{R}^C$ is a spatial filter and $\mathbf{r}_i \in \mathbb{R}^L$ is the evoked response of length L samples, both associated with the i -th stimulus. The matrix $\mathbf{M}_i \in \mathbb{R}^{L \times T}$ is a Toeplitz representation of the stimulus sequence, encoding the onset, duration, and possible overlap of evoked responses [6]. Similar to previous work [7, 12], a 300 ms response to a short flash, a 300 ms response to a long flash, and a 300 ms response to the onset of stimulation at the start of a trial were modeled at 120 Hz, such that $L = 3 \cdot 0.3 \cdot 120 = 108$. All unsupervised adaptive rCCA classifiers considered in this work use Equation 1 for decoding. They differ only in how the spatial filter $\mathbf{w}_i$ and temporal response $\mathbf{r}_i$ are learned and updated, potentially using previously classified trials.

Instantaneous rCCA: The instantaneous rCCA (Algorithm 1) does not use information from previous trials and thus treats each trial independently. For a given trial j , it estimates sequence-specific spatial filters $\mathbf{w}_i$ and evoked responses $\mathbf{r}_i$ for each of the $N = 20$ candidate stimulus sequences $i \in 1, \dots, N$, using only the current trial $\mathbf{X}_j$:

$$\mathbf{w}_i, \mathbf{r}_i = \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i) \quad (2)$$

Consequently, both the parameter estimation in Equation 2 and the prediction in Equation 1 are performed on the same data $\mathbf{X}_j$, once for each candidate stimulus. Effectively, classification is done by selecting the stimulus that yields the highest model fit to the observed EEG.

Cumulative rCCA: The cumulative rCCA (Algorithm 2) incorporates information from previous trials and can therefore improve classification performance over time as more data become available. That is, for a given trial j , it estimates sequence-specific spatial filters $\mathbf{w}_i$ and evoked responses $\mathbf{r}_i$ for each of the $N = 20$ candidate stimulus sequences $i \in 1, \dots, N$, using all data observed up to and including the current trial $\mathbf{X}_j$. This setting introduces a key challenge: rCCA is inherently supervised and requires labeled training data for

Algorithm 2 Cumulative rCCA

Require: stimulus structure matrices $\{\mathbf{M}_i\}_{i=1}^N$

- 1: **for** trial $j = 1$ to K **do**
- 2: **for** stimulus $i = 1$ to N **do**
- 3: $\mathbf{S}_j \leftarrow [\mathbf{X}_0, \dots, \mathbf{X}_j]$
- 4: $\mathbf{D}_i \leftarrow [\mathbf{M}_{\hat{y}_0}, \dots, \mathbf{M}_{\hat{y}_{j-1}}, \mathbf{M}_i]$
- 5: $\mathbf{w}_i, \mathbf{r}_i \leftarrow \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{S}_j, \mathbf{r}_i^\top \mathbf{D}_i)$
- 6: $\rho_i \leftarrow \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 7: **end for**
- 8: $\hat{y}_j \leftarrow \arg \max_i \rho_i$
- 9: **end for**

calibration. In the absence of ground-truth labels in a calibration-free setting, a pseudo-labeling approach is used, where the predicted labels from previous trials are treated as true labels. The cumulative learning objective is then given by:

$$\mathbf{w}_i, \mathbf{r}_i = \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{S}_j, \mathbf{r}_i^\top \mathbf{D}_i) \quad (3)$$

Here, $\mathbf{S}_j = [\mathbf{X}_0, \dots, \mathbf{X}_j]$ denotes the concatenation of all EEG trials up to the current trial j , and $\mathbf{D}_i = [\mathbf{M}_{\hat{y}_0}, \dots, \mathbf{M}_{\hat{y}_{j-1}}, \mathbf{M}_i]$ is the corresponding concatenation of stimulus structure matrices, using pseudo-labels $\hat{y}_k$ for $k < j$ and the current candidate i .

By definition, the first trial is still processed using instantaneous rCCA (Equation 2) as no previous trials are available. From the second trial onward, the model progressively accumulates training data and updates its parameters accordingly (Equation 3). Importantly, prediction for each trial is always performed using Equation 1, which considers only the current trial data, while learning leverages the full history.

Confidence-weighted cumulative rCCA: The cumulative rCCA uniformly aggregates data from all previous trials to update its parameters, implicitly assuming that all pseudo-labels are correct. In practice, this assumption may be violated. Pseudo-labels may be unreliable due to noise in the EEG signals, even with a well-specified model. Or, pseudo-labels may be unreliable due to model inaccuracies even when the EEG data is of good quality, for instance when too little data are available.

To reduce the impact of unreliable pseudo-labels, confidence-weighted cumulative rCCA (Algorithm 3) is introduced using a confidence-weighting mechanism like used in UMM [14]. The key idea is to assign greater influence to trials that are classified with high confidence and to downweight trials with low confidence during parameter updates.

The confidence c_j of trial j is defined as the normalized correlation margin between the highest winning correlation ρ_w and the second-highest runner-up correlation ρ_r , similar to the margin dynamic stopping method used by Thielen et al. [6]:

$$c_j = \frac{\rho_w - \rho_r}{\sigma}, \quad (4)$$

Algorithm 3 Confidence-weighted cumulative rCCA

Require: stimulus structure matrices $\{\mathbf{M}_i\}_{i=1}^N$

- 1: **for** trial $j = 1$ to K **do**
- 2: **for** stimulus $i = 1$ to N **do**
- 3: $\mathbf{w}_i, \mathbf{r}_i \leftarrow \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 4: $\rho_i \leftarrow \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 5: **end for**
- 6: $c_j \leftarrow (\rho_w - \rho_r) / \sigma$
- 7: **for** stimulus $i = 1$ to N **do**
- 8: $\mathbf{S}_j \leftarrow [c_0 \mathbf{X}_0, \dots, c_j \mathbf{X}_j]$
- 9: $\mathbf{D}_i \leftarrow [c_0 \mathbf{M}_{\hat{y}_0}, \dots, c_{j-1} \mathbf{M}_{\hat{y}_{j-1}}, c_j \mathbf{M}_i]$
- 10: $\mathbf{w}_i, \mathbf{r}_i \leftarrow \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{S}_j, \mathbf{r}_i^\top \mathbf{D}_i)$
- 11: $\rho_i \leftarrow \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 12: **end for**
- 13: $\hat{y}_j \leftarrow \arg \max_i \rho_i$
- 14: **end for**

Here, σ is the standard deviation of all N correlations except the highest winning correlation ρ_w . Thus, c_j increases when the winning candidate is well separated from the runner-up, and when the overall non-maximum correlation distribution is narrow, both indicating higher classification confidence.

These confidence values are used as linear weights in the cumulative learning process. Specifically, the accumulated data matrices in Equation 3 are redefined as $\mathbf{S}_j = [c_0 \mathbf{X}_0, \dots, c_j \mathbf{X}_j]$ and $\mathbf{D}_i = [c_0 \mathbf{M}_{\hat{y}_0}, \dots, c_{j-1} \mathbf{M}_{\hat{y}_{j-1}}, c_j \mathbf{M}_i]$, such that high-confidence trials contribute more strongly to the parameter estimation.

Since the confidence c_j of the current trial must be available prior to model updating, confidence values for all trials are obtained using an instantaneous rCCA. The resulting correlation scores are then used to compute c_j , after which the model is updated accordingly with confidence weighting (Equation 3).

Confidence-weighted cumulative rCCA with post hoc re-analysis: Confidence-weighted cumulative rCCA reduces the influence of unreliable pseudo-labels by downweighting low-confidence trials. Still, even with reduced weights, such trials still contribute to learning and may introduce errors if their pseudo-labels are incorrect.

By construction, cumulative rCCA learns incrementally across trials, such that the model is expected to improve over time. This opens the possibility of correcting earlier mistakes by reclassifying past trials using a later, presumably more accurate model, a process referred to as post hoc re-analysis [15, 16]. In principle, this strategy can rectify early misclassifications and improve the reliability of subsequent updates, since later model adaptations are then based on more accurate pseudo-labels.

In confidence-weighted cumulative rCCA with post hoc re-analysis (Algorithm 4), after classifying the current j -th trial and updating the model, all previous trials are reclassified by applying the updated model via Equation 1, and their pseudo-labels are replaced accordingly, affecting model updates for subsequent trials.

Algorithm 4 Confidence-weighted cumulative rCCA with post hoc re-analysis

Require: stimulus structure matrices $\{\mathbf{M}_i\}_{i=1}^N$

- 1: **for** trial $j = 1$ to K **do**
- 2: **for** stimulus $i = 1$ to N **do**
- 3: $\mathbf{w}_i, \mathbf{r}_i \leftarrow \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 4: $\rho_i \leftarrow \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 5: **end for**
- 6: $c_j \leftarrow (\rho_w - \rho_r) / \sigma$
- 7: **for** stimulus $i = 1$ to N **do**
- 8: $\mathbf{S}_j \leftarrow [c_0 \mathbf{X}_0, \dots, c_j \mathbf{X}_j]$
- 9: $\mathbf{D}_i \leftarrow [c_0 \mathbf{M}_{\hat{y}_0}, \dots, c_{j-1} \mathbf{M}_{\hat{y}_{j-1}}, c_j \mathbf{M}_i]$
- 10: $\mathbf{w}_i, \mathbf{r}_i \leftarrow \arg \max_{\mathbf{w}_i, \mathbf{r}_i} \text{corr}(\mathbf{w}_i^\top \mathbf{S}_j, \mathbf{r}_i^\top \mathbf{D}_i)$
- 11: $\rho_i \leftarrow \text{corr}(\mathbf{w}_i^\top \mathbf{X}_j, \mathbf{r}_i^\top \mathbf{M}_i)$
- 12: **end for**
- 13: $\hat{y}_j \leftarrow \arg \max_i \rho_i$
- 14: $\mathbf{w}, \mathbf{r} \leftarrow \mathbf{w}_{\hat{y}_j}, \mathbf{r}_{\hat{y}_j}$
- 15: **for** trial $k = 1$ to $j - 1$ **do**
- 16: $\hat{y}_k \leftarrow \arg \max_i \text{corr}(\mathbf{w}^\top \mathbf{X}_k, \mathbf{r}^\top \mathbf{M}_i)$
- 17: **end for**
- 18: **end for**

Please note that the confidence values c_j are not affected by this post hoc re-analysis, because they are generally estimated via an instantaneous rCCA.

Analysis: This study systematically compared the performance of four calibration-free decoding methods based on an unsupervised adaptive rCCA: instantaneous rCCA (rcca_i), cumulative rCCA (rcca_c), confidence-weighted cumulative rCCA (rcca_cw), and confidence-weighted cumulative rCCA with post hoc re-analysis (rcca_cwp). All methods were evaluated on an existing c-VEP dataset, with trials processed in their original chronological order to simulate an online decoding scenario and retaining any non-stationarity present in the data.

Performance was quantified using decoding curves, defined as classification accuracy as a function of trial duration. Classification accuracy was evaluated at sixteen trial durations: from 1.05 s to 8.4 s in steps of 1.05 s, from 10.5 s to 16.8 s in steps of 2.1 s, and from 18.9 s to 31.5 s in steps of 4.2 s.

Statistical significance of paired differences between methods at a particular trial duration was assessed using a one-sided Wilcoxon signed-rank test.

RESULTS

For all four methods, the classification accuracy as a function of decoding time, referred to as the decoding curve, is shown in Figure 1. In addition, Table 1 reports the earliest decoding time at which a given grand-average accuracy level is reached, and Table 2 summarizes the grand-average accuracy at selected decoding times. Overall, all methods show a clear and monotonic improvement in performance with increasing decoding

Table 1: **Grand average decoding time.** Listed are the trial durations in seconds to reach a certain grand average classification accuracy.

	0.70	0.80	0.85	0.90	0.95
rcca_i	6.30	8.40	10.50	14.70	27.30
rcca_c	3.15	3.15	3.15	5.25	7.35
rcca_cw	3.15	3.15	4.20	4.20	6.30
rcca_cwp	3.15	3.15	4.20	4.20	6.30

Table 2: **Grand average classification accuracy.** Listed are the grand average classification accuracy reached at distinct trial durations in seconds.

	1.05	2.10	4.20	10.50	31.50
rcca_i	0.07	0.27	0.60	0.87	0.96
rcca_c	0.25	0.54	0.87	0.97	0.97
rcca_cw	0.22	0.55	0.91	0.97	0.97
rcca_cwp	0.24	0.59	0.93	0.97	0.97

time, indicating that access to longer EEG segments consistently enhances classification accuracy.

A first prominent observation is that instantaneous rCCA (rcca_i) yields lower performance than each of the cumulative variants. In particular, its decoding accuracy is significantly lower than that of cumulative rCCA (rcca_c) up to and including 23.1 s ($p < .05$). At longer decoding times (27.3 s and 31.5 s), the differences are no longer statistically significant ($p = .199$ and $p = .463$, respectively), meaning that instantaneous rCCA can reach equal ceiling performance with sufficiently long single-trials. Interestingly, no statistically significant differences are observed between confidence-weighted cumulative rCCA (rcca_cw) and cumulative rCCA (rcca_c) at any decoding time. While rcca_cw shows a small absolute improvements over rcca_c at some time points (2.1 s and 4.2–6.3 s) and slightly worse performance at others (1.05 s and 3.15 s), these effects are minor and disappear for decoding times of 7.35 s and longer, after which both methods yield essentially identical performance.

In contrast, improvements are observed for confidence-weighted cumulative rCCA with post hoc re-analysis (rcca_cwp). Compared to rcca_c, rcca_cwp achieves significantly higher accuracy at 4.2 s (0.93 vs. 0.87, $p = .010$) and 6.30 s (0.96 vs. 0.94, $p = .043$), with no significant differences at other time points ($p > .05$). Importantly, rcca_c is never significantly better than rcca_cwp. Similarly, when compared to rcca_cw, the post hoc re-labeled variant performs significantly better from 2.1 s to 5.25 and 7.35 s to 8.4 s ($p < .05$), while no significant differences are observed at other decoding times. Again, at no time point does rcca_cw outperform rcca_cwp.

DISCUSSION

In this study, an extension of adaptive rCCA was proposed that incorporates confidence-weighted model updates and post hoc re-analysis. Using an existing c-VEP dataset, four calibration-free decoding strategies

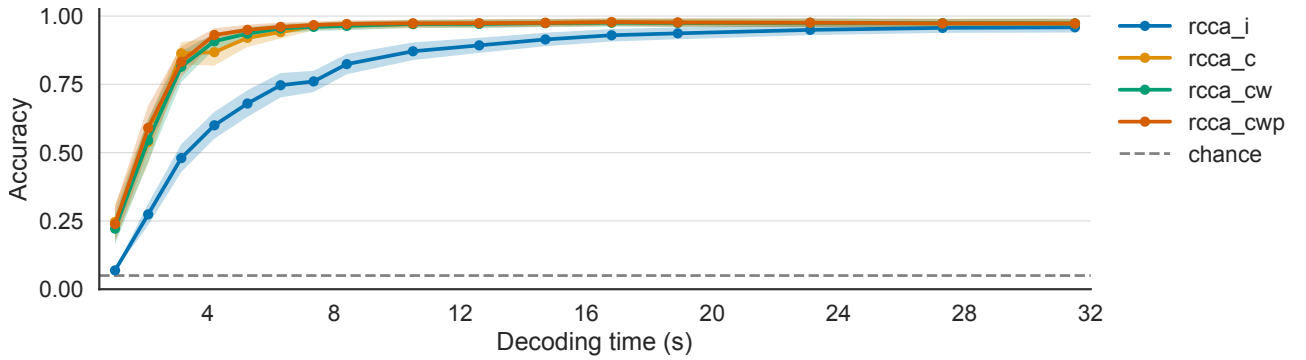


Figure 1: **Grand average decoding curve.** The mean calibration-free classification accuracy and its standard error are shown across varying single-trial durations from 1.05 s to 31.5 s. Curves indicate performances for instantaneous rCCA (rcca_i), cumulative rCCA (rcca_c), confidence-weighted cumulative rCCA (rcca_cw), and confidence-weighted cumulative rCCA with post hoc re-analysis (rcca_cwp). The dashed gray line denotes the theoretical chance level (0.05 %).

were evaluated: instantaneous rCCA, cumulative rCCA, confidence-weighted cumulative rCCA, and confidence-weighted cumulative rCCA with post hoc re-analysis. These methods differ in how, and to what extent, previously classified data are exploited for model adaptation. An important observation is that, while in general yielding lower performance than its cumulative alternatives, instantaneous rCCA ultimately reaches a similar ceiling performance as cumulative rCCA, given sufficient data. Because instantaneous rCCA relies exclusively on the current trial, it does not benefit from cross-trial learning. This makes instantaneous rCCA the simplest and most robust variant with respect to non-stationarities, as it adapts solely to local signal statistics.

Consistent with earlier findings [12, 13], cumulative rCCA significantly outperforms instantaneous rCCA. This confirms that learning from previously classified trials substantially improves model performance. However, cumulative rCCA relies on pseudo-labels, assuming that previously predicted labels are correct, which introduces the risk that unreliable labels degrade model estimates.

To mitigate this issue, the confidence-weighted cumulative rCCA was introduced, which downweights low-confidence trials and emphasizes high-confidence ones during model updates. Contrary to expectations and to earlier results with UMM [14], this confidence-weighting did not yield significant performance gains over vanilla cumulative rCCA. Although rCCA and UMM are structurally similar, differences in the confidence metric and in how weighting is applied (e.g., on averages or also covariance estimates) may explain this discrepancy.

While confidence-weighting reduces the influence of unreliable trials, such trials may still carry incorrect pseudo-labels. Therefore, post hoc re-analysis was introduced, which reconsiders previously classified trials after each model update. It is assumed that a more informed model produces more accurate pseudo-labels. Indeed, confidence-weighted cumulative rCCA with post hoc re-analysis significantly outperformed both vanilla and confidence-weighted cumulative rCCA at shorter trial durations, indicating that iterative correction can meaning-

fully improve calibration-free decoding, in line with other studies showing such improvement [15, 16].

It is worth noting that the instantaneous and cumulative rCCA results reported here are slightly higher than those in earlier work [12], despite using the same dataset and methods. The only difference lies in the downsampling rate: 120 Hz in the present study versus 180 Hz previously. The lower sampling rate may have reduced model capacity and the risk of overfitting, leading to marginally improved performance.

Although the proposed extensions demonstrate the feasibility of confidence-weighting and post hoc re-analysis for calibration-free rCCA, several open questions remain. First, the confidence metric used here, the normalized correlation margin, was inspired by earlier dynamic stopping work [6] and by UMM [14]. While it clearly separated correct from incorrect predictions, its lack of impact on performance suggests that alternative confidence measures may be more effective. Furthermore, the confidence-weighting was applied at the trial-level decoding in this work. Instead, UMM applies the confidence-weighting at the epoch-level which may influence the impact it has on the decoding process [14].

Second, confidence values were estimated using instantaneous rCCA to ensure availability for the current trial. An alternative is to derive confidence from the most recent cumulative model, although preliminary experiments in this study did not show improvements.

Third, post hoc re-analysis was limited to updating pseudo-labels only. If confidence values were estimated using cumulative rCCA, these too could be re-estimated, potentially yielding additional benefits.

Fourth, the cumulative variants maintain a complete history of all previous trials. This is memory-intensive and may become impractical in long sessions. Moreover, retaining all past data may be suboptimal under non-stationary conditions. Future work could explore sliding windows like first-in–first-out buffers, or exponential forgetting mechanisms to balance memory constraints, robustness as well as adaptability.

Finally, both this and previous studies benchmark

calibration-free methods using offline datasets [12, 13]. While this enables controlled comparisons, generalization to real-world use must ultimately be validated in online experiments, as demonstrated in earlier work [7]. In such calibration-free online scenario, decoding curves cannot be estimated, and static stopping times cannot be pre-calibrated. Instead, dynamic stopping strategies are required, ideally also in a calibration-free manner. The normalized correlation margin used here could serve this purpose, although it requires a predefined threshold [6]. Alternative dynamic stopping approaches that avoid explicit calibration [7] warrant further investigation.

CONCLUSION

This study evaluated calibration-free adaptive decoding for c-VEP BCIs by extending reconvolution canonical correlation analysis (rCCA) with confidence-weighted updates and post hoc re-analysis. Four variants were compared: instantaneous rCCA, cumulative rCCA, confidence-weighted cumulative rCCA, and confidence-weighted cumulative rCCA with post hoc re-analysis. The results confirm that cumulative learning substantially improves decoding performance over instantaneous classification. Furthermore, while confidence-weighting alone did not significantly outperform vanilla cumulative rCCA, the addition of post hoc re-analysis yielded consistent improvements, demonstrating that iterative correction of pseudo-labels is an effective strategy in calibration-free settings.

Overall, these findings show that adaptive rCCA can be made more robust by explicitly addressing pseudo-label uncertainty, thereby further reducing the need for calibration and advancing practical plug-and-play c-VEP BCIs.

ACKNOWLEDGEMENTS

This work was supported by the Dutch Research Council (grant 024.005.022). This project has received funding from the European Union's research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101118964.

REFERENCES

- [1] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): A literature review. *Journal of Neural Engineering*. 2021;18:061002.
- [2] Gao S, Wang Y, Gao X, Hong B. Visual and auditory brain-computer interfaces. *IEEE transactions on Biomedical Engineering*. 2014;61(5):1436–1447.
- [3] Thielen J. Addressing BCI inefficiency in c-VEP-based BCIs: A comprehensive study of neurophysiological predictors, binary stimulus sequences, and user comfort. *Biomedical Physics & Engineering Express*. 2025.
- [4] Dehais F, Castillos KC, Ladouce S, Clisson P. Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry EEG code-VEP BCI. *Journal of Neural Engineering*. 2024;21(6):066023.
- [5] Tangermann M et al. Learning from small datasets—review of workshop 6 of the 10th international BCI meeting 2023. *Journal of Neural Engineering*. 2025.
- [6] Thielen J, Broek P van den, Farquhar J, Desain P. Broad-band visually evoked potentials: Re(con)volution in brain-computer interfacing. *PLOS One*. 2015;10(7):e0133797.
- [7] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials for brain-computer interface. *Journal of Neural Engineering*. 2021;18(5):056007.
- [8] Turi F, Gayraud T, Clerc M. Auto-calibration of c-VEP BCI by word prediction. *HAL Open Science*. 2020.
- [9] Zheng L, Dong Y, Tian S, Pei W, Gao X, Wang Y. A calibration-free c-VEP based BCI employing narrow-band random sequences. *Journal of Neural Engineering*. 2024;21:026023.
- [10] Santamaría-Vázquez E, Martínez-Cagigal V, Ruiz-Gálvez R, Martín-Fernández A, Pascual-Roa B, Hornero R. Towards calibration-free user-friendly c-VEP-based BCIs: An exploratory study using deep-learning. In: *Converging Clinical and Engineering Research on Neurorehabilitation V*. Springer Nature Switzerland: Cham, 2025, 685–689.
- [11] Behboodi M, Kinney-Lang E, Etemad A, Kirton A, Abou-Zeid H. Leveraging foundation models for calibration-free c-VEP BCIs. In: *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2025, 4564–4571.
- [12] Thielen J, Sosulski J, Tangermann M. Exploring new territory: Calibration-free decoding for c-VEP BCI. In: *9th Graz Brain-Computer Interface Conference 2024*. 2024, 325–330.
- [13] Thielen J, Tangermann M. Exploring new territory II: Calibration-free decoding for ERP BCI. In: *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2025, 3788–3793.
- [14] Sosulski J, Tangermann M. UMM: Unsupervised mean-difference maximization. *arXiv preprint arXiv:2306.11830*. 2023.
- [15] Kindermans PJ, Schreuder M, Schrauwen B, Müller KR, Tangermann M. True zero-training brain-computer interfacing—an online study. *PLOS One*. 2014;9(7):e102504.
- [16] Hübner D, Verhoeven T, Schmid K, Müller KR, Tangermann M, Kindermans PJ. Learning from label proportions in brain-computer interfaces: Online unsupervised learning with guarantees. *PLOS One*. 2017;12(4):e0175856.
- [17] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. Version 3. 2023. [Online]. Available: <https://doi.org/10.34973/9txv-z787>.

Realtime Class Balancing for Efficient Decoder Adaptation in Asynchronous Brain-Computer Interface

Rémi Souriau¹, Félix Martel¹, Thomas Costecalde¹, Serpil Karakas¹, Tetiana Aksenova¹

¹Univ. Grenoble Alpes, CEA, Leti, Clnatec, Grenoble, France

E-mail: tetiana.aksenova@cea.fr

ABSTRACT: Brain-Computer Interfaces (BCIs) allow direct control of external effectors via brain activity decoding. For real-world use, BCIs must operate in an asynchronous mode that enables users to initiate or withhold system use, while maintaining low false positive rates (FPR) and high decoding accuracy. Class balance is crucial for efficient model training. The RSW-NPLS algorithm has been proposed recently to balance classes during online and incremental learning. However, perfect balancing may not be optimal for asynchronous BCIs, as they require a highly stable idle state. We propose a class balancing strategy that prioritizes the idle state to reduce its false activations of the other states, and evaluated it on four databases from tetraplegic and paraplegic users. It significantly reduced the FPR (between -55.36% and -70.08%) with minimal loss of accuracy (between -3.68% and -10.51%).

INTRODUCTION

A Brain-Computer Interface (BCI) enables a person to control an effector, such as a wheelchair, an exoskeleton [1] or an electrical stimulator [2], by performing motor imagery or attempting movements. This control is achieved by recording the person's neuronal activity. Brain signals are processed, and a decoder predicts the user's intention to control an effector. Motor BCI systems can be used to compensate lost functions or in rehabilitation programs [3].

Efficient decoding of user's motor intention remains a major challenge to guarantee satisfactory performance of BCI systems. Numerous decoding algorithms were proposed to address this challenge. However, many of them are limited to operating in a synchronous mode within controlled laboratory environments. In such experiments, the activation time of the BCI system is predetermined and fixed by the system. In contrast, autonomous BCIs suitable for everyday use should support an asynchronous mode, including an idle state (or resting state). Users should be able to activate and to deactivate the BCI system at their own will, and it is essential to prevent / minimize the unintended effectors activation. This makes the Idle State (IS) distinct from other Active States (AS). In addition, practical BCI systems should be compact (i.e., portable) and capable of online operation for both prediction and decoder training. The learning algorithm has to

work online to take into consideration the user feedback, and has to support learning since new states can be added to the decoder after some update iterations.

Some tasks may be complex and require longer training times than other states. In addition, participant training often require rest periods to avoid user fatigue. For these reasons, the experimenter cannot properly balance the different states, leading to class imbalance in the training data and degraded decoding performance [4]. In the case of real-time, streaming learning of the BCI decoder, class balancing must be performed online. This poses a challenge because not all data are known, and therefore the final class imbalance is unknown. Despite of the importance of class balancing, this problem has been rarely addressed in BCI studies, *e.g.* [4–6]. In [4], the proposed Recursive Sample Weighted N-way Partial Least Squares regression (RSWNPLS) is integrated to a Hidden Markov Model (HMM). The HMM-RSWNPLS algorithm is well adapted to online updates supporting class balancing for multi-class classification during incremental learning, while being compact. It incorporates a weighting mechanism in order to balance each class during incremental learning of the decoding model, and significantly improves the classification accuracy compared to the same algorithm without balancing mechanism. However, HMM-RSWNPLS, and more generally algorithms to control class balance in BCIs [4–6], treat the IS like any other state. This solution may be sub-optimal for asynchronous BCI where preventing false activation of the system is one of critical issue.

In this study, we propose a class balancing strategy to target any class ratio, while incrementally learning the decoding model. More specifically, we studied the impact of balancing towards the IS to promote a better stability and limit false activations of the BCI. The proposed strategy was integrated into the HMM-RSWNPLS and tested in a pseudo-online manner on 4 different databases recorded from one tetraplegic and two paraplegic participants. The unwanted activation of the BCI is quantified with the False Positive Rate (FPR). We evaluated variations of the FPR and the decoding accuracy in function of the balance between the IS and AS. Results show that the proposed algorithm can reduce significantly the FPR without significantly degrading the balanced accuracy. It also decreases the global latency and the noise level of IS predictions.

MATERIALS AND METHODS

Decoders for Asynchronous BCI:

Most BCI decoders consist in classification models where each class (or state) corresponds to a task that the user can perform. Numerous classifiers have been evaluated, including both static and dynamic approaches. Sample-wise static classifiers like linear discriminant analysis, support vector machine, logistic regression, k-nearest neighbors, or deep neural networks have been widely tested [7, 8]. Since asynchronous BCI decoders aim to decode temporal signal, dynamic classifiers are preferred [9] to capture temporal dependencies in recorded signals. In [10] tests show that taking the dynamic of data into account improves the decoder performance. HMM [1, 2] and Hierarchical HMM [11] have been proposed to decode user intents in an online manner. One of the main issues in asynchronous BCI systems is ensuring the accurate decoding of the IS to prevent false activations, where the system performs an action when the user does not intend it. Several studies have proposed BCI systems designed for asynchronous operation to control various effectors, such as robotic arms [12], wheelchairs [13], or environmental control systems [14]. These studies have explored diverse strategies for managing the IS. In [12], the authors used a threshold-based method to distinguish IS from AS on ElectroEncephaloGram (EEG) signals. Another approach uses a heuristic rule to detect the IS [13], and if the IS is not detected, a classifier is applied to find the desired AS. Another strategy uses hierarchical classifiers [11], a first classification model separates the IS from the AS, followed by a subsequent classification of the AS. The presented approaches aim to separate the IS from AS to minimize false activations. However, these methods do not allow to control the trade-off between decoding accuracy and false activations, and may inadvertently favor IS over AS too much. As a result, the decoder could become more likely to overlook an AS, potentially leading to a decline in overall accuracy.

Studies cited below evaluate model accuracy and the number or ratio of false activations, but, to the best of our knowledge, none incorporate a mechanism in the decoder to allow users to control the trade-off between accuracy and false activation limitation. In this study, the HMM-RSWNPLS algorithm, proposed in [4] is adapted to control the ratio between the IS and the AS to limit false activations and maintain a good classification accuracy in asynchronous BCI.

HMM-RSWNPLS algorithm:

The HMM is a dynamic classifier adapted for multi-class classification tasks. It is composed of two stochastic processes: the latent process $\{Y_t\}_{t=1}^T$ and the observation $\{X_t\}_{t=1}^T$ process which refers to, respectively, the user intent and the neural features. Two probabilistic laws are learnt to model the dynamic evolution of these processes: the emission law that provides $\Pr(X_t|Y_t)$ and the transition law $\Pr(Y_t|Y_{t-1})$. The transition law is modelled using

a transition matrix $A \in \mathbb{R}^{(K \times K)}$, where K is the number of states/classes of the model. Each coefficient of the matrix $a_{ij} = \Pr(Y_t = j|Y_{t-1} = i)$ is learnt by counting the transitions during the training session. The RSWNPLS [4] is used to learn the emission model. This algorithm learns a linear relationship between the user intent Y_t and the neural feature X_t in an incremental way. Data are acquired buffer by buffer and are discarded after each update of the model to keep the memory load stable.

The model allows to compute a probability for each state at every time index t , resulting in a prediction probability function over time for each state [1]. First, the output of the RSWNPLS is computed resulting in an estimation vector $\hat{Y}_t$. The probability of the k -th state $\Pr(Y_t = k|X_t)$ is then deduced from $\hat{Y}_t$ using the soft-max function. In this study, a scaling parameter θ is applied in soft-max function [15]: $\Pr(Y_t|X_t) = \text{soft-max}(\theta \times \hat{Y}_t)$. The prediction probability $\Pr(Y_t = k|X_{1:t})$ (where $X_{1:t}$ is the sequence of observation from the initial time step to t) is computed using the Bayes' rule and the HMM's assumption (first-order Markov model).

To compensate potential class imbalance, the RSWNPLS weights samples during the learning of the model. The weighting procedure is designed such that weighting data is equivalent to resampling the data, but can use non-integer values, which allows for greater flexibility in strategies to balance classes. Let $N_u^{(k)}$ be the cumulated size of the class k at iteration update u , $n_u^{(k)}$ be the size of the class k in the current buffer at iteration update u , λ be the forgetting factor, and $w_u^{(k)}$ be the weight applied to each element in the class k of the current buffer at iteration update u . For each k , the update rules of $N_u^{(k)}$ is given by:

$$N_u^{(k)} = \lambda N_{u-1}^{(k)} + w_u^{(k)} n_u^{(k)}. \quad (1)$$

To balance classes in [4], the weight applied to the current majority class is equal to 1. Weights of other classes are computed relatively to the majority one:

$$w_u^{(k)} = \frac{N_u^{(maj)} - \lambda N_{u-1}^{(k)}}{n_u^{(k)}} \quad (2)$$

where $N_u^{(maj)}$ is the size of the current majority class:

$$N_u^{(maj)} = \max_k (\lambda N_{u-1}^{(k)} + n_u^{(k)}). \quad (3)$$

A saturation is applied to $w_u^{(k)}$ to limit the maximum value of weights to avoid a few samples getting excessively high weights.

Modification on the RSWNPLS to control the class imbalance ratio:

The weight computation presented in the previous section only allows to balance all classes without any distinction. The weighting computation is modified in this study to control the class imbalance during the model learning. A target ratio between classes is defined, and the weighting

computation aims to maintain the ratio of between classes during the incremental learning of the decoder.

Let $\{r^{(k)}\}_{k=1}^K$, subject to $\sum_{k=1}^K r^{(k)} = 1$ and $r^{(k)} > 0$ for all k , be the desired ratio of each of the K classes. The updated weight computation is given by:

$$w_u^{(k)} = \frac{r^{(k)} N_u^{(maj)} - \lambda N_{u-1}^{(k)}}{n_u^{(k)}} \quad (4)$$

where $N_u^{(maj)}$ is the size of the current majority class given by:

$$N_u^{(maj)} = \max_k \left(\frac{\lambda N_{u-1}^{(k)} + n_u^{(k)}}{r^{(k)}} \right). \quad (5)$$

By replacing $w_u^{(k)}$ in equation (1) with its expression from equation (4), one can show $N_u^{(k)} / \sum_j N_u^{(j)} = r^{(k)}$. This weight computation allows to control the class imbalance between states. Similarly to the generic algorithm [4], a saturation is imposed on the weights to bound their maximal value.

Database:

Four databases have been used to test the proposed strategy. The first two come from the “BCI and Tetraplegia” clinical trial (ClinicalTrials.gov, NCT02550522) [1], the third database come from the “STIMO-BSI” clinical trial (clinicaltrials.gov, NCT04632290) [2], and the last database has been recorded in the context of the “THINK2GO” clinical trial (clinicaltrials.gov, NCT06243952, no publication yet). All participants were implanted with either one or two WIMAGINE [16] implant(s) designed to record multi-channels ECoG signals with a sampling frequency of 586 Hz. Tab. 1 provides a description of each database.

In BCI-Exo5, the participant was trained to control a virtual exoskeleton. Neural features are classified among five states: right hand translation (AS_{RH}), left hand translation (AS_{LH}), right wrist rotation (AS_{RW}), left wrist rotation (AS_{LW}) and IS. In BCI-Exo8 the same participant was trained to control a virtual exoskeleton with eight states: the same five states plus left grasp (AS_{LG}), right grasp (AS_{RG}), and walk (AS_{RUN}). In BSI-1, the participant was trained to control an implanted electrical epidural stimulator to activate their own left leg and right leg to walk. Data were classified among three states: left hip (AS_{LHp}), right hip (AS_{RHp}) flexion and IS. In BSI-2, the participant had to control an implanted electrical epidural stimulator to use their own legs and walk similarly to the participant of the BSI-1 database.

During each recording session, an experimenter was next to the participant. The experimenter cued the participant to perform specific tasks (or to idle) through a software. The participant had to perform the mental task associated to the cued task, which is considered as the ground truth of motor intent for machine learning purposes. The number of trials per state was not fixed in advance, each state was repeated as many times as necessary until the experimenter judged that the model was effectively decoding

the task in question.

Features extraction:

The Continuous Complex Wavelet Transform (CCWT) using the Morlet mother wavelet was used to compute neural features every 100ms. A decimation step was then applied to reduce the temporal dimensionality and set the sampling frequency to 10 Hz. The experimenter configured the frequency settings and the time length of the epoch (number of 100ms slices) thanks to the recording software.

The model’s input neural features formed a tensor $X_t \in \mathbb{R}^{T \times F \times C}$ with T epoch duration (number of 100ms slices), F is the number of frequencies and C the number of channels.

Performance evaluation:

From the saved databases, we retrain models with different desired balance between the IS and AS. The desired balance ratio was encoded with the following vector $\tilde{r} = [\tilde{r}^{(IS)} \ 1 \ \dots \ 1]$, where $\tilde{r}^{(IS)}$ is the ratio of the IS among AS. Different values for $\tilde{r}^{(IS)}$ were tested: from 1 to 5 with a 0.5 step. The desired ratio vector of each state is then normalized to ensure the sum of ratio is equal to 1: $r = \tilde{r} / \|\tilde{r}\|$. The vector r is then used in equation (4-5).

Leave-one-out cross-validation was employed for each database, wherein all but one recording session were used for model training, and the held-out session was used for evaluation.

Models are trained in a pseudo-online configuration. The data are presented to the learning algorithm sequentially, as if the model was trained in real-time. The advantage of pseudo-online testing is that it enables the evaluation of a model without the necessity of conducting a real experiment with a user, thereby reducing both cost and time. The limitation of this approach is that it does not incorporate user feedback during the model’s learning process as it does during online experiments. Four criteria are used to validate our approach, evaluating the decoder accuracy, false activations via the false positive rate, noise level in the prediction, and latency of the decoding.

The balanced accuracy ($BAcc$) metric is used to evaluate the model’s ability to classify data correctly. Increasing $\tilde{r}^{(IS)}$ will favor the prediction of the IS in comparison with AS. It is necessary to ensure the algorithm does not degrade excessively the $BAcc$ while decreasing the number of false activations. This metric is insensitive to class imbalance in the testing set.

The false positive rate (FPR) measures how much the model predicts AS while the IS is the desired state. For this criterion, merge all the AS to form the positive class and the IS is considered as the negative class. By increasing $\tilde{r}^{(IS)}$, we expect the FPR to decrease and tend to 0. The FPR is given by: $FPR = 1 - \text{Rec}^{(IS)}$ where $\text{Rec}^{(IS)}$ refers to the recall of the IS.

In addition, the noise level (NL) of the IS probability prediction is evaluated. The NL is estimated by computing the Shannon entropy of the probability prediction. In the ideal case where the prediction probability is binary, the Shannon entropy of the predicted probability should

Table 1: Description of databases.

Name of database	Number of implants	Implant(s) location	Motor impairment	Number of sessions	Mean duration of session	Recorded X months after implantation
BCI-Exo5	2	Sensorimotor cortex see [1], figure 2	Tetraplegia	8	46 min	16 to 17
BCI-Exo8				10	29 min	31 to 32
BSI-1	2	Sensorimotor cortex see [2], figure 1.b	Paraplegia	26	5 min	2 to 6
BSI-2	1	Sensorimotor cortex on the interhemispheric fissure	Paraplegia	24	14 min	1 to 4

be equal to zero. Noisy predictions lead to increase the Shannon entropy. We propose to modify the Shannon entropy function to estimate the NL of the IS:

$$NL = \frac{\log(2)}{e^{-1}} \times \frac{1}{T} \sum_{t=1}^T \tilde{Y}_{t,y=IS}^{(IS)} \log_2 \left(\tilde{Y}_{t,y=IS}^{(IS)} \right), \quad (6)$$

where $\tilde{Y}_{t,y=IS}^{(IS)}$ is the probability prediction of the IS at the time iteration t when the current desired state is the IS. The modification of the Shannon entropy (mean instead of the sum) and adding the coefficient $\frac{\log(2)}{e^{-1}}$ bounds the criterion between 0 and 1, but keeps it equal to 0 when $\tilde{Y}_{t,y=IS}^{(IS)}$ is binary. Note that the noise level is only computed for the samples corresponding to the desired IS, to evaluate the stability of the IS.

Additionally, the latency measures the delay between the cue given by the experimenter and the prediction of the cued state. The latency is defined as the time shift maximizing the cross correlation, given by:

$$Latency = \operatorname{argmax}_{\tau} \left(\sum_k E \left[Y_t^{(k)} \tilde{Y}_{t-\tau}^{(k)} \right] \right), \quad (7)$$

with $\tau > 0$.

The latency has an impact on the $BACC$, the FPR and the NL . To compute these criteria, the prediction signal is shifted after computing the global latency to “align” the prediction signal with the target signal. The $BACC$, the FPR and the NL are computed using the shifted prediction as in [17].

In addition to criteria evaluating the decoder performance, the class imbalance ratio (CIR) is provided to measure the class imbalance in the training dataset. The CIR is equal to the size of the majority class divided by the size of the minority class, taking into account the weights $w_u^{(k)}$ at each iteration [4]. A CIR equal to 1 means the training database has been perfectly balanced. It is expected that CIR tends toward $\tilde{r}^{(IS)}$ for all experiments.

CONFIGURATION

Tab. 2 gives the feature configuration for each database. The tensor size is given in the format $[T, F, C]$ with T epoch duration (number of 100ms slices), F is the number of frequencies and C the number of channels.

From the study on the HMM-RSWNPLS [4], we empirically chose a maximum weight equal to 10, to help the model balance the classes easily without having weights

Table 2: Feature settings.

Database	Tensor size	Frequencies (Hz)
BCI-Exo5	[10, 15, 64]	[10:10:150]
BCI-Exo8	[10, 15, 64]	[10:10:150]
BSI-1	[2, 24, 64]	[2 5:5:100 125 150 200]
BSI-2	[2, 29, 32]	[2 10:5:100 110:10:200]

large enough to bias the model and degrade its performance.

A test on unused sessions from the BSI-1 database shows that the scaling parameter $\theta = 3$ decreases the latency by 33% and keeps the balanced accuracy stable (+0.08%). We fixed $\theta = 3$ for all the computations with all databases.

RESULTS

Fig. 1 displays the distribution of each metric. In each plot, the distributions are grouped by database and presented in ascending order of $\tilde{r}^{(IS)}$. The Mann-Whitney U-test is performed to evaluate the significance of between the first distribution ($\tilde{r}^{(IS)} = 1$) and the other distributions ($\tilde{r}^{(IS)} > 1$) except for the CIR .

The RSWNPLS succeeds to control the class imbalance for the BSI-1 and the BSI-2 for every ratio $\tilde{r}^{(IS)}$. For the BCI-Exo5, the RSWNPLS succeeds to control the class balance for $\tilde{r}^{(IS)} \leq 4$ (see Fig. 1C). However, for $\tilde{r}^{(IS)} > 4$, the RSWNPLS fails to make the CIR equal to $\tilde{r}^{(IS)}$. This observation is even more pronounced with the BCI-Exo8 where the RSWNPLS never reaches the expected CIR .

The mean $BACC$ (Fig. 1A) tends to decrease as the ratio increases for all databases. The decrease is significant with regards to the Mann-Whitney U-test only after $\tilde{r}^{(IS)} = 3$. Medians and dispersion of the FPR (Fig. 1D) and the NL (Fig. 1B) appear to decrease exponentially as $\tilde{r}^{(IS)}$ increases. In three databases, the difference in these scores is already significant when $\tilde{r}^{(IS)}$ increases from 1 to 1.5. The latency (Fig. 1E) decreases slightly when $\tilde{r}^{(IS)}$ increases. The median latency increases in the BCI-Exo5 database for $\tilde{r}^{(IS)} > 4$. Change in the latency distribution are significant with regards to the Mann-Whitney U-test only for $\tilde{r}^{(IS)} > 3$, except for the BCI-Exo8 databases.

The Fig. 2 illustrates two examples of prediction centered around one cue of the IS, for $\tilde{r}^{(IS)}$ equal to 1 and 3. These examples show the probability prediction of the IS is smoother when $\tilde{r}^{(IS)} = 3$. Also, there is less confusion between the IS and AS when the cue is IS. Note that some AS predictions are noisier since the model favors IS predictions.

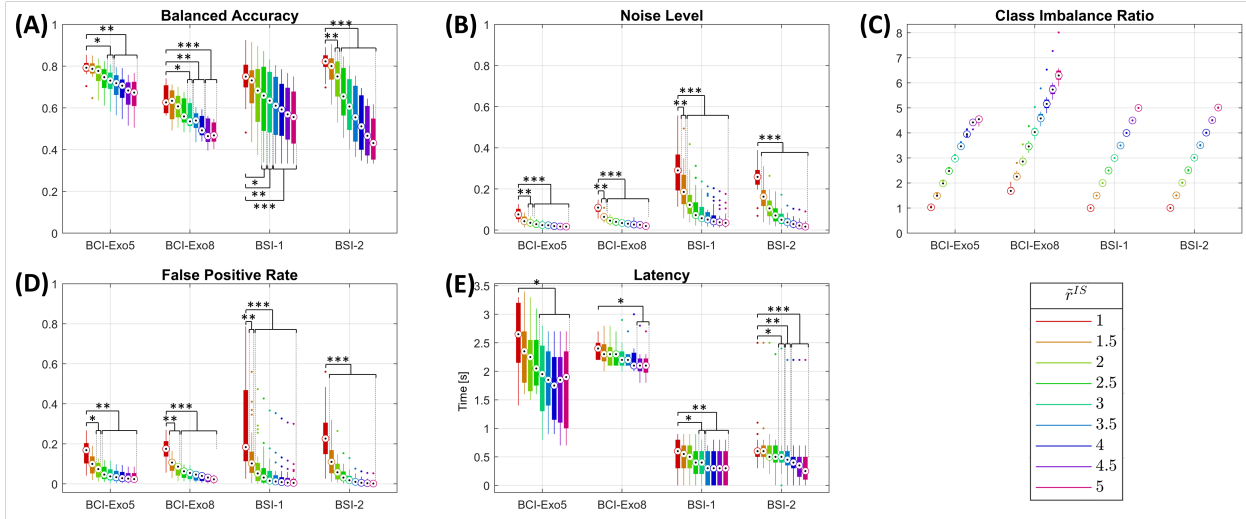


Figure 1: Performance criteria per database (A: $BACC$, B: NL , C: CIR , D: FPR and E: $Latency$). In each plot, the four group correspond to the four databases. Boxplots are given for each $\tilde{r}^{(IS)}$ from 1 to 5. Note that for the CIR that the dispersion is very low, but multiple points are displayed. The Mann-Whitney U-test is performed between the first boxplot and each other to evaluate the hypothesis that the two distributions are different. Significant p-values are displayed (* when $p < 0.05$, ** when $p < 0.01$ and *** when $p < 0.001$).

DISCUSSION

A primary goal of recent BCI development is to create home-use systems that can be used outside of a laboratory setting and without external assistance. Recent publications addressed the related issues. In particular, the HMM-RSWNPLS algorithm was proposed in [4], aimed to simplify real-time model training by automatically balancing the classes using sample weighting. Previously, experimenters had to manually ensure class balance during model training [1, 2]. The HMM-RSWNPLS relieves experimenters of this task representing a step toward unassisted BCI use. One limitation of the weighting strategy proposed in [4] is that the IS is treated equally with all other states, potentially increasing the FPR of the system. For home-use, unwanted activations of the BCI system can be particularly perturbing and even dangerous, depending on the effector activated without the user's intent. This study explores a new weighting strategy that favors the IS in model predictions and limits false system activations.

Test results show targeting class balance while training the decoder may provide better performance in asynchronous BCI compared to a perfect class balance ($\tilde{r}^{(IS)} = 1$). The algorithm successfully reaches the desired balance between classes in databases with low number of states. For a large number of states, the class balance tends to the desired balance, but sometime fails to reach it. Some states may be missing for multiple update iterations (because the experimenter was focusing on other states), which makes it impossible for the algorithm to balance missing states until they are trained again. The algorithm reduces the FPR and the NL of the IS. It also slightly reduces the latency. All of these improvements are made with a minimal loss in accuracy. A $\tilde{r}^{(IS)}$ between 1.5 to 2.5 seems to be a good trade-off to reduce

the FPR with minor degradation of the $BACC$. For example, $\tilde{r}^{(IS)} = 2$ degrades by 3.68% to 10.51% the $BACC$ and decreases the FPR by 55.36% to 70.08% in comparison with $\tilde{r}^{(IS)} = 1$. It also allows to have a model with a lower latency. Note that the optimal ratio IS/AS can vary from one user to another.

In this study, the proposed class balancing strategy was integrated into the HMM-RSWNPLS, but it may be adapted for any online training algorithm to train model in an incremental way. Additionally, the proposed algorithm could be tested on other signals recorded with other modality (e.g. EEG or intracortical EEG).

Model performance was evaluated using multiple criteria. To the best of our knowledge, there is no consensus in the literature about which criterion to use to evaluate the latency and the noise level of a prediction probability. Other criteria could be explored to evaluate the latency and the noise level. Multiple criteria analysis complicates decision-making. To address this issue, future studies will consider a multi-criteria decision analysis to facilitate the selection of the most appropriate hyperparameter value.

ACKNOWLEDGMENTS

We would like to thank all the members of the Brain Computer Interface project, composed of the multidisciplinary technical and clinical teams at Clinatéc (CEA-LETI and University Grenoble Alpes), as well as the NeuroRestore team for conducting the experiments with the participants. This work has been supported by the European Commission (EIC 2021-TransitionChallenges-01-01 ReverseParalysis 101057450), Fonds Clinatéc and its sponsors Malakoff Humanis, Covéa and Klésia (BCI4GRASP) and Institut Carnot Leti.

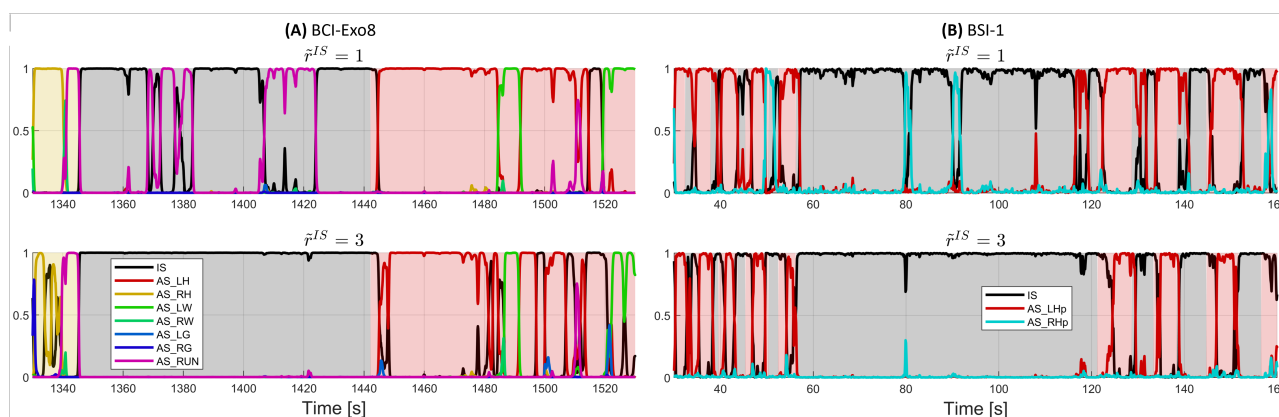


Figure 2: Example of predictions from BCI-Exo8 (A) and BSI-1 (B). Background color show the desired states and the curves represent the predicted probability of each state. Top plots provide results for $\tilde{r}^{IS} = 1$ and bottom plots provide results for $\tilde{r}^{IS} = 3$.

REFERENCES

- [1] Moly A et al. An adaptive closed-loop ECoG decoder for long-term and stable bimanual control of an exoskeleton by a tetraplegic. *Journal of Neural Engineering*. 2022;19(2):026021.
- [2] Lorach H et al. Walking naturally after spinal cord injury using a brain–spine interface. *Nature*. 2023;618(7963):126–133.
- [3] Patrick-Krueger KM, Burkhart I, Contreras-Vidal JL. The state of clinical trials of implantable brain–computer interfaces. *Nature Reviews Bioengineering*. 2024:1–18.
- [4] Souriau R, Martel F, Aksenova T. Recursive sample weighted – N-way partial least squares for brain-computer interface decoder online learning with class imbalance. *Biomedical Signal Processing and Control*. 2026;113:109025.
- [5] Zhang L, Li C, Zhang R, Sun Q. Online semi-supervised learning for motor imagery EEG classification. *Computers in Biology and Medicine*. 2023;165.
- [6] Krell M, Wilshusen N, Seeland A, Kim S. Classifier transfer with data selection strategies for online support vector machine classification with class imbalance. *Journal of Neural Engineering*. 2017;14(2):025003.
- [7] Tibrewal N, Leeuwis N, Alimardani M. Classification of motor imagery EEG using deep learning increases performance in inefficient BCI users. *PLOS ONE*. 2022;17(7):e0268880.
- [8] Peksa J, Mamchur D. State-of-the-Art on Brain-Computer Interface Technology. *Sensors*. 2023;23(13):6001.
- [9] Haotian X, Anmin G, Jiangong L, Fan W, Peng D, Yunfa F. Online adaptive classification system for brain–computer interface based on error-related potentials and neurofeedback. *Biomedical Signal Processing and Control*. 2023;82:104554.
- [10] Ferdinand Q et al. ECoG-Based Movement Classification and Limbs 3D Translation Prediction : A Deep Learning Study. In: 2025 International Joint Conference on Neural Networks (IJCNN). Jun. 2025, 1–10.
- [11] Souriau R et al. Hierarchical Hidden Markov Model for Online Decoding in Brain-Computer Interface. In: 2024 32nd European Signal Processing Conference (EU-SIPCO). IEEE: Lyon, France, Aug. 2024, 1466–1470.
- [12] Zhou Y et al. Shared Three-Dimensional Robotic Arm Control Based on Asynchronous BCI and Computer Vision. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2023;31:3163–3175.
- [13] Yu Y, Liu Y, Jiang J, Yin E, Zhou Z, Hu D. An Asynchronous Control Paradigm Based on Sequential Motor Imagery and Its Application in Wheelchair Navigation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2018;26(12):2367–2375.
- [14] Zhang R et al. A BCI-Based Environmental Control System for Patients With Severe Spinal Cord Injuries. *IEEE Transactions on Biomedical Engineering*. 2017;64(8):1959–1971.
- [15] Xuan H, Yang B, Li X. Exploring the impact of temperature scaling in softmax for classification and adversarial robustness. 2025. arXiv: 2502.20604 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2502.20604>
- [16] Mestais CS, Charvet G, Sauter-Starace F, Foerster M, Ratel D, Benabid AL. WIMAGINE: Wireless 64-Channel ECoG Recording Implant for Long Term Clinical Applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2015;23(1):10–21.
- [17] Ferdinand Q et al. Ecog-based movement classification and limbs 3d translation prediction : A deep learning study. 2025 International Joint Conference on Neural Networks (IJCNN). 2025:1–10.

Combining Virtual Reality Feedback and Transcranial Alternating Current Stimulation to Enhance Restorative Brain-Computer Interfaces

Joana Saraiva^{1,2}, Daniela Esteves³, Christel Bour⁴, Sofia Fernandes^{1,2,‡}, Athanasios Vourvopoulos^{3,‡}

¹Department of Physics, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

²Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

³Institute for Systems and Robotics (ISR-Lisboa), Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

⁴Grenoble INP-UGA (Grenoble Institute of Technology - Grenoble Alpes University), Grenoble, France

‡These authors contributed equally to this work and are joint senior authors.

E-mail: fc58701@alunos.ciencias.ulisboa.pt

ABSTRACT: In stroke rehabilitation, brain–computer interfaces (BCIs) enable neurofeedback-based regulation of sensorimotor rhythms, promoting motor-related cortical reorganization. Integrating virtual reality (VR) into BCI paradigms enhances this process through immersive, multisensory feedback that strengthens sensorimotor engagement. In parallel, transcranial alternating current stimulation (tACS) may further modulate motor imagery (MI)–related oscillatory activity. Together, these approaches offer a multimodal framework to optimize neurorehabilitation outcomes. This study investigated the synergistic effects of a MI-based VR-BCI paradigm and tACS applied at the individual alpha frequency (IAF). Twelve healthy participants performed bimanual MI tasks in VR and non-VR conditions before and after sham or active tACS. Alpha event-related desynchronization (ERD) served as the primary outcome. VR elicited stronger contralateral and more widespread ERD, including parietal and occipital regions. tACS effects were subtle and, in VR, sometimes attenuated ERD.

INTRODUCTION

Neurological disorders rank as the primary driver of disability worldwide and frequently result in motor impairments that profoundly affect individuals' independence and quality of life. Since 1990, the burden of neurological disorders, measured in disability-adjusted life years (DALYs), has risen by 18% [1].

Conventional therapy strategies, such as physical therapy, often require residual voluntary motor function and frequent supervised sessions, which many severely impaired patients cannot achieve [2]. These limitations have motivated the development of neurotechnology-based interventions as complementary approaches that directly target the central nervous system.

Non-invasive Brain-computer interfaces (BCIs), which enable the control of external devices in real-time us-

ing neural activity, usually recorded through electroencephalography (EEG), allow users to control virtual reality (VR) avatars or robotic devices through motor imagery (MI), forming closed-loop systems that strengthen key motor pathways [3]. MI is the mental simulation of movement without muscle activation, engaging sensory and cognitive processes. Both MI and motor execution (ME) are known to induce event-related desynchronization (ERD), which is a reduction in EEG power in the alpha (8–12 Hz) and beta (12–30 Hz) bands, due to their desynchronization over sensorimotor areas [4]. This brain rhythm modulation that results from BCI training has shown to correlate with motor recovery [5, 6].

VR is increasingly integrated with BCIs to enhance embodiment and user engagement by delivering immersive, task-oriented feedback environments [7]. VR-BCI systems have been shown to strengthen cortical activation, with studies reporting robust contralateral event-related desynchronizations (ERDs) comparable to those observed during motor execution (ME) [8].

To further amplify these effects, VR-BCI training has been combined with non-invasive brain stimulation (NIBS). Repetitive transcranial magnetic stimulation (rTMS) enhances cortical excitability and strengthens ERD during VR-BCI training, though benefits appear to be short-term [6]. Transcranial electrical stimulation (tES) techniques, including transcranial direct current stimulation (tDCS) and transcranial alternating current stimulation (tACS), modulate neural activity using weak currents.

tDCS combined with VR feedback has yielded greater and clinically significant motor function improvements compared with conventional physical therapy [9]. Furthermore, MI-based VR-BCI paradigms coupled with tDCS have shown potential to modulate sensorimotor rhythms and support motor recovery, although existing evidence remains preliminary [10, 11]. Reported effects include modulations in alpha and beta rhythms and gains

in motor performance. Nonetheless, these findings are tempered by important limitations, such as small sample sizes, lack of control conditions and the use of bilateral montages that may attenuate interhemispheric differences [10, 11]. Overall, the effectiveness of tDCS appears to depend strongly on stimulation parameters and spatial targeting, underscoring the need for more optimized and personalized approaches.

tACS is particularly suited for personalized stimulation, because it can target intrinsic oscillations such as the individual alpha frequency (IAF), defined as the frequency within the alpha band at which EEG power is maximized, and which is strongly correlated with task preparation and engagement [6, 12]. tACS has been shown to enhance MI performance and hybrid BCI classification (combining MI and steady-state visual evoked potentials) [13], suggesting a potential to amplify contralateral ERDs induced by VR-BCI.

Given the limited exploration of this combination, the present study proposes a hybrid approach integrating VR-BCI with tACS applied at the participants' IAF measured during resting-state EEG, aiming to determine whether this personalized stimulation enhances MI-related BCI features beyond VR-BCI alone, thereby supporting more effective neurorehabilitation.

MATERIALS AND METHODS

Participant recruitment: Twelve healthy, right-handed participants (six females; mean age 21.08 ± 1.16 years) were recruited. None had prior experience with BCI systems and three (S07, S10 and S12) had previously participated in non-invasive tES studies without adverse effects. Written informed consent was obtained in accordance with the 1964 Declaration of Helsinki. This study was approved by the Ethics Committee of Instituto Superior Técnico (CE-IST) with Ref. no. 22/2024.

EEG acquisition: EEG was recorded using a 32 active electrode's amplifier (Liveamp32, Brain Products GmbH, Gilching, Germany), positioned according to the international 10–20 system. Signals were acquired through the BrainVision Recorder at a sampling rate of 250 Hz, with electrode impedance maintained below 10 k Ω (Figure 1(A)(i)).

VR setup: The NeuRow environment [2] is a first-person perspective upper-limb MI training paradigm that renders two virtual hands performing left- and right-sided rowing movements [6, 7] (Figure 1(B)(ii)). The visual feedback for this study was presented on a computer screen.

Electric field (EF) modeling: EF simulations were performed using the example head model in *SimNIBS 4.1.0*, comprising six tissue compartments (scalp, skull, cortical gray matter (GM), cortical white matter (WM), cerebrospinal fluid (CSF) and eyes) represented by a finite element mesh with approximately 4.76×10^6 tetrahedral elements, with standard conductivity values provided by *SimNIBS* [14].

Table 1: Comparison of the EF magnitude (*magnE*) and focality between the ideal (C3, C1, C2 and CP2) and selected (CP3, C1, C2 and CP4) montages. Percentile values refer to the maximum EF intensity in the GM.

Montage	magnE (V/m)			Focality ($\times 10^4 \text{ mm}^3$)	
	99.9%	99.0%	95.0%	75%	50%
Ideal	0.422	0.322	0.241	0.809	6.020
Selected	0.423	0.336	0.258	1.080	8.230
Δ (%)	0.237	4.348	7.054	33.498	36.711

Four circular electrodes ($\pi \text{ cm}^2$) were modeled with a 1 mm electrode silicone rubber pad ($\sigma = 29.4 \text{ S/m}$) and a 2.5 mm saline-soaked sponge layer ($\sigma = 1.0 \text{ S/m}$). The maximum total injected current was set to 4 mA (maximum 1 mA per electrode), with up to four active electrodes.

The primary motor cortex (C3/C4) was defined as the stimulation target using the *SimNIBS* optimization script with coordinates expressed in Montreal Neurological Institute (MNI) space [15]. Since C3/C4 were required for EEG, the ideal montage (anodes: C1 and C3; cathodes: C2 and CP2) was adjusted by replacing C3 with CP3 and, for symmetry, CP2 with CP4, yielding, at targets, comparable EF magnitudes with some loss of focality (Table 1).

Given the low frequencies used in tACS, a quasi-static approximation was assumed, treating tissues as predominantly resistive and the EF as linearly proportional to the injected current. Session-specific EF magnitudes were thus estimated by scaling a reference EF of 0.256 V/m (obtained from *SimNIBS* simulations of the selected montage and defined as the mean within 15 mm radius spherical ROIs centered on left and right M1 [15]) according to the applied current (Table 3). Since EF magnitudes around 0.1–0.2 V/m are suggested to induce neuromodulatory effects [16], sessions reaching at least 0.1 V/m are highlighted.

tACS setup: tACS was delivered using the *StarStim* device (Neuroelectronics, Barcelona, Spain) with four channels (anodes - C1, CP3; cathodes - C2, CP4) through saline-soaked sponges ($\pi \text{ cm}^2$). The reference was placed on the earlobe (Figure 1(A)(ii)).

Stimulation frequency was set to each participant's IAF, defined as the average IAF value from C3, Cz and C4 during the 2-minute eyes-open resting-state EEG, consistent with the task design. Intensity was determined through a threshold test starting at 300 μA and increasing in 100 μA steps until sensations were reported. The applied intensity corresponded to the level immediately below this threshold.

Protocol: Each session lasted approximately one hour and participants completed the VR and No VR sessions in randomized order, separated by at least 48 hours to ensure safety with tACS exposure.

Sessions began with a 4-minute resting-state EEG acquisition period to determine the participant's IAF, followed by the threshold test described in Section *tACS montage and setup*. Three MI tasks were performed per session, each with 40 randomized trials per side (left or right). Each trial comprised a 2-second baseline, 1-

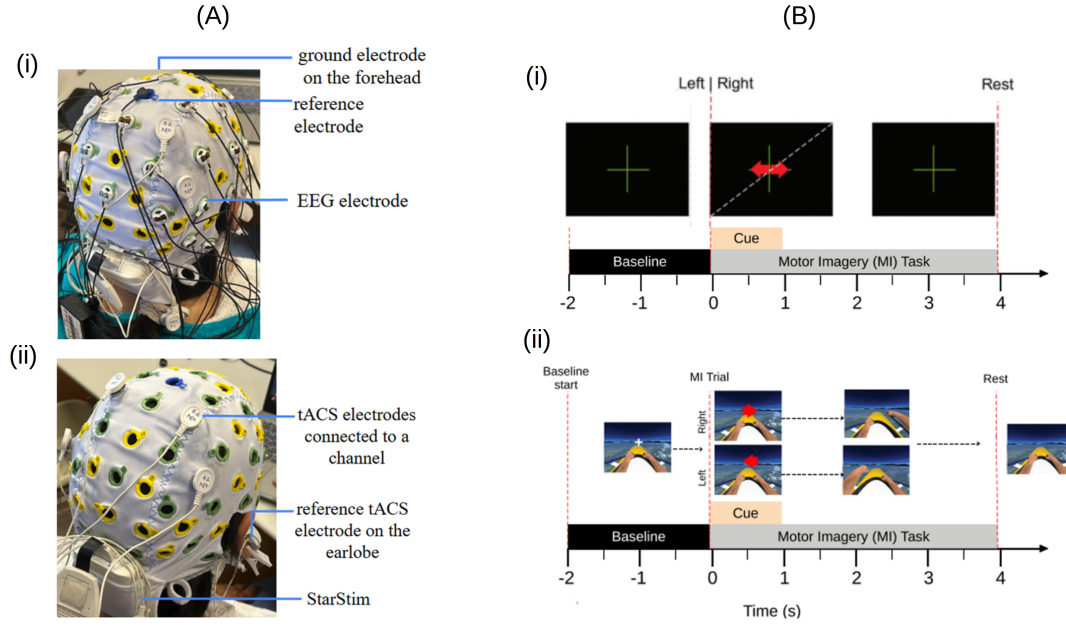


Figure 1: (A) Experimental setup illustrating (i) the EEG recording configuration with 32 active electrodes and (ii) the tACS setup. (B) Structure of each MI trial for tasks 1, 2 and 3: Each trial begins with a 2-second baseline, followed by a 1-second cue indicating left- or right-side MI, resulting in a 4-second MI trial. (i) No-VR condition, (ii) VR condition.

second cue and 4-second MI execution period, respectively (Figure 1(B)).

Task 1 served as training prior to any stimulation. Between tasks, participants received sham and active tACS for 15 minutes, with either eyes closed or fixated on a cross, in a blinded manner, such that Task 2 followed sham stimulation and Task 3 followed active tACS. Participants were informed they would undergo two stimulation periods without being told which was active tACS or sham. However, they were made aware of potential stimulation sensations, as this was necessary for the threshold test to determine whether they perceived it.

Participant questionnaires: Before data acquisition, participants completed an *Exclusion Criteria* questionnaire to screen for tACS contraindications [17] and the *Vividness of Movement Imagery* questionnaire (VMIQ), in a 7-point Likert scale, to assess their ability to generate vivid mental images during various MI tasks [18]. After each stimulation period, an *Adverse Effects* questionnaire was administered [19]. At the end of the VR session, an *Embodiment* questionnaire evaluated body ownership, self-location and agency [20].

Pre-processing: EEG data were pre-processed in MATLAB R2023b (The MathWorks, MA, United States) using EEGLAB v2025.0.0. Signals were downsampled to 128 Hz and band-pass filtered between 1–40 Hz. Independent Component Analysis (ICA) was applied to remove remaining artifactual components from the EEG data [7]. Components classified by *ICLabel* as ‘Muscle’ or ‘Eye’ with $\geq 90\%$ probability were rejected [7]. Next, visual inspection and the Artifact Subspace Reconstruction (ASR) method were used. Finally, data were re-referenced to the common average and epoched from

-2 to +4 seconds around cue onset.

ERD power computation: The Event-Related Spectral Perturbation (ERSP) was estimated from the time and frequency domains using a 3-cycle Morlet wavelet with a Hanning taper. Values were expressed in decibels (dB) across 0.5 Hz bins relative to the mean baseline power and converted to alpha ERD using Equation 1.

$$ERD(\%) = (10^{\frac{ERSP_{dB}}{10}} - 1) \times 100 \quad (1)$$

Lateralization index (LI) calculation: LI was calculated from alpha-band ERD over electrodes C3 and C4, as shown in Equation 2 [7].

$$LI_{ERD} = \frac{(ERD_{C3(L)} - ERD_{C4(L)}) + (ERD_{C4(R)} - ERD_{C3(R)})}{2} \quad (2)$$

A positive LI suggests stronger desynchronization in the hemisphere contralateral to the imagined movement: C4 for left (L) and C3 for right (R) MI [7].

Statistical analysis: Normality was evaluated using the Shapiro-Wilk test and was satisfied for all variables. Task-related differences in ERD power and LI were assessed using a one-way repeated-measures ANOVA ($\alpha = 0.05$) within each session with Bonferroni correction.

Linear mixed-effects modeling: Linear mixed-effects (LME) models were used to evaluate the influence of experimental factors on ERD power, with subject included as a random effect to account for inter-individual variability. Candidate models incorporated fixed effects for feedback, stimulation condition and intensity, electrode (C3/C4), MI side and VMIQ score, including relevant interactions. Models were compared using the Bayesian Information Criterion (BIC) and the best-fitting model included main effects of feedback modality and stimulation condition, as well as an interaction between electrode location and MI side.

RESULTS

In terms of ERD, most subjects are able to induce strong desynchronization across sessions and tasks (Figure 2), with generally stronger at contralateral electrodes (C3/C4 at the population level). For left MI, in No VR condition, sham weakened ERD and tACS restored it. A significant main effect of task was observed at contralateral C4 ($p = 0.042$), although Bonferroni-corrected post-hoc comparisons were not significant (Task 1 vs. Task 2: $p = 0.091$; Task 1 vs. Task 3: $p = 1.000$; Task 2 vs. Task 3: $p = 0.221$). In VR, sham strengthened contralateral ERD, whereas tACS hindered it at both electrodes and no significant task effects were detected. For right MI, effects were modest. In No VR, sham and tACS slightly enhanced ipsilateral ERD, with no significant task effects. In VR, sham weakened ipsilateral ERD and tACS slightly improved it, while contralateral ERD remained relatively stable across tasks, with no significant differences.

In terms of lateralization, LI values were generally positive, particularly in VR, indicating contralateral dominance (Figure 3). In No VR, median LI increased from Task 2 to Task 3, suggesting a possible tACS-related effect, with a trend toward a main effect of task ($p = 0.052$), although not significant. In VR, median LI values remained similar across tasks and no main effect of task was observed ($p = 0.734$). Overall, LI distributions were comparable between sessions and tasks.

In terms of the linear modeling, Table 2 reports significant fixed effects from the best-fitting LME model. Compared to C3, ERD at C4 was significantly more negative, indicating stronger desynchronization. Sham stimulation was associated with less negative ERD, reflecting reduced desynchronization. A significant electrode $\times$ MI side interaction showed that ERD at C4 during right MI was attenuated relative to main effects.

Finally, in terms of the neuromodulatory effects, Table 3 indicates that few sessions reached EF magnitudes (0.1 V/m) commonly associated with neuromodulation [16]. Such values were observed in both sessions for S03, S04 and S11, and only in the VR session for S05, S07 and S08. The highest EF was 0.154 V/m (S07, VR), while the lowest was 0.051 V/m , observed in multiple sessions.

Table 2: Statistically significant fixed effects from the best-fitting LME model predicting ERD.

	Estimate	p -value
(Intercept)	-17.203	0.009
C4	-7.251	0.020
Sham	6.737	0.013
C4*Right_MI	19.052	< 0.001

DISCUSSION

To our knowledge, this is the first study to examine the synergistic effects of personalized-frequency tACS and a VR-BCI paradigm on the induction of ERD during MI. We aimed to determine whether adding tACS to VR-based MI enhances BCI outcomes during MI tasks.

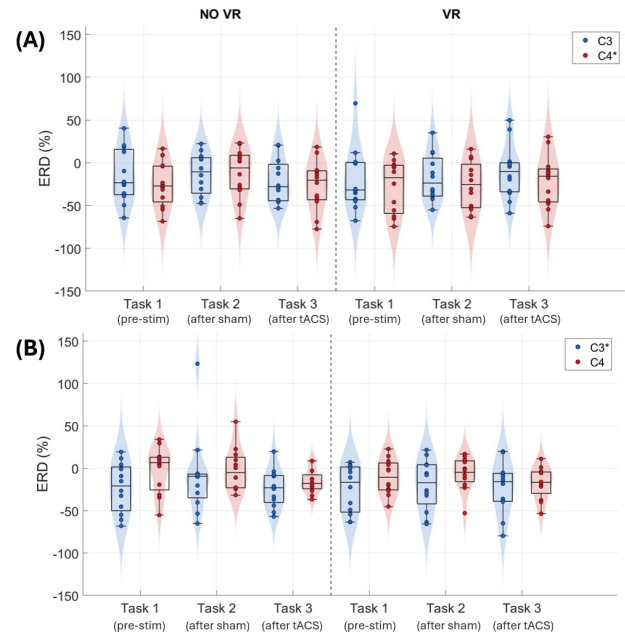


Figure 2: Boxplots showing ERD values at electrodes C3 and C4 during left (A) and right MI (B). Each boxplot presents both sessions (No VR and VR), each with three boxes representing the different tasks. The asterisk (*) indicates the contralateral electrode relative to the imagined movement.

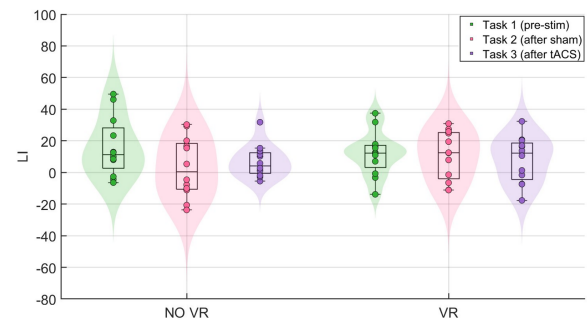


Figure 3: Boxplot of LI values comparing No VR and VR sessions, each with the three tasks.

Across analyses, substantial inter-subject variability was evident, consistent with previous literature [21–23], and reflected in the ERD distributions (Figures 2 and 3). Desynchronization was not restricted to primary motor areas but extended to parietal and occipital regions, especially in VR. This widespread pattern likely reflects increased reliance on visuospatial processing, possibly due to the relative complexity and unfamiliarity of the rowing task [24]. In line with prior findings [25], MI engages motor and premotor regions but also recruits visual areas, particularly when motor representations are less established. Stronger contralateral ERD was observed during left MI, potentially reflecting that, although imagining the rowing movement was generally challenging, imagery involving the non-dominant hand (the left hand, given that all participants were right-handed) may have been perceived as even more demanding. Instead of reducing engagement, this increased difficulty may have prompted greater effort, leading to stronger contralateral

Table 3: IAF (Hz) and tACS (μA) and EF magnitudes (V/m) for each subject under both VR and No VR sessions. Highlighted cells indicate sessions in which the delivered EF exceeded $0.1 V/m$.

	NO VR			VR		
	IAF (Hz)	tACS magn (μA)	EF (V/m)	IAF (Hz)	tACS magn (μA)	EF (V/m)
S01	10.260	300	0.077	9.800	300	0.077
S02	9.490	200	0.051	9.540	200	0.051
S03	10.160	500	0.128	10.160	400	0.102
S04	10.310	400	0.102	10.280	500	0.128
S05	10.180	200	0.051	10.290	400	0.102
S06	9.460	200	0.051	9.570	300	0.077
S07	9.330	300	0.077	8.960	600	0.154
S08	10.630	300	0.077	10.510	400	0.102
S09	10.360	200	0.051	9.930	200	0.051
S10	10.050	200	0.051	10.380	200	0.051
S11	9.820	500	0.128	9.690	400	0.102
S12	9.540	300	0.077	9.670	300	0.077

desynchronization regardless of spatial focality.

Consistent with the literature ([7]), ERD was generally stronger at contralateral electrodes, indicating preserved physiological lateralization (see Figures 2 and 3). Moreover, VR sessions tended to elicit stronger ERDs consistent with previous studies [8, 26] that show that MI together with motor-observation (MO) lead to greater brain activation. In the No VR sessions, tACS appeared to enhance desynchronization compared with sham, particularly during left MI, where a significant task effect at C4 was observed ($p = 0.042$), although post-hoc comparisons were not significant. In contrast, no significant ERD effects were found in VR, suggesting that VR feedback may attenuate or mask tACS-related modulation. A similar pattern was observed for LI (Figure 3): a trend toward increased contralateral dominance after tACS emerged in No VR, but not in VR. This visual trend is consistent with the statistical analysis, which revealed a trend toward a main effect of task on LI in the No VR session that, however, did not reach the predefined level of significance ($p = 0.052$). Despite limited statistical significance here, prior studies show that tES, particularly tACS, can enhance BCI classification accuracy [13] and strengthen ERDs [27]. Our results suggest that while tACS may facilitate the ERD during MI under conventional conditions, its interaction with VR feedback may attenuate or reverse these effects, warranting further investigation.

The LME analysis confirmed stronger ERD at C4, suggesting that left MI elicited stronger ERD than right MI) and revealed a significant interaction between electrode and MI side, consistent with contralateral organization. Sham stimulation was associated with weaker ERD, supporting evidence that sham is not physiologically inert and may influence cortical activity over time through expectancy or sensory factors.

Estimated EF magnitudes (presented in Table 3) varied substantially, and only a minority of sessions exceeded commonly cited neuromodulatory thresholds

[16]. Higher EF was not consistently associated with stronger ERD. Notably, several VR sessions with higher EF coincided with those in which pronounced counter-productive effects of tACS on ERD were observed. Although this co-occurrence does not imply causality, it supports the possibility that higher tACS intensities may interact unfavorably with VR feedback. While no universal EF threshold exists and individual sensitivity varies, the need to preserve blinding led to reduced stimulation intensities in some participants, likely limiting neuromodulatory impact and increasing variability.

CONCLUSION

This study evaluated whether adding personalized-frequency tACS to a VR-based MI-BCI paradigm enhances ERD modulation. VR-BCI elicited stronger and more widespread contralateral ERDs compared to non-VR MI. In contrast, tACS effects were modest and primarily observable in the absence of VR, with variable or attenuated responses during immersive feedback. EF magnitude was not clearly related to ERD strength.

Overall, while tACS influenced MI-related dynamics, its added value within VR-BCI frameworks remains uncertain. Larger studies with optimized stimulation protocols are required to clarify its role in hybrid neurorehabilitation strategies.

ACKNOWLEDGMENTS

This research was supported by Bolsas Novos Talentos of the Fundação Calouste Gulbenkian and the Fundação para a Ciência e Tecnologia (FCT, <https://ror.org/00snfq58>) through the LARSyS grant (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020), the NOISyS project (DOI: 10.54499/2022.02283.PTDC) and the IBEB's funding UID/00645/2025 (<https://doi.org/10.54499/UID/00645/2025>). For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

REFERENCES

- [1] Feigin VL et al. World stroke organization (wso): Global stroke fact sheet 2022. *International Journal of Stroke*. 2022;17(1):18–29.
- [2] Vourvopoulos A, Jorge C, Abreu R, Figueiredo P, Fernandes JC, Badia S Bermudez i. Efficacy and brain imaging correlates of an immersive motor imagery be-driven vr system for upper limb motor rehabilitation: A clinical case report. *Frontiers in human neuroscience*. 2019;13:244.
- [3] Cervera MA et al. Brain-computer interfaces for post-stroke motor rehabilitation: A meta-analysis. *Annals*

of Clinical and Translational Neurology. 2018;5(5):651–663.

[4] Pfurtscheller G, Silva F Lopes da. Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.

[5] Ray AM, Figueiredo TD, López-Larraz E, Birbaumer N, Ramos-Murguialday A. Brain oscillatory activity as a biomarker of motor recovery in chronic stroke. *Human brain mapping*. 2020;41(5):1296–1308.

[6] Afonso M, Sánchez-Cuesta F, González-Zamorano Y, Romero JP, Vourvopoulos A. Investigating the synergistic neuromodulation effect of bilateral rtms and vr brain-computer interfaces training in chronic stroke patients. *Journal of Neural Engineering*. 2024;21(5).

[7] Vourvopoulos A, Fleury M, Blanco-Mora DA, Fernandes JC, Figueiredo P, Badia S Bermúdez i. Brain imaging and clinical outcome of embodied vr-bci training in chronic stroke patients: A longitudinal pilot study. *Research in Biomedical Engineering and Technology*. 2024;11(4):193–209.

[8] Batista D, Caetano G, Fleury M, Figueiredo P, Vourvopoulos A. Effect of head-mounted virtual reality and vibrotactile feedback in erd during motor imagery brain-computer interface training. *Research in Biomedical Engineering and Technology*. 2024;11(1–2):11–20.

[9] Llorens R et al. Effectiveness of a combined transcranial direct current stimulation and virtual reality-based intervention on upper limb function in chronic individuals post-stroke with persistent severe hemiparesis: A randomized controlled trial. *Journal of NeuroEngineering and Rehabilitation*. 2021;18(108).

[10] Lima JPS, Silva LA, Delisle-Rodriguez D, Cardoso VF, Nakamura-Palacios EM, Bastos-Filho TF. Unraveling transformative effects after tdc and bci intervention in chronic post-stroke patient rehabilitation—an alternative treatment design study. *Sensors*. 2023;23(23).

[11] Blanco Mora DA et al. Toward methodologies for motor imagery enhancement: A tdc-bci study. *Research in Biomedical Engineering and Technology*. 2024;11(3):110–124.

[12] Wischniewski M, Alekseichuk I, Opitz A. Neurocognitive, physiological, and biophysical effects of transcranial alternating current stimulation. *Trends in Cognitive Sciences*. 2023;27(2):189–205.

[13] Li Z, Zhang R, Li W, Li M, Chen X, Cui H. Enhancement of hybrid bci system performance based on motor imagery and ssvep by transcranial alternating current stimulation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024;32:3222–3230.

[14] Thielscher A, Antunes A, Saturnino GB. Field modeling for transcranial magnetic stimulation: A useful tool to understand the physiological effects of tms? In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Milan, Italy, 2015, 222–225.

[15] Mayka MA, Corcos DM, Leurgans SE, Vaillancourt DE. Three-dimensional locations and boundaries of motor and premotor cortices as defined by functional brain imaging: A meta-analysis. *NeuroImage*. 2006;31(4):1453–1474.

[16] Kasten FH, Duecker K, Maack MC, Meiser A, Herrmann CS. Integrating electric field modeling and neuroimaging to explain inter-individual variability of tacs effects. *Nature Communications*. 2019;10(5427).

[17] Antal A et al. Low intensity transcranial electric stimulation: Safety, ethical, legal regulatory and application guidelines. *Clinical Neurophysiology*. 2017;128(9):1774–1809.

[18] Roberts R, Callow N, Hardy L, Markland D, Bringer J. Movement imagery ability: Development and assessment of a revised version of the vividness of movement imagery questionnaire. *Journal of Sport and Exercise Psychology*. 2008;30(2):200–221.

[19] Brunoni AR, Amadera J, Berbel B, Volz MS, Rizzorio BG, Fregni F. A systematic review on reporting and assessment of adverse effects associated with transcranial direct current stimulation. *International Journal of Neuropsychopharmacology*. 2011;14(8):1133–1145.

[20] Peck TC, Gonzalez-Franco M. Avatar embodiment. a standardized questionnaire. *Frontiers in Virtual Reality*. 2021;1.

[21] Rimbart S, Trocellier D, Lotte F. Is event-related desynchronization variability correlated with bci performance? In: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine). Rome, Italy, 2022, 163–168.

[22] Kwon M, Cho H, Won K, Ahn M, Jun SC. Event-related desynchronization (erd) may not be correlated with motor imagery bci performance. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Miyazaki, Japan, 2018, 1133–1137.

[23] Wriessnegger SC, Müller-Putz GR, Brunner C, Sburlea AI. Inter- and intra-individual variability in brain oscillations during sports motor imagery. *Frontiers in Human Neuroscience*. 2020;14.

[24] Olsson CJ, Jonsson B, Larsson A, Nyberg L. Motor representations and practice affect brain systems underlying imagery: An fmri study of internal imagery in novices and active high jumpers. *The Open Neuroimaging Journal*. 2008;2:5–13.

[25] Burianová H et al. Multimodal functional imaging of motor imagery using a novel paradigm. *NeuroImage*. 2013;71:50–58.

[26] Nunes JD et al. Brain activation by a vr-based motor imagery and observation task: An fmri study. *PLoS ONE*. 2023;18(9):e0291528.

[27] Kasten FH, Herrmann CS. Transcranial alternating current stimulation (tacs) enhances mental rotation performance during and after stimulation. *Frontiers in Human Neuroscience*. 2017;11.

Efficient Decoding of the Code-Modulated Motion Visual Evoked Potential (c-MVEP) for BCIs

H.A. Scheppink^{1,2}, S. Bialonski^{2,3,4}, M. Tangermann⁵, J. Thielen⁵, I. Volosyak^{1,2}

¹Faculty of Technology and Bionics, Rhine-Waal University of Applied Sciences, Kleve, Germany

²Department of Informatics and Data Science, Graduate School for Applied Research in North Rhine-Westphalia, Bochum, Germany

³Department of Medical Engineering and Technomathematics

⁴Institute for Data-Driven Technologies

FH Aachen University of Applied Sciences, Jülich, Germany

⁵Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

E-mail: ivan.volosyak@hochschule-rhein-waal.de

ABSTRACT: The code-modulated motion visual evoked potential (c-MVEP) paradigm visually stimulates targets using motion in pseudo-random intervals. Due to its code-modulated nature, c-MVEP shares aspects of the code-modulated visual evoked potential (c-VEP). This study investigated whether the c-MVEP can be decoded using similar classification approaches as used for c-VEP. Several canonical correlation analysis (CCA)-based methods were evaluated, with increasingly constrained assumptions about the neural data, to reduce model complexity. These included ensemble and global spatial filters, as well as template estimation by averaging across trials, cycles, classes, or stimulus events. Consistent with findings in c-VEP, we found that the event-based approaches, with events defined by the rising or falling edge of the stimulus sequence, achieved a higher decoding accuracy with fewer training data for calibration than the commonly used trial-based approach. This study shows the potential for decoding the c-MVEP with high accuracy, without requiring extensive training.

INTRODUCTION

A brain-computer interface (BCI) uses brain signals to enable users to control a BCI application without requiring muscular activity [1]. The fastest and highest performance non-invasive BCIs for communication use steady-state visual evoked potential (SSVEP) or code-modulated visual evoked potential (c-VEP), usually based on high-contrast flickering, which is visually and mentally fatiguing, and causes eye-strain during longer use [2, 3].

To address this, the steady-state motion visual evoked potential (SSMVEP) was proposed [2, 4], using frequency-based movement of the stimuli, such as rotation or radial zooming. Chai et al. showed users may prefer SSMVEP over the SSVEP flickering [5].

However, the decoding accuracy and information transfer rate (ITR) of SSMVEP are not on par with SSVEP and

c-VEP [6]. Furthermore, while c-VEP is a robust and reliable paradigm for all users [7], SSMVEP was less reliable, as not all users can reach sufficient performance [6]. These findings motivated us to integrate the comfort of SSMVEP with the robustness of c-VEP, leading us to propose the code-modulated motion visual evoked potential (c-MVEP) [8]. Specifically, the novel c-MVEP paradigm uses pseudo-random stimulation sequences with optimal autocorrelation properties from c-VEP, as well as the comfortable smooth motion characteristics of SSMVEP. In this study, we investigate whether the c-MVEP can be decoded similarly to the c-VEP. We evaluate (1) an ensemble canonical correlation (CCA) approach with class-specific spatial filters [9] and (2) a CCA restricted to one global spatial filter. For both, (a) templates are estimated by trial averaging. For (2), we additionally compared averaging across (b) stimulus sequences and (c) classes, exploiting the property of shifted m-sequences [3].

To further reduce calibration requirements, we use the (3) linear superposition model and reconvolution-CCA approach proposed by Thielen et al. [10], which represents sequence responses as the sum of event-related responses. This allows averaging over stimulus events, such as single flashes. Since each sequence consists of many events, reducing model complexity by event-level averaging decreases the number of parameters, while increasing the effective number of training samples per recording. As a result, the same recording delivers more training samples, which allows for a shorter calibration.

This study investigates the applicability of the assumptions made in c-VEP for c-MVEP. Specifically, we work towards an efficient decoding approach, with only a few free parameters and a small amount of training data to achieve high decoding accuracies.

MATERIALS AND METHODS

Data: We used a subset of previously recorded data

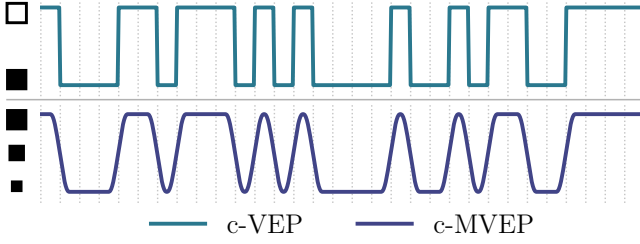


Figure 1: Stimulation waveforms. The stimulation waveforms for c-VEP (top) and c-MVEP (bottom). Both use the same 31-bit m-sequence, with c-MVEP using a smoothing of the edges for continuous stimulus size modulation (illustrated by three zoom levels), whereas for c-VEP it translates to binary black-white flicker (illustrated by two flicker states). The dotted vertical lines represent each bit.

from 19 participants (14 males, 5 females, 23.7 ± 3.9 years, range 19-35) with normal or corrected-to-normal vision, and four had previous experience with a BCI system [8]. The study was designed and conducted in accordance with the Helsinki Declaration and was approved by the ethical committee of the medical faculty of Duisburg-Essen University (24-11957-BO).

The EEG data was acquired at a sampling frequency F_s of 600 Hz with a g.USBamp amplifier (g.tec, Schiedlberg, Austria) using sixteen gel-based passive scalp electrodes placed according to the international 10-5 system ($P_7, P_3, P_z, P_4, P_8, PO_7, PO_3, PO_z, PO_4, PO_8, O_1, O_z, O_2, O_9, I_z$, and O_{10}), with reference at C_z and ground at AF_z . Impedances were kept below 5 k Ω . A 50 Hz notch filter and 1–40 Hz fourth-order Butterworth band-pass filter were applied to the raw data. No outlier removal was done, all epochs and channels were kept for analysis.

The experiment was conducted on a Dell Precision desktop (Windows 11 Education, Intel Core i9 3.70 GHz, NVIDIA RTX 3070). Visual stimuli were presented on an Acer Nitro XV252Q F LCD screen (1920 $\times$ 1080 pixels) with a refresh rate of 360 Hz.

The original experiment included four conditions: SSVEP, SSMVEP, c-VEP, and c-MVEP. Our analyses use the c-VEP and c-MVEP data only. Furthermore, we use the original calibration data for an offline analysis.

During this calibration session, four black square stimuli were shown on a gray background, each with a visual angle of $4.34^\circ \times 4.34^\circ$, with 7.26° distance between the stimuli, at a viewing distance of 60 cm. The participants fixated on the cued target until the next cue appeared. The cue was presented for 1.5 s, followed by a 1.0 s return to the original stimulus color prior to the onset of flickering or motion. The stimulation in a trial lasted for a total of three stimulus cycles, $3 \times 1.55 = 4.65$ s. Each condition included six four-trial blocks, with each target cued once per block (left to right), leading to 24 trials per condition.

Stimulus Generation: For c-VEP, we used $c_1 = 1000110111010100001001011001111$ as the 31-bit m-sequence for the first target. The second, third, and fourth target c_2, c_3 , and c_4 used a 2-bit, 4-bit, and 7-bit right-circular shifted version of c_1 , respectively. These m-sequences were displayed at a 20 Hz presentation rate to

flash the stimuli between black ('0') and white ('1'), creating a stimulus cycle that is $31/20 = 1.55$ s long.

For c-MVEP, the same m-sequences were used. However, a binary sequence for motion would cause abrupt size changes without apparent motion. Thus, the sequences were made continuous by smoothing the transitions between zeros and ones, while preserving the original timing and duration properties of the m-sequence, similar to the approach used in a code-modulated auditory evoked potential (c-AEP) study [11]. This resulted in pseudo-random sequences with unchanged event onset and duration but smoother transitions, thereby improving the perception of motion.

Smoothing of the edges was implemented using a raised sine function on the sequence upsampled to the refresh rate of the screen, where each bit corresponded to 18 frames. For each rising edge ($0 \rightarrow 1$) and falling edge ($1 \rightarrow 0$), a transition window equal to one bit duration (50 ms) was placed around each transition. Inside a window, the sharp edge was replaced by a smooth sinusoidal function, while outside the sequence remained unchanged (see Fig. 1). The autocorrelation of the smoothed sequences closely resembled that of the original m-sequence, with a value of 1 at zero lag, and approximately $-1/N$ at all other lags, except at lag 1, where it was 0.12. These smoothed m-sequences were used to modulate a radial zooming motion, mapping '1' to the full stimulus size and '0' to half its size, again with a stimulus cycle of 1.55 s.

Classification: Several template-matching classifiers have previously been proposed to predict the attended target from the recorded EEG data, specifically versions of the average-based reference CCA (denoted as aCCA) [3] and reconvolution CCA (denoted as rCCA) [12]. Both classify an unseen single-trial $\mathbf{X} \in \mathbb{R}^{C \times T}$ of C channels and T samples by selecting the template response $\mathbf{t}_i \in \mathbb{R}^T$ that maximizes the Pearson's correlation with the spatially filtered data using spatial filter $\mathbf{w} \in \mathbb{R}^C$:

$$\hat{y} = \arg \max_i \rho(\mathbf{w}^\top \mathbf{X}, \mathbf{t}_i). \quad (1)$$

The methods differ in how they estimate the spatial filter $\mathbf{w}$ and template responses $\mathbf{t}_i$ for each of the N classes:

aCCA: CCA learns two spatial filters $\mathbf{w}$ and $\mathbf{v}$:

$$\arg \max_{\mathbf{w}, \mathbf{v}} \rho(\mathbf{w}^\top \mathbf{S}, \mathbf{v}^\top \mathbf{T}), \quad (2)$$

with the concatenated multichannel calibration data $\mathbf{S}$, and the corresponding stacked templates $\mathbf{T}$.

We evaluated four established modes of aCCA: First, we tested an ensemble approach (**aCCA-Ens**), in which individual spatial filters are determined for each stimulus class, that is, the CCA is trained for each class i , leading to spatial filters $\mathbf{w}_i$ and $\mathbf{v}_i$.

Let $\mathbf{X}_i^{(k)} \in \mathbb{R}^{C \times T}$ be the k -th trial of class i , with $i = 1, \dots, N$ and $k = 1, \dots, K_i$ trials of class i . Then, for each class i , we define the concatenated multichannel calibration data $\mathbf{S}_i = [\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(K_i)}]$ and the corresponding stacked templates $\mathbf{T}_i = [\bar{\mathbf{X}}_i, \dots, \bar{\mathbf{X}}_i]$, where the

class-average across trials is $\bar{\mathbf{X}}_i = \frac{1}{K_i} \sum_{k=1}^{K_i} \mathbf{X}_i^{(k)}$. Then, aCCA-Ens solves Eq. 2 for each class separately, leading to $(\mathbf{w}_i, \mathbf{v}_i) = \arg \max_{\mathbf{w}, \mathbf{v}} \rho(\mathbf{w}^\top \mathbf{S}_i, \mathbf{v}^\top \mathbf{T}_i)$, yielding one spatial filter per class $\mathbf{w}_i$. Classification is realized by $\hat{y} = \arg \max_i \rho(\mathbf{w}_i^\top \mathbf{X}, \mathbf{t}_i)$, with template response $\mathbf{t}_i = \mathbf{w}_i^\top \bar{\mathbf{X}}_i$.

Secondly, we test aCCA with one universal spatial filter, assuming that the classes only differ temporally, but share the same spatial source. This leads to the conventional aCCA that learns one spatial filter and estimates templates by averaging full trials (**aCCA-Trl**). Specifically, it solves Eq. 2 once, with $\mathbf{S} = [\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_N^{(K_N)}]$, and $\mathbf{T} = [\bar{\mathbf{X}}_1, \dots, \bar{\mathbf{X}}_N]$. The obtained universal spatial filter $\mathbf{w}$ is used for classification as in Eq. 1, with $\mathbf{t}_i = \mathbf{w}^\top \bar{\mathbf{X}}_i$.

Thirdly, we test an aCCA that estimates templates by averaging stimulus cycles, rather than trials, where each trial consisted of three cycles. We assume that the response across cycles is the same, such that averaging can be done over the cycles (**aCCA-Cyc**). Let each trial k consist of D cycles, and denote the d -th cycle of trial k of class i as $\mathbf{X}_i^{(k,d)} \in \mathbb{R}^{C \times T_c}$ with T_c samples in a cycle. Instead of averaging over full trials, we average over all cycles of a class $\bar{\mathbf{X}}_i = \frac{1}{K_i D} \sum_{k=1}^{K_i} \sum_{d=1}^D \mathbf{X}_i^{(k,d)}$. We solve Eq. 2

once, with $\mathbf{S} = [\mathbf{X}_1^{(1,1)}, \dots, \mathbf{X}_N^{(K_N,D)}]$ and $\mathbf{T} = [\bar{\mathbf{X}}_1, \dots, \bar{\mathbf{X}}_N]$. Classification is done as Eq. 1, with $\mathbf{t}_i = \mathbf{w}^\top \bar{\mathbf{X}}_i$.

Fourthly, we test an aCCA that learns a universal template across classes exploiting their lag differences (**aCCA-Lag**). We assume a shared template, because each class has the same underlying m-sequence, but only differs in its temporal shift τ_i . Thus, for calibration, we can align all cycles to zero lag via $\tilde{\mathbf{X}}_i^{(k,d)} = \mathcal{S}_{-\tau_i}(\mathbf{X}_i^{(k,d)})$, where $\mathcal{S}_{\tau_i}$ denotes a circular temporal shift by $-\tau_i$ samples, and then estimate a single canonical template as $\bar{\mathbf{X}}_0 = \frac{1}{\sum_{i=1}^N K_i D} \sum_{i=1}^N \sum_{k=1}^{K_i} \sum_{d=1}^D \tilde{\mathbf{X}}_i^{(k,d)}$. The class-specific templates are generated by circular shifts $\bar{\mathbf{X}}_i = \mathcal{S}_{\tau_i}(\bar{\mathbf{X}}_0)$. We construct the concatenated multichannel calibration data $\mathbf{S} = [\mathbf{X}_1^{(1,1)}, \dots, \mathbf{X}_N^{(K_N,D)}]$ and the corresponding stacked templates $\mathbf{T} = [\bar{\mathbf{X}}_1^{(K_1 D)}, \dots, \bar{\mathbf{X}}_N^{(K_N D)}]$, where $\bar{\mathbf{X}}_i^{(K_i D)}$ denotes $K_i D$ repetitions of $\bar{\mathbf{X}}_i$. This is again used to solve Eq. 2 to obtain $\mathbf{w}$. For classification we use template response $\mathbf{t}_i = \mathbf{w}^\top \bar{\mathbf{X}}_i$ in Eq. 1.

rCCA: The previous approaches averaged at the trial or cycle level. However, we can potentially assume that the response to a single event (e.g., a flash) within such a cycle is the same, no matter where and when it is presented. That is, given the linear superposition hypothesis, we assume that the response to a sequence of events is the linear summation of the responses to individual events. This is formulated in reconvolution CCA (rCCA) [12].

For rCCA, however, different event definitions can be used, such as the duration of the on/off states, the on and/or off states themselves, the rising and/or falling edges, or the stimulus sequence. In turn, rCCA learns the temporal response function to each of the defined events.

Table 1: Number of parameters and minimum training data. Listed are the number of parameters and the minimum required training data of each of the classifiers. Here, the number of classes $N = 4$, the number of channels $C = 16$, the sampling frequency $F_s = 600$, the number of event-related time points $L = 0.3 \cdot F_s$, and the number of events $E = 2$ for rCCA-Edge and $E = 1$ for rCCA-ID.

Classifier	Mode	N. Parameters	Min. Data
aCCA-Ens	Trial	$N \cdot C + N \cdot 4.65 \cdot F_s$ (11176)	N trials $4 \cdot 4.65 \cdot F_s$ (11160)
aCCA-Trl	Trial	$C + N \cdot 4.65 \cdot F_s$ (11176)	N trials $4 \cdot 4.65 \cdot F_s$ (11160)
aCCA-Cyc	Cycle	$C + N \cdot 1.55 \cdot F_s$ (3736)	N cycles $4 \cdot 1.55 \cdot F_s$ (3720)
aCCA-Lag	Cycle Lag	$C + 1.55 \cdot F_s$ (946)	1 cycle $1.55 \cdot F_s$ (930)
rCCA-Edge	Event Lag	$C + E \cdot L$ (376)	E events $L \cdot E$ (360)
rCCA-ID	Event Lag	$C + E \cdot L$ (196)	E events $L \cdot E$ (180)

We evaluated two event definitions for rCCA: firstly, we used the rising and falling edges (**rCCA-Edge**) to determine events. This has been shown to generalize across binary sequences [7, 10]. This event definition consists of an event for when the stimulus rises from 0 to 1 (c-VEP) or smoothly grows in size (c-MVEP), and when it falls from 1 to 0 (c-VEP) or shrinks in size (c-MVEP). The rCCA then learns separate responses for the two events. Secondly, because the c-MVEP is using a continuous motion path, we evaluated the direct convolution of the stimulus sequence and one temporal response function, scaled based on the amplitude of the sequence (**rCCA-ID**).

The rCCA can be applied on single cycles, so we again have $\mathbf{X}_i^{(k,d)} \in \mathbb{R}^{C \times T_c}$. A Toeplitz structure matrix $\mathbf{M}_i \in \mathbb{R}^{M \times T_c}$ is constructed, where $M = E \cdot L$, with E denoting the number of events, and L the length of the modeled response per event, here 300 ms so $L = 0.3 \cdot f_s$. rCCA-Edge has $E = 2$, corresponding to the rising and falling edge, and rCCA-ID has $E = 1$, corresponding to the stimulation input. The structure matrix $\mathbf{M}_i$ models the onset, duration, and overlap of responses to each event in the sequence of the i -th class. Then, we solve Eq. 3 with concatenated calibration data $\mathbf{S} = [\mathbf{X}_1^{(1,1)}, \dots, \mathbf{X}_N^{(K_N,D)}]$ and stacked structure matrices $\mathbf{M} = [\mathbf{M}_1, \dots, \mathbf{M}_N]$, where $\mathbf{M}_i$ is repeated $K_i D$ times.

The general rCCA can be formulated as follows:

$$\arg \max_{\mathbf{w}, \mathbf{r}} \rho(\mathbf{w}^\top \mathbf{S}, \mathbf{r}^\top \mathbf{M}) \quad (3)$$

where both the spatial filter $\mathbf{w} \in \mathbb{R}^C$ for C channels and event-related response(s) $\mathbf{r} \in \mathbb{R}^M$ for M event-related time points are optimized. For classification, template response $\mathbf{t}_i = \mathbf{r}^\top \mathbf{M}_i$ is used in Eq. 1. Code for rCCA is available at <https://github.com/thijor/pyntbci>.

Analysis: The data were analyzed using a leave-one-block-out cross-validation, using six folds. In each of the folds, learning curves were estimated, where the amount of training data, i.e., train-time, started at a minimum and increased to using all available training data. All the

methods have different requirements regarding the minimum training data needed. Hence, the starting point and step size differ between methods (see Tab. 1).

The training data contained five trials of 4.65 s for each of the N classes, where each trial consisted of three cycles of 1.55 s. For aCCA-Ens and aCCA-Trl, the learning steps ranged from $N \times 4.65$ s (at least a full trial of 4.65 s per class) to $5 \times N \times 4.65$ s in steps of $N \times 4.65$ s, since they require at least one trial per class. The aCCA-Cyc used learning steps ranging from $N \times 1.55$ s (at least one full cycle of 1.55 s per class) to $5 \times 3 \times N \times 1.55$ s in steps of $N \times 1.55$ s, because it requires at least one cycle per class. The aCCA-Lag, rCCA-Edge and rCCA-ID used learning steps from 1.55 s to $5 \times 3 \times N \times 1.55$ s in steps of 1.55 s, because they can generalize across classes. Note that rCCA-Edge and rCCA-ID could use even less training data, as long as some data is present for each event.

To ensure identical training data across classifiers, cycles were added in a trial-like manner. Specifically, the cycle of another trial was added before adding the next cycle within a trial. This was done until we had a trial of all classes, that is $N \times 1.55$, after which we continued by adding the second cycle of each of the first N trials. This ensured that, for any given amount of calibration data, the classifiers were evaluated using identical data. For each fold, the testing data contained N full-trials, one for each class, used to estimate the performance of the classifier (see bci-lab.hochschule-rhein-waal.de/en/acc.html).

Statistical significance was tested using pairwise Wilcoxon tests, with Bonferroni correction.

RESULTS

We systematically compared the classification accuracy of the six classifiers across different amounts of training data while testing on a full trial (see Fig. 2 and Tab. 2). Note that the methods require different minimum training data, resulting in different learning curves onsets: 18.6 s (aCCA-Ens, aCCA-Trl), 6.2 s (aCCA-Cyc), 1.55 s (aCCA-Lag, rCCA-Edge, rCCA-ID), although the rCCA methods could start even earlier, see Tab. 1. For the selected time-points, p -values were Bonferroni-adjusted for the number of pairwise comparisons (3 at 1.55 s; 6 at 6.2 s; 15 at 18.6 and 93.0 s).

c-MVEP: We first evaluate at 1.55 s, which is the minimum data for aCCA-Lag and both reconvolution-based methods. The traditional methods (aCCA-Trl, and aCCA-Ens) and the cycle-based aCCA are not included here due to insufficient data. rCCA-Edge (46.7%) and rCCA-ID (43.6%) did not differ significantly ($p > .879$). However, both reconvolution-based methods outperformed aCCA-Lag (30.5%; $p < .04$).

Secondly, at 6.2 s, also sufficient for aCCA-Cyc with one cycle per class, reconvolution based on the rising and falling edges (76.5%) outperformed the direct convolution with the stimulus sequence (62.1%; $p = .037$). Both reconvolution-based approaches and aCCA-Lag (74.3%) outperformed aCCA-Cyc (31.8%; $p < .002$). Any other

Table 2: **Offline decoding versus learning time.** Mean decoding accuracy (in %) and standard error for each classifier at four learning times, where minimum data requirements were met. Grey rows indicate c-VEP, white rows c-MVEP.

Classifier	1.55 s	6.2 s	18.6 s	93.0 s
aCCA-Ens	-	-	41.9 (3.1)	98.9 (0.8)
	-	-	36.8 (2.1)	81.1 (5.2)
aCCA-Trl	-	-	41.7 (3.4)	99.8 (0.2)
	-	-	41.0 (2.8)	89.5 (3.6)
aCCA-Cyc	-	42.3 (4.9)	99.8 (0.2)	100.0 (0.0)
	-	31.8 (2.5)	83.1 (4.7)	96.9 (1.3)
aCCA-Lag	36.6 (4.1)	92.3 (3.6)	99.8 (0.2)	100.0 (0.0)
	30.5 (3.2)	74.3 (5.8)	93.9 (2.4)	98.2 (0.7)
rCCA-Edge	72.4 (6.4)	99.1 (0.6)	100.0 (0.0)	100.0 (0.0)
	46.7 (5.3)	76.5 (5.0)	89.9 (3.2)	96.9 (1.2)
rCCA-ID	56.6 (6.9)	93.6 (2.4)	99.6 (0.3)	99.6 (0.4)
	43.6 (4.9)	62.1 (6.2)	77.4 (4.4)	90.1 (3.0)

comparison was not significant ($p > .05$).

Thirdly, we compared at 18.6 s, where the calibration data contains a full trial for each class, hence all classifiers can be considered. The aCCA-Lag outperformed all other classifiers ($p < .05$). Reconvolution using the edge event (89.9%) outperformed the direct convolution (77.4%; $p = 0.005$) and both traditional methods (36.8% and 41.0%; $p < .005$), but had no significant difference with the cycle-based aCCA (83.1%; $p = 0.098$). The cycle-based aCCA and the direct convolution outperformed both traditional methods ($p < .002$). The cycle-based aCCA and the direct convolution did not differ significantly, nor did the traditional methods ($p > .05$).

Fourthly, we compared the methods at 93.0 s, the maximum training duration. Using universal templates across classes (98.2%) results in significantly higher decoding accuracies than direct convolution with the stimulus sequence (90.1%; $p = .037$). Additionally, using the ensemble approach (81.1%) yields significantly lower performances than the trial (89.5%), cycle (96.9%), and lag-based (98.2%) approaches and reconvolution with the edge event (96.9%; $p < .03$). None of the other comparisons were significant ($p > .05$).

c-VEP: Overall, the methods show a similar pattern for c-VEP as for c-MVEP, but with higher accuracy for c-VEP (Tab. 2 and Fig. 2). We compared the c-VEP methods using the same learning time points.

First, at 1.55 s, rCCA-Edge (72.4%) significantly outperformed rCCA-ID ($p = .031$). Both reconvolution-based approaches achieved significantly higher decoding accuracies than the lag-based aCCA (36.6%; $p < .02$).

Secondly, with one cycle per class, estimating templates by averaging over cycles (42.3%) resulted in significantly lower accuracies than estimating a universal template across classes (92.3%) or using the reconvolution approaches (99.1% and 93.6%; $p < .001$). Any other comparison did not show a significant difference ($p > .05$).

Thirdly, with one trial per class, both traditional approaches (41.9% and 41.7%) reached significantly lower accuracies than all other methods (mean accuracies between 99.6% and 100.0%; $p < .001$). Any other com-

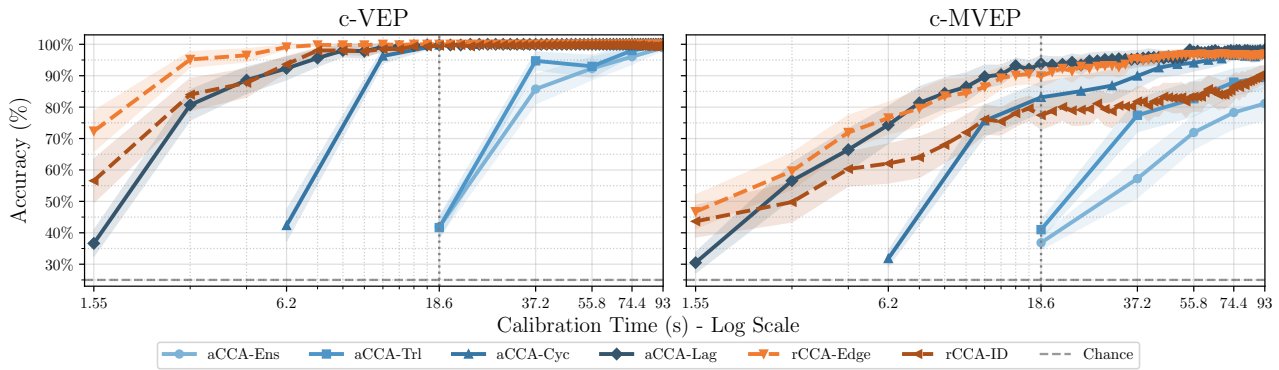


Figure 2: **Offline Learning Curves.** Grand average decoding accuracy and standard error for each classifier across learning time. Left: c-VEP, right: c-MVEP. Each classifier starts from its minimum data requirement to the maximum amount of training data available, in its individual step sizes. Vertical dotted lines represent a step size of a cycle (1.55 s). The vertical solid line at 6.2 s indicates the data included one cycle per class. Solid lines after 18.6 s indicate a step size of a full trial per class.

parison was not statistically significant ($p > .05$).

Fourthly, we evaluated when all training data was used for calibration (93.0 s). Classification accuracies for all classifiers were between 98.9% and 100.0%, and showed no significant difference ($p > .05$).

c-VEP versus c-MVEP: Finally, we compared the accuracies of the best-performing classifiers for c-MVEP and c-VEP at two learning time points. At 18.6 s, with one full trial for each class, c-MVEP reached the highest decoding accuracies when learning universal templates across classes, with a mean accuracy of 93.86%. For c-VEP, rCCA-Edge performed best, reaching 100%. A pairwise Wilcoxon showed that c-VEP significantly outperformed c-MVEP ($p = .009$). When using all training data for calibration, at 93.0 s, learning universal templates across classes for c-MVEP resulted in a mean decoding accuracy of 98.25%. For c-VEP, rCCA-Edge again reached 100%, which was significantly higher than c-MVEP (pairwise Wilcoxon; $p = .01$).

DISCUSSION

We investigated decoding approaches for the proposed code-modulated motion visual evoked potential (c-MVEP), using smoothed m-sequences for motion-based BCI stimulation. Given its code-modulated nature, we hypothesized that c-VEP decoding assumptions transfer to c-MVEP, and indeed observed improved decoding accuracy when applying complexity-reducing assumptions. First, the traditional decoding approaches aCCA-Ens and aCCA-Trl, both needing to estimate a substantial amount of parameters, achieve moderate performance for c-MVEP, with 81.1% and 89.5%, respectively, on the maximum available training data (93.0 s). Additionally, as the learning curves continued to increase, suggesting that even more training data was required, the findings confirm the inefficiency of these traditional methods.

Second, assuming identical responses across cycles, the aCCA-Cyc at 18.6 s and 93.0 s reached 83.1% and 96.9% for c-MVEP. With only a trial per class, the averaging over cycles provides effectively more data, while at the

same time needing to estimate less parameters than the traditional methods at full training data, suggesting that cycle averaging can substantially improve results.

Third, aCCA-Lag learned a universal template across classes by exploiting m-sequence lag differences. This reduced model complexity and achieved 74.3% accuracy for c-MVEP with 6.2 s of calibration, increasing to 93.9% with 18.6 s, suggesting that fewer parameters and effectively more training samples, improve accuracy and learning speed.

Fourth, we applied reconvolution, which assumes identical event responses, and as such further lowers model complexity. rCCA-Edge achieved high decoding accuracies with minimal training data: c-MVEP reached 76.5% with only 6.2 s, improved to 89.9% at 18.6 s, and to 96.9% at maximum data. For c-VEP, one stimulus cycle of 1.55 s was sufficient to reach 72.4%, 6.2 s to reach 99.1%, after which it converged to 100%.

Finally, we modeled responses as the linear convolution of the stimulus and one impulse response function, the model with the smallest complexity. However, this may have led to underfitting: For c-MVEP, rCCA-ID reached 43.6% with one cycle, 62.1% with one cycle per class, 77.4% with one trial per class and, finally, 90.1% at maximum available data (93.0 s). For c-VEP, this was 56.6% with one cycle, 93.6% with one cycle per class, and converged to 99.6% at one trial per class.

For c-VEP, the results are similar to those observed in prior work [9, 10], with all methods reaching close to 100% with sufficient training data.

In summary, reconvolution-based approaches appear well-suited for c-MVEP decoding and allow model calibration with minimal training data. Consistent with prior work [7, 10], they show superior performance for c-VEP. Interestingly, c-MVEP achieved significantly lower accuracies than c-VEP. We believe that at least two factors require further investigation. First, the radial zooming amplitude was limited to 50% of the stimulus size. Second, c-VEP used full-contrast flicker, while c-MVEP had a lower target-to-background contrast. Both may have caused weaker c-MVEP responses. Future studies should

investigate a fairer comparison in stimulus contrast.

Despite c-MVEP achieving lower performance than c-VEP, we believe it offers a viable alternative when high-contrast flickering is undesirable, impractical, or uncomfortable. It may be particularly suitable for low-light settings, sensitive users, or extended BCI use [4]. Therefore, despite the lower performance, c-MVEP broadens the range of code-modulated protocols.

Interestingly, for both c-MVEP and c-VEP, the rCCA-ID model yielded lower accuracies than rCCA-Edge. These findings suggest that rCCA-ID may be overly constrained, estimating too few parameters and thereby underfitting the response. Specifically, the assumption of full linearity with one temporal response function scaled by stimulus amplitude seems violated. For both contrast and motion reversals, it seems there are sub-additive properties that should be modeled carefully. While the rCCA-Edge that encodes stimulus contrast performed well, future studies should investigate the event model further. This highlights the trade-off between model simplicity and sufficient representational capacity.

For c-MVEP, we showed that it can be accurately decoded after calibration on only one cycle per class. Increasing this to one trial per class, the aCCA-Lag achieved a decoding accuracy of 93.9%. This is a promising outlook for future c-MVEP BCI applications, particularly when combined with continuous cumulative learning to improve the classifier during online use. Specifically, the user could use the system with minimal calibration, while the decoder progressively refines and optimizes the learned parameters during online operation. While this study shows clear results for constraining the decoding model, the evaluations were based on a single offline dataset. Replication on other data and an online study should be performed to further validate these findings. Furthermore, future work should address a larger number of classes beyond the used 4 targets in this work.

CONCLUSION

In this study, we introduced a motion stimulus into the c-VEP domain, leading to the code-modulated motion visual evoked potential (c-MVEP). We evaluated different classifiers, each having a smaller parameter space and requiring less training data. With minimal training data, a model learning a universal template across classes and a reconvolution-based model using stimulus events, achieved high decoding accuracies after just 6.2 s of calibration data. This demonstrates that the c-MVEP can be effectively decoded without requiring extensive amounts of training data for calibration.

ACKNOWLEDGMENTS

This work was supported by the European Union's research and innovation program under the Marie Skłodowska-Curie grant agreement No 101118964, the "The Friends of the University Rhine-Waal - Campus Cleve" association, and the Dutch Research Council (NWO) (project number 024.005.022).

AUTHOR CONTRIBUTIONS

HS: Conceptualization, Methodology, Formal Analysis, Writing - original draft - review & editing; JT: Conceptualization, Methodology, Writing - review & editing, Supervision; MT: Methodology, Writing - review & editing, Supervision; SB: Writing - review & editing; IV: Writing - review & editing, Funding acquisition

REFERENCES

- [1] Wolpaw J, Wolpaw EW. Brain-Computer Interfaces Principles and Practice. Oxford University Press (2012).
- [2] Xie J, Xu G, Wang J, Li M, Han C, Jia Y. Effects of Mental Load and Fatigue on Steady-State Evoked Potential Based Brain Computer Interface Tasks: A Comparison of Periodic Flickering and Motion-Reversal based visual attention. *PLOS ONE*. 2016;11(9):e0163426.
- [3] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-Computer Interfaces Based on Code-Modulated Visual Evoked Potentials (c-VEP): A Literature Review. *Journal of Neural Engineering*. 2021;18(6):061002.
- [4] Xie J, Xu G, Wang J, Zhang F, Zhang Y. Steady-State Motion Visual Evoked Potentials Produced by Oscillating Newton's Rings: Implications for Brain-Computer Interfaces. *PLoS ONE*. 2012;7(6):e39707.
- [5] Chai X, Zhang Z, Guan K, Liu G, Niu H. A Radial Zoom Motion-Based Paradigm for Steady State Motion Visual Evoked Potentials. *Frontiers in Human Neuroscience*. 2019;13.
- [6] Volosyak I, Rezeika A, Benda M, Gembler F, Stawicki P. Towards solving of the Illiteracy phenomenon for VEP-based brain-computer interfaces. *Biomedical Physics & Engineering Express*. 2020;6(3):035034.
- [7] Thielen J. Addressing BCI inefficiency in c-VEP-based BCIs: A comprehensive study of neurophysiological predictors, binary stimulus sequences, and user comfort. *Biomedical Physics & Engineering Express*. 2025;11(4):045017.
- [8] Scheppink H, Herpers R, Thielen J, Volosyak I. Beyond Flickering: Introducing Code-Modulated Motion Visual Evoked Potentials for Brain-Computer Interfacing. *arXiv preprint arXiv:2605.15801*.
- [9] Gembler F, Volosyak I. A Novel Dictionary-Driven Mental Spelling Application Based on Code-Modulated Visual Evoked Potentials. *Computers*. 2019;8(2):33.
- [10] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. *Journal of Neural Engineering*. 2021.
- [11] Scheppink H, Ahmadi S, Desain P, Tangermann M, Thielen J. Towards auditory attention decoding with noise-tagging: A pilot study. In: *Proceedings of the 9th Graz Brain-Computer Interface Conference 2024*. 2024, 337–342.
- [12] Thielen J, Broek P van den, Farquhar J, Desain P. Broad-Band Visually Evoked Potentials: Re(con)volution in Brain-Computer Interfacing. *PLOS ONE*. 2015;10(7):e0133797.

Advancing NIRS-Based Insights into Motor Imagery and Execution of Swallowing: Effects of Artifact Correction Methods and Hemodynamic Validation

S.E. Kober¹, V.I. Engele², T. Kanatschnig¹

¹ Department of Psychology, University of Graz, Graz, Austria

² Department of Psychiatry, Psychotherapy, Psychosomatic and Medical Psychology, University Hospital for Psychiatry II, Medical University of Innsbruck, Innsbruck, Austria

E-mail: silvia.kober@uni-graz.at

ABSTRACT: Motor imagery (MI) of swallowing has proven to be a promising mental strategy for brain-computer interface (BCI) and neurofeedback applications aimed at treating swallowing difficulties. Previous near-infrared spectroscopy (NIRS) studies demonstrated comparable hemodynamic responses over the inferior frontal gyrus during motor execution (ME) and MI of swallowing, particularly for deoxygenated hemoglobin. However, due to the reliance on manual artifact rejection, it remained unclear whether these findings were influenced by motion artifacts. In this study, 33 healthy young adults were tested, and their hemodynamic responses during ME and MI of swallowing were assessed using NIRS. By applying advanced and automated motion artifact correction methods, such as wavelet filtering and short-distance channel regression, the results were largely consistent with previous studies that used manual artifact rejection. This validates earlier findings and confirms that the observed activation patterns in the hemodynamic response were not primarily caused by motion artifacts, demonstrating the robustness of prior conclusions. When directly comparing motion artifact correction methods (manual rejection, wavelet filtering, and short-distance channel regression), wavelet filtering effectively reduced concentration changes in oxygenated hemoglobin compared to manual correction, indicating fewer motion artifacts. In contrast, short-distance channel regression showed no significant effects, underscoring the need for further optimization of correction methods and the importance of selecting appropriate artifact correction techniques in NIRS studies investigating ME and MI of swallowing.

INTRODUCTION

Previous studies using near-infrared spectroscopy (NIRS) to investigate the hemodynamic response during motor execution (ME) and motor imagery (MI) of swallowing have shown that ME leads to an increase in HbO (oxygenated hemoglobin) and HbR (deoxygenated hemoglobin) in frontal brain regions, whereas MI results in a decrease in HbO and an increase in HbR, likely

reflecting inhibition of active swallowing [1–4]. The strongest activations were observed in the inferior frontal gyrus (IFG) for both ME as well as MI [1, 2, 5]. Furthermore, neurofeedback studies demonstrated that it is possible to upregulate HbR in the IFG during MI of swallowing, suggesting that MI-based neurofeedback could be a promising tool for the treatment of dysphagia [3, 5, 6].

Although NIRS is a robust method for measuring brain activity, it is susceptible to unsystematic variations caused by physiological noise and motion artifacts. The signal, which originates from light passing through the scalp and skull to the cortex, contains not only the hemodynamic response from neuronal activation but also noise from cardiac signals, arterial blood pressure oscillations (Meyer waves), respiration, and other physiological changes in extracortical layers. While high-frequency noise can be filtered out, low-frequency changes like Meyer waves are harder to detect and may correlate with the signal [7–10]. Additionally, motion artifacts, caused by optode displacement or scalp movements from jaw or eyebrow activity, can distort the signal in various forms, such as spikes, baseline shifts, or low-frequency variations. These artifacts may occur as isolated events or overlap with task-related hemodynamic responses, making them difficult to distinguish [9, 11].

Especially swallowing movements leading to jaw movements produce significant motion artifacts in NIRS signals due to strong temporal muscle activity [12]. Previous studies addressing neuronal correlates of swallowing relied on visual inspection and manual rejection of motion artifacts [1, 2, 13]. This approach is highly subjective and poses challenges in distinguishing motion artifacts from the hemodynamic response [8, 10, 11]. To ensure the comparability between ME and MI of swallowing and to establish MI as a potential tool for dysphagia therapy using it as a mental strategy in brain-computer interface (BCI) and neurofeedback applications, it is crucial to confirm that MI of swallowing elicits brain activation patterns comparable to those of ME.

Recent advancements in NIRS technology, such as the

use of short-distance (SD) channels and correction techniques like wavelet filtering, provide improved insights into the true hemodynamic response in signals affected by motion and physiological artifacts [8, 14, 15]. SD channels are a method used in NIRS studies to isolate and remove physiological noise originating from extracortical layers, such as the scalp and skull. Unlike standard long-distance channels, where the source and detector optodes are spaced approximately 3 cm apart to capture signals from both cortical and extracortical layers, SD channels place the detector close to the light source. This setup ensures that the detected signal originates only from extracortical layers [16, 17]. By simultaneously measuring signals at multiple distances, SD channels allow researchers to regress extracortical noise from cortical activation, improving signal quality [16, 18]. This approach complements other external instruments, such as pulse sensors, respiratory belts, or EMG, for extracting interfering measures from NIRS signals [7, 15, 19].

Wavelet filtering has been successfully applied as artifact correction method in speech studies, where motion artifacts were stimulus-dependent [8]. Wavelet filtering is a channel-by-channel approach that assumes signal components follow a normal distribution, with outliers attributed to motion artifacts [20]. Outliers, defined by the interquartile range, are set to zero, and the signal is recomposed using an inverse discrete wavelet transformation. This method preserves most of the frequency content and is effective for correcting spikes [8, 14].

The aim of this study was to validate previous findings [1–3] using a modern NIRS system equipped with SD channels and wavelet filtering, and to evaluate the effectiveness of these correction methods compared to manual artifact rejection. Therefore, we compared different NIRS artifact removal methods (SD, wavelet, manual rejection) and investigated their effects on the hemodynamic response elicited by ME and MI of swallowing. This study has two main objectives: (1) to validate and replicate previous findings on the hemodynamic response during ME and MI of swallowing using more advanced artifact handling methods (wavelet filtering and SD channel regression) and (2) to investigate the effectiveness of wavelet filtering and SD channel regression for motion artifact correction in NIRS signals.

MATERIALS AND METHODS

Participants: 33 healthy young adults between 18–35 years (20 females, 13 males) participated in the present study. All participants gave their written informed consent and the study was approved by the local ethics committee of the University of Graz (GZ. 39/4/63 ex 2020/21).

Task: In the present study, a motor execution (ME) and motor imagery (MI) task similar to [1, 2] were used, involving saliva swallowing instead of water. Participants performed 20 ME and 20 MI trials, each

lasting 15 seconds, presented in randomized order. During ME trials, participants swallowed their saliva 5–6 times at a comfortable pace, while during MI trials, they imagined swallowing at the same pace using a kinesthetic approach. Task instructions were displayed on a computer screen using PsychoPy [21], with the letter "A" indicating ME and "V" indicating MI, and triggers marking task onsets and offsets in the signal. Between tasks, a fixation cross was shown for 28–32 seconds, during which participants were instructed to avoid active swallowing. All participants received brief training before the task.

NIRS-recordings: The hemodynamic response during the ME/MI tasks was recorded using a NIRx Sport 2 system (NIRx Medizintechnik, GmbH, Berlin, Germany), a continuous wave system with LED sources emitting light at 760 and 850 nm wavelengths. The system measured changes in oxy- (HbO), deoxy- (HbR), and total hemoglobin in the cortex. The probe setup included eight source and seven detector optodes mounted on an EEG cap, ensuring a 30 mm distance between each source-detector pair (long-distance channels). Additionally, eight short-distance detectors were placed 8 mm from the source optodes (SD channels), requiring one long-distance detector to be repurposed. The sampling rate was set to 10.2 Hz, and NIRx software (NIRSite 2.0 and Aurora fNIRS 1.4) was used for channel configuration and data recording. The probe setup was positioned bilaterally above the IFG and included regions such as the dorsolateral prefrontal cortex, premotor and supplementary motor cortex, frontal eye fields, middle and superior temporal gyrus, subcentral area, primary somatosensory cortex, supramarginal and angular gyrus (Wernicke), and somatosensory association cortex, following previous NIRS studies on ME/MI of swallowing [1, 2]. Visualizations of the probe setup (see Fig. 1) and recorded data were generated using AtlasViewer [22], a MATLAB-based software package. Channel pairs S2-D4, S4-D3, and S4-D4 over the left hemisphere and channel pairs S6-D7, S8-D6, and S8-D7 over the right hemisphere covered the IFG (Fig. 1).

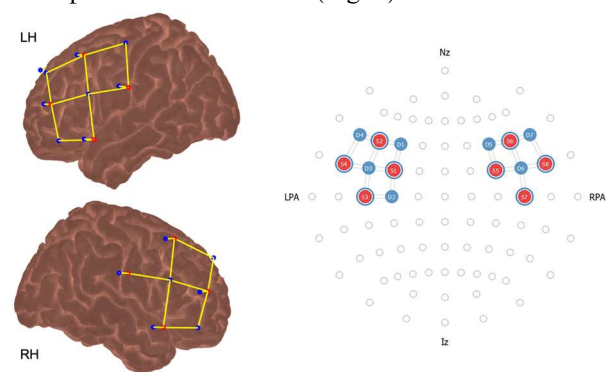


Figure 1: The NIRS probe was positioned over the left (LH) and right (RH) hemispheres. Red markers represent source optodes, blue markers indicate both long- and short-distance detectors, yellow lines illustrate corresponding source-detector pairs.

NIRS data preprocessing: NIRS data preprocessing was performed using the MATLAB-based program Homer2 (Huppert et al., 2009), comparing four artifact correction methods. Raw optical density (OD) data was converted into OD changes using the Homer2 function *hmrIntensity2OD*, and channels with low OD were discarded via *enPruneChannels*. Motion artifacts were corrected either with wavelet filtering (*Wav and Wav + SD*) using *hmrMotionCorrectWavelet* ($iqr = 0.1$) or manually (*manual* and *manual + SD*).

Motion artifacts were further detected and excluded using *hmrMotionArtifact* and *enStimRejection*. All pathways applied high-pass (0.01 Hz) and low-pass (0.50 Hz) filters via *hmrBandpassFilt* to remove physiological noise, followed by conversion to HbO and HbR concentration changes using *hmrOD2Conc*. For SD-channel regression, *hmrDeconvHRF_DriftSS* was used to remove hemodynamic drift. HbO and HbR changes were baseline-referenced to 5 seconds before stimulus onset (-5 to 0 seconds).

Block averages were calculated for -5 to 40 seconds around stimulus onset using *hmrBlockAvg*. After visual inspection, two participants were excluded due to poor data quality, leaving 31 participants. Data was averaged for Task (5–15 seconds) and Pause (15–25 seconds) intervals for ME and MI separately, using Homer's *Export Mean Results* function.

Statistical analysis: To assess the topographical distribution of significantly activated brain areas, paired sample *t*-tests were conducted comparing the activation interval (5–25 seconds) to the baseline (-5–0 seconds) for each channel, condition (ME and MI), and hemoglobin type (HbO and HbR). Significant changes were identified using the false discovery rate (FDR) method to correct for multiple comparisons [23]. For this analysis, only data processed with wavelet filtering and SD channel regression was included to compare the current findings with previous NIRS studies that relied solely on manual artifact rejection (e.g., [1]) in order to determine whether different artifact handling techniques yield varying results.

Regions of interest (ROIs) were defined based on [1, 2] for comparability, targeting channels associated with the left (channel pairs: S2-D4, S4-D3, S4-D4) and right IFG (channel pairs: S6-D7, S8-D6, S8-D7; see Fig. 1).

To analyze HbO and HbR changes during ME and MI, two separate 2x2x2 repeated-measures ANOVAs were performed with within-subject factors *Task* (ME vs. MI), *Time* (Task vs. Pause), and *Hemisphere* (Left vs. Right), using HbO and HbR concentration changes as dependent variables. Again, to confirm whether earlier findings that relied solely on manual artifact correction (e.g., [1]) were influenced by motion artifacts, this analysis included only data processed using wavelet filtering and SD channel regression to validate those results.

Additionally, two 2x2 repeated-measures ANOVAs were conducted to compare motion artifact correction methods, with factors *Correction Method* (Wavelet vs. Manual) and *SD-Channel Regression* (Yes vs. No). HbO and HbR concentration changes during the ME task (5–

15 seconds) were used as dependent variables, as motion artifacts are most prominent during active swallowing. To account for differences in hemodynamic response function (HRF) scaling across correction methods, HbO and HbR values were standardized via z-transformation within subjects across all methods.

RESULTS

The FDR analysis revealed that the strongest NIRS signal changes were observed over the bilateral IFG, mainly in the channel pairs merged for the left and right IFG ROIs (e.g., for ME HbO, the five strongest significant channel pairs were S8-D7, S4-D4, S2-D3, S4-D3, and S8-D6; number of sign. channel pairs ME HbO: 18; MI HbO: 18; ME HbR: 3; MI HbR: 2). HbO increased during ME but not during MI, while HbR increased during ME as well as MI (see Fig. 2).

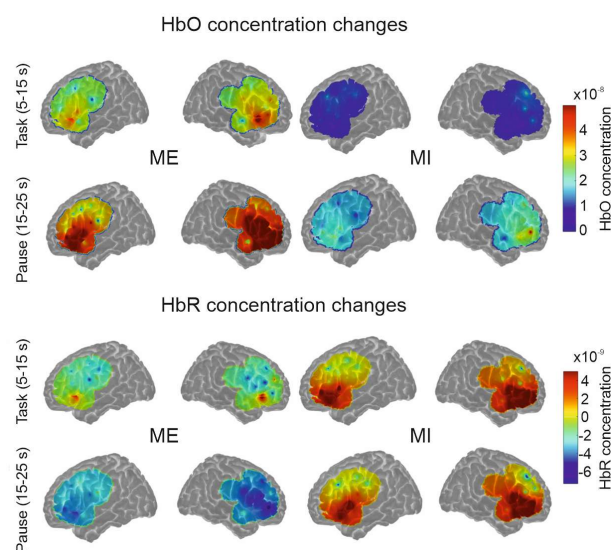


Figure 2: Topographical distributions of HbO and HbR changes averaged for the Task and the Pause interval, presented separately for the motor execution (ME) and motor imagery (MI) conditions.

For HbO, the 2x2x2 repeated-measures ANOVA for NIRS data preprocessed with wavelet filtering and SD channel regression revealed a significant main effect for *Task* ($F(1,29) = 27.61, p < .001, \eta^2 = .488$), indicating higher activation during ME ($M = .063 \text{ m(mol/l)} * \text{mm}$, $SE = .005$) compared to MI ($M = .015 \text{ m(mol/l)} * \text{mm}$, $SE = .008$). Additionally, a significant main effect for *Time* was found ($F(1,29) = 38.77, p < .001, \eta^2 = .572$), with higher HbO levels during the Pause interval ($M = .055 \text{ m(mol/l)} * \text{mm}$, $SE = .007$) compared to the Task interval ($M = .023 \text{ m(mol/l)} * \text{mm}$, $SE = .004$). A significant interaction between *Hemisphere* and *Time* ($F(1,29) = 15.46, p < .001, \eta^2 = .348$) showed greater activation in the right hemisphere ($M = .061 \text{ m(mol/l)} * \text{mm}$, $SE = .008$) compared to the left hemisphere ($M = .023 \text{ m(mol/l)} * \text{mm}$, $SE = .005$) during the Pause interval (see Fig. 2). The interaction between *Task* and *Time* almost reached significance ($F(1,29) = 4.00, p = .055, \eta^2$

= .121), suggesting higher HbO during Pause ($M = .082$ m(mol/l) * mm, $SE = .008$) compared to Task ($M = .045$ m(mol/l) * mm, $SE = .004$) for ME, while MI showed a slight increase during Pause ($M = .029$ m(mol/l) * mm, $SE = .010$) compared to Task ($M = .002$ m(mol/l) * mm, $SE = .007$). All other effects were not statistically significant: *Hemisphere* ($F(1,29) = 2.02$, $p = .166$, $\eta_p^2 = .065$), *Task*Hemisphere* ($F(1,29) = 1.76$, $p = .195$, $\eta_p^2 = .057$), and *Task*Time*Hemisphere* ($F(1,29) = .00$, $p = .959$, $\eta_p^2 = .000$).

The analysis of HbR concentration changes revealed a significant main effect of *Task* ($F(1,29) = 9.72$, $p < .01$, $\eta_p^2 = .251$), with higher activation during MI ($M = .011$ m(mol/l) * mm, $SE = .005$) compared to ME ($M = -.004$ m(mol/l) * mm, $SE = .002$). Additionally, a significant interaction between *Task* and *Time* was found ($F(1,29) = 5.67$, $p < .05$, $\eta_p^2 = .163$), showing a reduction in HbR concentration during the Pause interval for ME ($M = -.007$ m(mol/l) * mm, $SE = .003$) compared to the Task interval ($M = .001$ m(mol/l) * mm, $SE = .002$). Fig. 2 illustrates this interaction effect. No other effects reached statistical significance: *Time* ($F(1,29) = 1.81$, $p = .189$, $\eta_p^2 = .059$), *Hemisphere* ($F(1,29) = 0.27$, $p = .610$, $\eta_p^2 = .009$), *Time*Hemisphere* ($F(1,29) = .00$, $p = .948$, $\eta_p^2 = .000$), *Task*Hemisphere* ($F(1,29) = 1.50$, $p = .230$, $\eta_p^2 = .049$), and *Task*Time*Hemisphere* ($F(1,29) = 0.01$, $p = .910$, $\eta_p^2 = .000$).

The 2x2 repeated-measures ANOVA for HbO comparing the different artifact correction methods revealed a significant main effect of *Correction Method* ($F(1,29) = 35.36$, $p < .001$, $\eta_p^2 = .549$), with manual correction ($M = .612$ z-transf. m(mol/l) * mm, $SE = .103$) resulting in higher HbO levels compared to wavelet filtering ($M = -.612$ z-transf. m(mol/l) * mm, $SE = .103$). The main effect of *SD-Channel Regression* was marginally significant ($F(1,29) = 3.11$, $p = .088$, $\eta_p^2 = .097$), suggesting slightly higher HbO levels without SD regression ($M = .062$ z-transf. m(mol/l) * mm, $SE = .035$) compared to with SD regression ($M = -.062$ z-transf. m(mol/l) * mm, $SE = .035$). The interaction between *Correction Method* and *SD-Channel Regression* was not statistically significant ($F(1,29) = 2.18$, $p = .150$, $\eta_p^2 = .070$).

For HbR, no significant effects were observed for *Correction Method* ($F(1,29) = 2.07$, $p = .161$, $\eta_p^2 = .067$), *SD-Channel Regression* ($F(1,29) = 0.626$, $p = .435$, $\eta_p^2 = .021$), or the interaction between *Correction Method* and *SD-Channel Regression* ($F(1,29) = 0.23$, $p = .638$, $\eta_p^2 = .008$).

DISCUSSION

This study aimed to replicate and extend previous NIRS findings on cortical activation during ME and MI of swallowing, which has the potential to be a useful mental strategy in BCI and neurofeedback applications to treat dysphagia symptoms [2, 3, 5, 6]. Earlier studies showed that ME increases both HbO and HbR levels, while MI leads to increased HbR and decreased HbO, with the strongest changes observed in the IFG [1–3]. Given the possibility of motion artifacts caused by simultaneous

muscle activation during swallowing [12], this study applied advanced artifact correction methods, including wavelet filtering and SD-channel regression, to validate and refine previous results. These techniques are effective in reducing noise unrelated to cortical activation [8, 14] and were compared to manual artifact rejection used in earlier studies.

When employing advanced motion artifact reduction techniques, such as wavelet filtering and regressing SD channel information from the NIRS signal, the topographical distribution of HbO and HbR remains consistent with previous NIRS studies that relied solely on manual artifact rejection. The strongest signal changes were observed over the IFG [1, 2]. These results successfully replicate earlier findings, demonstrating that prior findings were not merely influenced by jaw movement artifacts during ME of swallowing [1, 2].

This study successfully replicated and extended previous findings on cortical activation during ME and MI of swallowing, confirming that ME leads to higher HbO levels, while MI results in increased HbR and decreased HbO. These findings are consistent with earlier research [1, 2] and align with studies on MI of other movements, which also report weaker HbO activation during MI compared to ME [24]. The reduced HbO activation during MI may be attributed to the absence of actual movement and sensory feedback, as well as inhibitory mechanisms that prevent unintentional swallowing during MI [1, 25]. Interestingly, HbO levels during MI decreased during the task but increased steadily during the resting period, a pattern commonly observed in MI of swallowing but not in MI of other movements [26, 27]. This may reflect the complexity and reflexive nature of swallowing, which involves effortful inhibitory processes [1].

The study also found that HbO levels were significantly higher during the Pause interval following both ME and MI, indicating a prolonged hemodynamic response compared to other movements [26]. This prolonged time course is consistent with earlier studies on swallowing [1, 2] and may be explained by the duration and complexity of the swallowing process itself [28].

Regarding lateralization, HbO levels during the resting period were higher in the right hemisphere, while no hemispheric differences were observed during the tasks. Previous studies have reported predominantly bilateral activation during swallowing [1], but neuroimaging studies suggest lateralization effects depending on the type of bolus swallowed. For example, water swallowing tends to activate the right hemisphere more strongly, while saliva swallowing shows left-dominant activation [4, 29]. The right lateralization observed in this study during voluntary saliva swallowing may reflect insula activation, as suggested by [30], and could have been masked by motion artifacts in earlier studies. However, the lateralization effect during resting periods was small and may also be influenced by interindividual differences in gustatory sensations, as the anterior insula processes taste stimuli [29].

In contrast to HbO, MI resulted in stronger HbR

concentration changes compared to ME, with HbR levels decreasing over time during ME (comparing Task and Pause intervals). Similar findings were reported in previous research on ME and MI of swallowing, although these effects were primarily observed during a pause interval, while during the task itself, HbR levels were higher for ME than MI [1]. However, as [1] did not apply advanced motion artifact correction methods, such as wavelet filtering or SD-channel regression, the elevated HbR concentrations during active swallowing in their study may have been influenced by motion artifacts caused by muscle activity during swallowing [12, 31]. To sum up, motion artifacts caused by muscle activity during swallowing [12, 31] were addressed in this study using advanced correction methods, including wavelet filtering and SD-channel regression. These methods effectively reduced noise unrelated to cortical activation, providing a clearer NIRS signal. The results were largely consistent with previous studies that relied on manual artifact rejection [1–4], validating earlier findings and confirming that the observed activation patterns in the HRF were not primarily due to motion artifacts. This demonstrates the robustness of prior conclusions and highlights the value of advanced artifact correction techniques in NIRS research.

Furthermore, this study compared different motion artifact correction methods for NIRS signals during ME of swallowing, focusing on wavelet filtering, SD-channel regression, and manual artifact rejection. Wavelet filtering significantly reduced HbO signal changes compared to manual rejection, indicating fewer motion artifacts. This method has been effective in speech studies with task-related artifacts (Brigadoi et al., 2014), but in the present study, the chosen iqr parameter (0.1) may have been overly conservative, potentially reducing the hemodynamic response itself [32]. Adjusting the iqr could help preserve the characteristics of the HRF during swallowing.

For HbR, no significant effects were observed for wavelet filtering or SD-channel regression, consistent with previous findings that HbR is more robust against physiological noise [7]. SD-channel regression, applied after wavelet filtering, showed no additional benefits, likely due to the already corrected data. While SD-channels are effective in reducing physiological noise from the scalp [11, 15], their role in addressing motion artifacts caused by muscle activity remains unclear.

The study also highlighted limitations in the sensitivity of the analysis to detect effects of correction methods. Measures like signal-to-noise ratios (SNR) or trial retention rates could provide better insights into the effectiveness of artifact correction techniques, especially in clinical populations where trial numbers are limited [2]. Visual inspection revealed no rejected trials for wavelet filtering, while manual rejection excluded numerous trials, which may be problematic in studies with fewer trials.

Overall, wavelet filtering proved effective in improving HbO signal quality during swallowing, while SD-channel regression showed potential for enhancing signal quality

without compromising data. Future studies should consider applying SD-channels alongside wavelet filtering to further refine NIRS signals in swallowing research [11, 14].

CONCLUSION

In summary, the present study successfully replicated earlier findings on ME and MI of swallowing [1–3], while applying automatic motion artifact correction methods. Hemodynamic signal changes during ME and MI were largely comparable in their topographical distribution and exhibited a prolonged time course. However, differences in HbR concentration changes between ME and MI suggest that motion artifacts may have influenced results in previous studies to some extent [1, 2].

Regarding motion artifact correction, wavelet filtering reduced HbO concentration changes, indicating a decrease in motion artifacts, while no significant effects of SD-channel regression were observed. Future research should explore different measures and parameter settings, such as the iqr, to optimize correction methods. In conclusion, the study replicated most earlier findings on ME and MI of swallowing [1, 2]. However, the observed differences between manual artifact rejection and wavelet filtering highlight the critical importance of selecting appropriate motion artifact correction methods in NIRS studies investigating swallowing.

REFERENCES

- [1] Kober SE, Wood G. Changes in hemodynamic signals accompanying motor imagery and motor execution of swallowing: A near-infrared spectroscopy study. *NeuroImage* 2014;93:1–10.
- [2] Kober SE, Bauernfeind G, Woller C, Sampl M, Grieshofer P, Neuper C et al. Hemodynamic Signal Changes Accompanying Execution and Imagery of Swallowing in Patients with Dysphagia: A Multiple Single-Case Near-Infrared Spectroscopy Study. *Front. Neurol* 2015;6:1–10.
- [3] Kober SE, Gressenberger B, Kurzmann J, Neuper C, Wood G. Voluntary modulation of hemodynamic responses in swallowing related motor areas: A near-infrared spectroscopy based neurofeedback study. *PLoS ONE* 2015;10(11):1–17.
- [4] Kober SE, Hinterleitner V, Bauernfeind G, Neuper C, Wood G. Trainability of hemodynamic parameters: A near-infrared spectroscopy based neurofeedback study. *Biological psychology* 2018;136:168–80.
- [5] Kober SE, Grössinger D, Wood G. Effects of Motor Imagery and Visual Neurofeedback on Activation in the Swallowing Network: A Real-Time fMRI Study. *Dysphagia* 2019.
- [6] Kober SE, Weiss S, Grieshofer P, Wood G. NIRS-BASED NEUROFEEDBACK TRAINING TO TREAT DYSPHAGIA: Verlag der Technischen Universität Graz; 2019.

- [7] Kirilina E, Yu N, Jelzow A, Wabnitz H, Jacobs AM, Tachtsidis I. Identifying and quantifying main components of physiological noise in functional near infrared spectroscopy on the prefrontal cortex. *Frontiers in Human Neuroscience* 2013;7:864.
- [8] Brigadoi S, Ceccherini L, Cutini S, Scarpa F, Scatturin P, Selb J et al. Motion artifacts in functional near-infrared spectroscopy: a comparison of motion correction techniques applied to real cognitive data. *NeuroImage* 2014;85 Pt 1(0 1):181–91.
- [9] Gagnon L, Yücel MA, Boas DA, Cooper RJ. Further improvement in reducing superficial contamination in NIRS using double short separation measurements. *NeuroImage* 2014;85 Pt 1(0 1):127–35.
- [10] Jahani S, Setarehdan SK, Boas DA, Yücel MA. Motion artifact detection and correction in functional near-infrared spectroscopy: a new hybrid method based on spline interpolation method and Savitzky-Golay filtering. *Neurophotonics* 2018;5(1):15003.
- [11] Yücel MA, Selb J, Boas DA, Cash SS, Cooper RJ. Reducing motion artifacts for long-term clinical NIRS monitoring using collodion-fixed prism-based optical fibers. *NeuroImage* 2014;85 Pt 1(0 1):192–201.
- [12] Schecklmann M, Mann A, Langguth B, Ehlis A-C, Fallgatter AJ, Haeussinger FB. The Temporal Muscle of the Head Can Cause Artifacts in Optical Imaging Studies with Functional Near-Infrared Spectroscopy. *Front. Hum. Neurosci* 2017;11:456.
- [13] Kober SE, Wood G. Hemodynamic signal changes during saliva and water swallowing: a near-infrared spectroscopy study 2018;23 IS -:015009-23-7.
- [14] Yücel MA, Lühmann AV, Scholkmann F, Gervain J, Dan I, Ayaz H et al. Best practices for fNIRS publications. *Neurophotonics* 2021;8(1):12101.
- [15] Chen W-L, Wagner J, Heugel N, Sugar J, Lee Y-W, Conant L et al. Functional Near-Infrared Spectroscopy and Its Clinical Application in the Field of Neuroscience: Advances and Future Directions. *Front. Neurosci.* 2020;14:724.
- [16] Gagnon L, Perdue K, Greve DN, Goldenholz D, Kaskhedikar G, Boas DA. Improved recovery of the hemodynamic response in diffuse optical imaging using short optode separations and state-space modeling. *NeuroImage* 2011;56(3):1362–71.
- [17] Saager R, Berger A. Measurement of layer-like hemodynamic trends in scalp and cortex: implications for physiological baseline suppression in functional near-infrared spectroscopy. *JBO* 2008;13(3):34017.
- [18] Brigadoi S, Cooper RJ. How short is short? Optimum source-detector distance for short-separation channels in functional near-infrared spectroscopy. *Neurophotonics* 2015;2(2):25005.
- [19] Raggam P, Bauernfeind G, Wriessneger SC. NICA: A Novel Toolbox for Near-Infrared Spectroscopy Calculations and Analyses. *Front. Neuroinform.* 2020;14:26.
- [20] Molavi B, Dumont GA. Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological measurement* 2012;33(2):259–70.
- [21] Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H et al. PsychoPy2: Experiments in behavior made easy. *Behavior research methods* 2019;51(1):195–203.
- [22] Aasted CM, Yücel MA, Cooper RJ, Dubb J, Tsuzuki D, Becerra L et al. Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. *Neurophotonics* 2015;2(2):20801.
- [23] Singh AK, Dan I. Exploring the false discovery rate in multichannel NIRS. *NeuroImage* 2006;33(2):542–9.
- [24] Hétu S, Grégoire M, Saimpont A, Coll M-P, Eugène F, Michon P-E et al. The neural network of motor imagery: an ALE meta-analysis. *Neuroscience and biobehavioral reviews* 2013;37(5):930–49.
- [25] Guillot A, Di Rienzo F, MacIntyre T, Moran A, Collet C. Imagining is Not Doing but Involves Specific Motor Commands: A Review of Experimental Data Related to Motor Inhibition. *Front. Hum. Neurosci* 2012;6.
- [26] Wriessneger SC, Kurzmann J, Neuper C. Spatio-temporal differences in brain oxygenation between movement execution and imagery: A multichannel near-infrared spectroscopy study. *International Journal of Psychophysiology* 2008;67(1):54–63.
- [27] Batula AM, Mark JA, Kim YE, Ayaz H. Comparison of Brain Activation during Motor Imagery and Motor Movement Using fNIRS. *Comput Intell Neurosci* 2017;2017:5491296.
- [28] Hamdy S, Mikulis D, Crawley A, Xue S, Law H, Henry S et al. Cortical activation during human volitional swallowing: an event related fMRI study. *Am J Physiol Gastrointest Liver Physiol* 1999;277:219–25.
- [29] Sörös P, Inamoto Y, Martin RE. Functional brain imaging of swallowing: An activation likelihood estimation meta-analysis. *Hum. Brain Mapp* 2009;30(8):2426–39.
- [30] Martin REGB, Gati J, Menon R. Cerebral Cortical Representation of Automatic and Volitional Swallowing in Humans. *Journal of Neurophysiology* 2001;85:938–50.
- [31] Zanato LE, Chiari B, Vieira M, Bommarito S. Study of the electrical activity of muscles: Masseter, temporal and supra-hyoid during swallowing. *Dent Oral Craniofac Res* 2016;3(1).
- [32] Di Lorenzo R, Pirazzoli L, Blasi A, Bulgarelli C, Hakuno Y, Minagawa Y et al. Recommendations for motion correction of infant fNIRS data applicable to multiple data sets and acquisition systems. *NeuroImage* 2019;200:511–27.

Inter-Expert Variability in Common Spatial Pattern Component Selection: Implication for Neurofeedback Applications

C. Dumas¹, C. Dussard¹, M.-C. Corsi^{1*}, N. George^{1*}

¹ Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, Inserm, CNRS, AP-HP, INRIA, CENIR MEG-EEG & NERV lab & Experimental Neurosurgery team, Hôpital de la Pitié Salpêtrière, Paris, France

* co-last authors

E-mail: Cassandra.dumas@icm-institute.org

ABSTRACT: Common Spatial Patterns (CSP) are widely used in motor imagery (MI)-based brain-computer interfaces (BCIs). While classification pipelines combine multiple components to optimize performance, MI-based neurofeedback (MI-NF) requires selecting a single spatial filter to drive feedback based on physiologically interpretable sensorimotor activity. This critical step is often implicit and relies on expert judgment. We quantified inter-expert variability in CSP selection using data from 20 subjects in a right-hand MI-NF experiment. Twenty-three BCI or neurophysiology experts independently selected the most physiologically relevant component among six CSP candidates per subject. Inter-expert agreement was fair overall (Fleiss' $\kappa = 0.256$) but varied substantially across subjects. Consensus did not systematically favor the top-ranked component, and selections spanned all six candidates. Although higher-ranked components elicited stronger agreement, substantial variability persisted. These findings identify single-component CSP selection as an overlooked source of methodological heterogeneity in MI-NF systems, highlighting the need for explicit and reproducible selection criteria.

INTRODUCTION

Common Spatial Pattern (CSP) decomposition is a widely used spatial filtering method in motor imagery (MI)-based brain-computer interfaces (BCIs) with electroencephalography (EEG) [1]. By maximizing variance differences between conditions, CSP enhances class separability. In classification-oriented BCIs, 2 or 3 pairs of components are typically selected to optimize discriminative performance, and multiple filters may be combined within machine learning pipelines [2]. In MI-based neurofeedback (MI-NF), however, the brain activity driving the feedback must reflect a physiologically interpretable neural process that subjects learn to voluntarily modulate [3]. The objective therefore extends beyond statistical discrimination to the extraction of a meaningful and controllable activity, typically expressed as event-related desynchronization (ERD) of the sensorimotor rhythm [4]. Traditional MI-NF commonly relies on predefined spatial filters, such as Laplacian derivations centered on C3/C4 electrodes, applied uniformly across subjects [3].

Although this ensures anatomical interpretability and methodological consistency, it overlooks substantial inter-individual variability in sensorimotor rhythm topography. CSP offers a subject-specific alternative. Yet, it typically yields multiple candidate components with high discriminative power. This creates the critical methodological challenge of determining which single component should drive feedback.

In practice, CSP selection may rely on quantitative criteria, including eigenvalues, mutual information, or cross-validation accuracy [5], or on qualitative neurophysiological judgment, such as the presence of focal sensorimotor activity consistent with motor cortex anatomy and known to be associated with MI [6]. However, identifying the most relevant component typically involves expert judgment, particularly when several components exhibit comparable discriminability or plausible spatial topography. This introduces a non-standardized, expert-dependent decision into system configuration.

Such expert-dependent selection may have broad implications. The selected spatial filter directly determines the neural signal fed back to the participant, thereby shaping the feedback representation and potentially influencing learning dynamics. Variability in expert selection translates into variability in feedback signals across studies, laboratories, and experimenters or clinicians, constituting an unrecognized source of methodological heterogeneity.

Previous research has examined numerous factors influencing MI-BCI performance, including signal processing choices, feedback modality, task instructions, and experimenter-participant interactions [8]. However, the influence of expert decisions on the internal configuration of NF systems remains largely unexplored. No study to date has systematically quantified inter-expert variability in CSP component selection for individual subjects in MI-NF.

The present study addresses this gap by characterizing agreement among 23 BCI and neurophysiology experts in selecting CSP components for right-hand MI-NF in 20 subjects. By quantifying inter-expert variability, this study empirically investigates the extent to which CSP selection reflects shared expert consensus versus expert-dependent methodological choice in face of inter-subject variability in MI-related ERD and CSPs.

MATERIALS AND METHODS

EEG Dataset

We analyzed the previously recorded EEG data from 20 healthy subjects (11 female; mean age $\pm$ SD: 25.8 ± 5.8 years). These data were acquired during a MI-NF CSP calibration protocol [9].

EEG was recorded with an actiCHamp Plus system (Brain Products GmbH) using a 32-active-electrode cap (ActiCAP snap) positioned according to the international 10–20 system. EEG signals were sampled at 1 kHz with a hardware band-pass filter of DC–280 Hz, referenced to Fz, with the ground at Fpz. Four electrodes (TP9, FT9, TP10, FT10) were excluded due to frequent artifact contamination, resulting in a final montage of 28 electrodes.

The calibration protocol included two conditions (rest and right-hand MI), each comprising 10 randomized trials of 8 s. During MI trials, subjects repeatedly imagined opening and closing their right hand while a virtual hand displayed similar movements on a computer screen, independent of neural activity.

CSP computation

To generate candidate spatial filters, EEG signals were notch-filtered at 50 Hz and band-pass filtered between 8 and 30 Hz, corresponding to the sensorimotor rhythm frequency band. No further artifact correction was applied, consistent with classical online BCI usage. CSP filters were estimated using the MNE-Python implementation [10]. Components were ranked according to their mutual information with class labels, and the six highest-rank CSP components were retained for each participant. For neurophysiological interpretation, CSP filters were converted to spatial patterns by projection into sensor space [2].

Expert Survey and CSP Selection Task

An anonymous online questionnaire was disseminated via learned societies, mailing lists, and specialized online communities. The call explicitly targeted researchers with expertise in either BCI or human electrophysiology; participation was voluntary and based on self-declared expertise. Twenty-three experts completed the survey. Non-BCI experts ($n=14$) were researchers with expertise in human EEG but without specific expertise in BCI systems. Among BCI experts, 6 out of 9 specifically worked with MI-based paradigms.

The questionnaire first collected background information, including field of expertise (open-ended), years of professional experience (< 5 years, 5–10 years, > 10 years), and self-rated familiarity with the CSP methodology (none, limited, moderate, high).

Experts then completed a CSP selection task for each of the 20 study subjects. The data for each subject were presented in a fixed randomized order identical across experts, and experts were blind to subjects' identifiers and characteristics. For each subject, experts viewed a composite figure including (i) a baseline-corrected ERD

topographical map computed from right-hand MI trials (power log-ratio relative to a -3 to -1 s baseline, averaged over 8–30 Hz and 0–8 s post-stimulus), and (ii) six candidate CSP components labeled A–F, corresponding to the six components ranked 1 to 6 by mutual information.

Experts were instructed to select the single CSP pattern they considered the most physiologically relevant for an online right-hand MI-NF application and to rate their confidence on a 5-point Likert scale (1 = very low confidence, 5 = very high confidence). Experts also indicated for each participant whether one CSP clearly stood out, several appeared equally plausible, or none appeared satisfactory. Instructions specified that CSP polarity is arbitrary and that no selection was considered correct or incorrect. Although CSP labels A–F reflected mutual-information ranking, the ranking criterion itself and its order were not explicitly disclosed, in order to encourage selections primarily based on neurophysiological interpretation rather than algorithmic metrics.

Inter-Expert Agreement and Association with CSP Rank

For each of the 20 subjects, we computed two metrics to quantify expert consensus across the six CSP candidates. First, inter-expert agreement was defined as the proportion of experts selecting the modal CSP (i.e., the most frequently selected component), ranging from 0 (no consensus) to 1 (perfect consensus). Second, choice dispersion was quantified using normalized Shannon entropy over the six CSP choices, ranging from 0 (perfect consensus) to 1 (maximal disagreement: uniform distribution across all six CSPs) [11].

Overall expert agreement across all 20 subjects was quantified using Fleiss' kappa (23 experts, 20 subjects, 6 CSPs), which measures inter-rater agreement beyond chance [12]. Kappa values were interpreted according to standard conventions: < 0 (no agreement), 0–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), 0.81–1.00 (almost perfect). Ninety-five percent confidence intervals were estimated using non-parametric bootstrap resampling across subjects (10,000 iterations).

To examine whether consensus strength was related to the CSP rank, we computed, for each subject, the mean rank of the CSP selected by experts (rank = 1–6 according to mutual information). The association between mean selected rank and inter-expert agreement (proportion of modal CSP selection) was evaluated using Spearman correlation. We checked that agreement was not driven by CSP-unfamiliar experts by re-calculating the modal CSP after excluding the experts that reported no familiarity with CSP ($n = 7$).

Spatial Consistency of Selected CSP Patterns

To assess the spatial consistency of expert selections, the CSP components selected by each expert for each subject were normalized to unit Euclidean norm (L2 normalization) and sign-aligned so that the C3 electrode

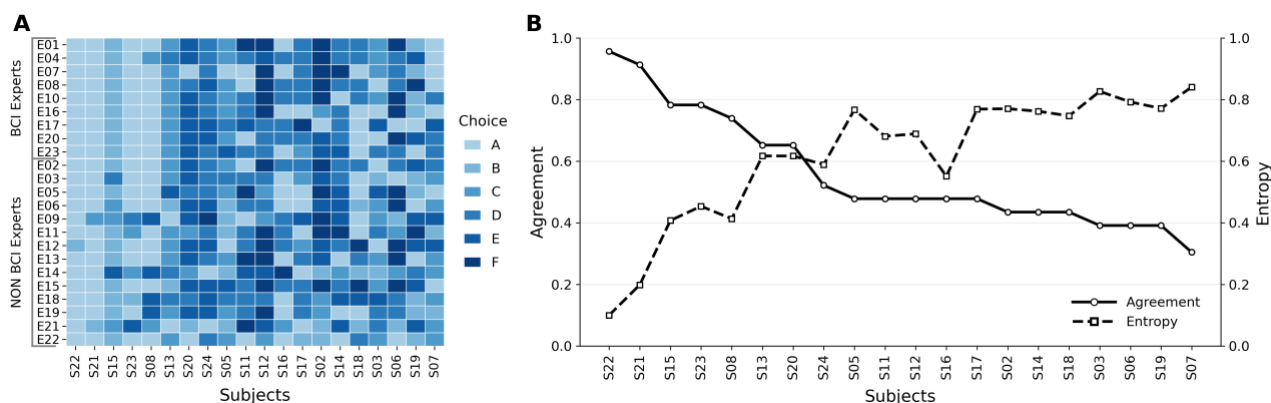


Figure 1: Inter-expert agreement on CSP selection. (A) Colormap showing the experts' (E01, E04, ... E22) component selections (A–F) for each subject (S22, S21, ... S07). Subjects are ordered from highest to lowest consensus (left to right columns). Experts are grouped according to their domain of expertise: BCI experts (top rows) and non-BCI experts (bottom rows). Colors represent the selected CSP component for each expert–subject pair. (B) Inter-expert agreement (in solid black line) and normalized Shannon entropy (in dashed black line) for each subject (in the same order as in (A)). The left y-axis represents agreement (modal selection proportion; 1 = full agreement, 0 = no agreement). The right y-axis represents entropy (0 = no dispersion, 1 = maximal dispersion across choices).

exhibited negative values, consistent with expected sensorimotor desynchronization [4]. For each subject, the mean and standard deviation (SD) of CSP values across experts, on each electrode, were computed to obtain topographical maps of inter-expert variability. Similarly, for each expert, the mean and SD of selected CSP values on each electrode across subjects were computed to assess within-expert spatial variability in CSP selection.

Influence of Expert Characteristics

At the expert level, we examined whether selection behavior was associated with research background, professional experience, and familiarity with the CSP methodology. For each expert, two outcome variables were computed: (i) individual agreement, defined as the proportion of selections matching the modal (consensus) CSP across the 20 subjects, and (ii) mean confidence, defined as the average confidence rating across all selections.

Experts were classified according to research background (BCI vs non-BCI) and group differences in expert agreement and mean confidence were evaluated using independent samples t-tests. Professional experience was categorized into two groups (<5 years vs ≥5 years) to ensure balanced group sizes, and group differences in agreement and confidence were assessed using independent samples t-tests. Familiarity with CSP was categorized into three ordered levels (none, limited, moderate–high), with moderate and high familiarity pooled together to ensure balanced group sizes. A linear regression analysis was conducted treating familiarity as an ordinal predictor to test for a linear trend in agreement and confidence across increasing familiarity levels.

Effect sizes were reported using Cohen's d for t-tests. Assumptions of normality (Shapiro–Wilk test) and homogeneity of variance (Levene's test) were verified prior to parametric testing and non-parametric tests were used as robustness checks where appropriate. The significance threshold was set at $\alpha = 0.05$. All analyses

were performed in Python 3.12. The analysis code is publicly available at: https://github.com/Dumas-C/CSP_Inter-Expert_Variability_Neurofeedback.

RESULTS

Inter-Expert Agreement and Association with CSP Rank

Across the 23 experts and 20 subjects, Fleiss' kappa indicated fair inter-rater agreement ($\kappa = 0.256$, 95% bootstrap CI [0.157–0.334]) according to standard benchmarks. At the subject level, inter-expert agreement (proportion of modal CSP selection) was on average 0.559 ± 0.187 , and normalized Shannon entropy was 0.618 ± 0.208 , indicating substantial variability in expert consensus strength across subjects (Fig. 1).

A significant negative association was observed between mean selected CSP rank and inter-expert agreement (Spearman $\rho = -0.48$, $p = 0.031$), indicating that stronger consensus tended to occur when experts converged on higher-rank CSP components.

Regarding decision clarity, experts indicated that one CSP clearly stood out for 6 subjects, several CSPs appeared equally plausible for 8 subjects, and no CSP appeared satisfactory for 6 subjects.

Spatial Consistency of Selected CSP Patterns

The subjects for whom inter-expert agreement was high were also those for whom SD across experts was low over the whole scalp, whereas subjects yielding lower inter-expert agreement were associated with larger variability (SD) across experts.

Importantly, this variability was not uniformly distributed but concentrated in frontal regions (Fig. 2A). At the subject level, inter-expert variability was approximately twofold higher over frontal electrodes (Fp1: 0.225 ± 0.100 ; Fp2: 0.190 ± 0.100) than over C3 (0.106 ± 0.043 ; mean $\pm$ SD across subjects).

At the expert level, selected CSP components were globally consistent across subjects, with mean spatial patterns predominantly centered over the left sensorimotor cortex. However, intra-expert variability across subjects was again particularly high over frontal electrodes (Fp1: 0.242 ± 0.082 ; Fp2: 0.228 ± 0.063 ; for comparison: C3: 0.155 ± 0.019 ; mean $\pm$ SD across experts), particularly among experts showing low agreement with the modal selection (Fig. 2B).

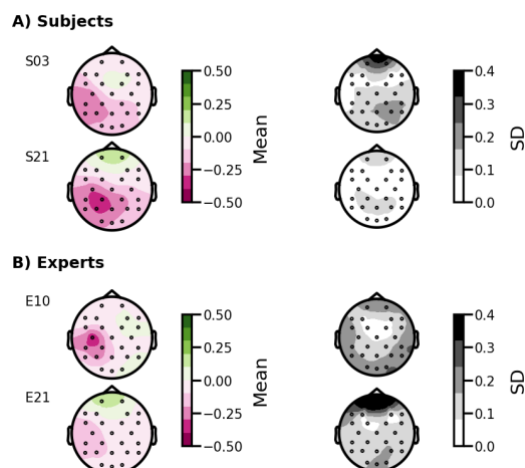


Figure 2: Topographical representation of the selected CSPs. The CSPs are displayed using a green–pink color scale, with green indicating positive weights and pink indicating negative weights. The variability (SD) of the selected CSPs is represented in grey maps. (A) Mean and SD of the CSPs selected by the experts for two representative subjects (S03 and S21). (B) Mean and SD of the CSPs selected by two representative experts (E10 and E21) across all subjects. This figure illustrates both inter-subject and inter-expert variability in the spatial distribution of the selected CSPs.

Influence of Expert Characteristics

Experts with BCI-related backgrounds ($n = 9$) demonstrated significantly greater individual agreement with the modal selection ($t(21) = 2.48$, $p = 0.022$, $d = 1.03$) and higher mean confidence ($t(21) = 2.87$, $p = 0.009$, $d = 1.20$) relative to non-BCI experts ($n = 14$). These effects were confirmed using non-parametric tests. Note, however, that the two groups differed in size and were small, which limits the statistical power of this comparison. Professional experience was categorized into two groups (<5 years, $n = 13$; ≥ 5 years, $n = 10$). No significant difference between these groups was observed for mean confidence ($t(21) = -0.88$, $p = 0.390$, $d = -0.38$). Individual agreement showed a non-significant trend toward lower values in less experienced experts ($t(21) = -2.01$, $p = 0.057$, $d = -0.86$). Familiarity with CSP technique was treated as an ordinal variable with three levels (none, $n = 7$; limited, $n = 9$; moderate/high, $n = 7$). Familiarity showed a significant positive linear trend with both individual agreement ($p = 0.037$) and confidence ($p = 0.037$) (Fig. 3).

When excluding CSP-unfamiliar experts ($n = 7$), the

modal CSP selection remained identical in 19 of 20 subjects (95%) ruling out that these CSP-unfamiliar experts could account for the results.

DISCUSSION

CSP is a widely used method in BCI applications, which enables defining individualized spatial features for feedback. While it may help refining the targeting of neural activity in MI-NF, the use of CSP in MI-NF requires selecting a single physiologically interpretable component [2]. This selection constitutes a critical methodological decision, as the chosen filter determines which neural process participants are trained to modulate. The present study quantifies inter-expert variability in CSP component selection to evaluate the level of expert consensus and its potential implications for MI-NF.

Fleiss' kappa indicated fair agreement ($\kappa = 0.256$), demonstrating that single-component CSP selection is neither trivial nor fully standardized. Importantly, in multi-category settings involving six response options and 23 independent raters, κ statistics are known to be conservative and sensitive to marginal distributions [13]. Thus, the observed value reflects substantial, structured agreement well beyond chance. At the subject level, expert agreement strength varied markedly: Some subjects elicited strong convergence, but most generated distributed selections across two or three plausible candidates, leading to a level of inter-expert agreement below .5 (less than half the experts selecting the modal component) in a majority of subjects. Normalized entropy analysis further showed that identical modal agreement proportions could mask distinct underlying selection structures. Together, these findings suggest that CSP selection operates as a semi-structured expert decision guided by partially shared evaluation principles rather than as a deterministic outcome of signal processing.

Inter-expert agreement showed a negative association with the mean selected CSP rank. This indicates that stronger consensus tended to emerge when experts converged on higher-rank components. This supports partial alignment between algorithmic discriminability and physiological plausibility. In subjects exhibiting clear contralateral sensorimotor modulation, the top-ranked CSP component appeared to coincide with physiologically meaningful activity [1]. However, this alignment was not systematic. In several cases, experts preferentially selected lower-ranked components, which had actually a reduced discriminative value between MI and rest in terms of mutual information, a measure often used for automatic selection of CSP in BCI settings. This behavior likely reflects situations in which top-rank components captured discriminative variance that was spatially diffuse or physiologically ambiguous. Rather than indicating a limitation of the CSP algorithm, these findings suggest that algorithmic ranking does not necessarily coincide with interpretability constraints

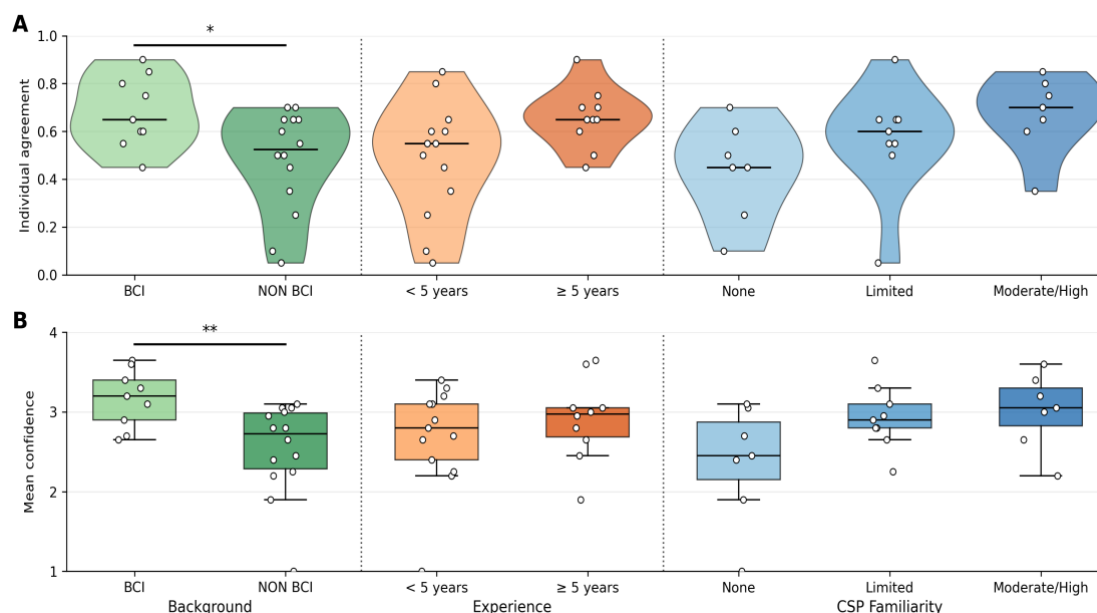


Figure 3: Influence of expert characteristics on their individual agreement and mean confidence. The individual expert agreement with the consensus (aka. modal CSP selection) and the mean confidence is compared across three expert characteristics: background (BCI vs. non-BCI, left plot in green), years of experience (< 5 vs. ≥ 5 years, middle plot in orange), and CSP familiarity (none, limited, moderate-to-high, right plot in blue). In each plot, white dots represent individual expert data points. (A) Individual expert agreement, defined as the proportion of subjects for whom an expert’s selected the consensus (aka. modal) CSP component. Violin plots show the distribution within each expert group; the central black line represents the median. (B) Mean confidence ratings of the experts. Boxplots display the interquartile range (Q1–Q3), the median (central black line), and whiskers extend to 1.5× IQR (with outliers excluded).

specific to NF applications.

Spatial consistency analysis further clarified sources of disagreement: Variability was predominantly concentrated over frontal regions. These regions are notably associated with ocular artifacts CSP decomposition is known to be sensitive to variance-dominant sources, including artifacts [14]. Although the present study did not directly validate artifact contributions, the observed spatial distribution is consistent with the hypothesis that ambiguous or artifact-prone components contribute to interpretative divergence. This raises practical considerations for real-time NF implementation. Incorporating online artifact correction strategies or physiologically constrained spatial filtering approaches may enhance the specificity of extracted components and increase inter-expert consensus [15]. Future studies should evaluate preprocessing effects on CSP selection variability.

Experts with BCI-focused research backgrounds reported higher confidence and demonstrated greater alignment with modal selections. This difference may reflect greater familiarity with MI datasets and CSP-derived spatial representations rather than superior neurophysiological knowledge. Treating familiarity as an ordered variable revealed a significant positive linear trend for both individual agreement and confidence. This indicates that greater familiarity with CSP is associated with more consistent and confident selection behavior. However, even among the experts familiar with CSPs, substantial variability persisted. When excluding CSP-unfamiliar experts, modal selections remained identical

in 19 of 20 subjects, confirming that observed variability was not driven by methodological inexperience. Moreover, experts reported that either several CSPs appeared equally plausible or none appeared satisfactory for 14 of 20 subjects, indicating that selection frequently involved navigating ambiguous candidates rather than identifying a single unequivocal solution. These findings have implications for methodological transparency and reproducibility in MI-NF studies. In practice, different laboratories may implement similar systems yet select distinct spatial filters, thereby training subjects on different neural signals. Whether such configuration differences influence learning and/or behavioral outcomes remains an open empirical question. It is noteworthy that, although CSP is widely used in MI-based classification BCIs, its use as a single spatial filter to directly drive neurofeedback remains relatively uncommon [16]. Most MI-NF paradigms rely on predefined anatomical filters. The need to select one physiologically interpretable component among several discriminative candidates may partly explain this limited adoption. Nonetheless, in a domain already characterized by substantial inter-individual variability and NF/BCI inefficiency [7], unreported variability in expert-driven system configuration may contribute to inconsistent outcomes across studies. Systematic reporting of CSP ranks, selection rationale, spatial topographies, time interval between calibration and real time application, and, when possible, open sharing of spatial filters would represent a minimal yet important step toward improving transparency and cross-study comparability.

A limitation of the present study is that it characterizes variability in expert choices without directly evaluating the functional consequences of these choices on NF performance. While we demonstrate substantial inter-expert variability in spatial filter selection, we do not assess if different CSP selections would lead to differences in subjects' learning or behavioral outcomes. It remains possible that multiple candidate components, despite spatial differences, yield comparable NF efficacy. Conversely, it is also plausible that certain selections enhance controllability or engagement of task-relevant neural processes. Determining whether inter-expert variability translates into meaningful differences in NF performance constitutes a critical next step.

More broadly, the observed pattern of structured but incomplete agreement suggests that experts rely on partially shared interpretative criteria when evaluating CSP components. Formalizing these criteria into explicit, reproducible guidelines or developing hybrid algorithmic approaches that integrate discriminability with physiologically informed constraints may help reduce latent configuration variability in MI-based NF / BCI systems. In particular, data-driven approaches such as transfer learning could reduce reliance on manual expert selection by leveraging across-subject reliable priors to identify physiologically relevant components more robustly. Such efforts could strengthen interpretability, reproducibility, and translational robustness of CSP-based NF approaches.

CONCLUSION

This study demonstrates that single-component CSP selection in MI-NF is not fully standardized and yields only fair inter-expert agreement. Although consensus often emerges for higher-rank components, substantial variability persists, particularly in spatially ambiguous cases. These findings identify CSP selection as a source of configuration-level variability in MI-NF systems. Implementing on-line physiological (ocular) artifact correction may help reduce this variability. Formalizing or automating the CSP selection step, and reporting selected components, would contribute to transparency and reproducibility in NF and BCI research.

ACKNOWLEDGEMENTS

This work was supported by the French National Research Agency (BETAPARK project, ANR-20-CE37-0012). The authors thank the BCI and neurophysiology experts who participated in the survey for their valuable contribution to this study.

REFERENCES

[1] Ramoser H, Müller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans Rehabil Eng.* 2000;8(4): 441-446
[2] Blankertz B, Tomioka R, Lemm S, Kawanabe M,

Müller KR. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process Mag.* 2008;25(1): 41-56
[3] Marzbani H, Marateb HR, Mansourian M. Neurofeedback: a comprehensive review on system design, methodology and clinical applications. *Basic Clin Neurosci.* 2016;7(2): 143-158
[4] Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol.* 1999;110(11): 1842-1857
[5] Ang KK, Chin ZY, Wang C, Guan C, Zhang H. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Front Neurosci.* 2012;6: 39
[6] Bennett JD, John SE, Grayden DB, Burkitt AN. A neurophysiological approach to spatial filter selection for adaptive brain-computer interfaces. *J Neural Eng.* 2021;18(2): 026006
[7] Zhang R, Li F, Zhang T, Yao D, Xu P. Subject inefficiency phenomenon of motor imagery brain-computer interface: influence factors and potential solutions. *Brain Sci Adv.* 2021;6(3): 224-241
[8] Pillette L, Roc A, N'Kaoua B, Lotte F. Experimenters' influence on mental-imagery based brain-computer interface user training. *Int J Hum Comput Stud.* 2021;149: 102603
[9] Dussard C, Pillette L, Dumas C, Pierrieau E, Hugueville L, Lau B, et al. Influence of feedback transparency on motor imagery neurofeedback performance: the contribution of agency. *J Neural Eng.* 2024;21(5): 056029
[10] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. *Front Neuroinform.* 2013;7: 267
[11] Shannon CE. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27(3): 379-423
[12] Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5): 378-382
[13] Almehrizi RS. Coefficient Lambda for Interrater Agreement Among Multiple Raters: Correction for Category Prevalence. *Educ Psychol Meas.* 2025;00131644251380540
[14] Yong X, Ward RK, Birch GE. Robust common spatial patterns for EEG signal preprocessing. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vancouver (BC), Canada; 2008. p. 2087-2090
[15] Dumas C, Corsi MC, Dussard C, Grosselin F, George N. Automatic ocular artifact correction in electroencephalography for neurofeedback. In: Proceedings of the 18th International Conference on Bio-inspired Systems and Signal Processing; 2025 Mar; Porto, Portugal
[16] Lioi G, Butet S, Fleury M, Bannier E, Lécuyer A, Bonan I, et al. A multi-target motor imagery training using bimodal EEG-fMRI neurofeedback: A pilot study in chronic stroke patients. *Front Hum Neurosci.* 2020;14:37

Resting-State EEG Predictors for BCI Performance: A Scoping Review

T. Settgast¹, R. Patel², A. Kübler¹

¹ Institute of Psychology, Section Intervention Psychology, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

² Electronic and Electrical Engineering, University College London, London, United Kingdom

E-mail: tomko.settgast@uni-wuerzburg.de

ABSTRACT: This scoping review summarizes studies investigating the association between resting-state EEG and subsequent BCI performance. Following PRISMA-ScR guidelines, we systematically mapped the literature to provide a comprehensive overview. We discuss the results within the context of BCI performance prediction, emphasizing the importance of resting-state brain activity, particularly for patient-centered research.

INTRODUCTION

Resting-state brain activity influences subsequent stimulus-induced activity, and vice versa [1,2]. Additionally, resting-state brain activity relates to subsequent task performance [3,4]. Therefore, resting-state brain activity contains valuable information about a subject's cognitive state [5] that is relevant to stimulus processing and task performance. Recently, there have been calls to recognize this user's state [6] and to use resting-state dynamics to optimize brain-computer interface (BCI) use [7].

This scoping review maps experimental studies relating resting-state electroencephalography (EEG) to subsequent BCI performance, discussing these results in the broader context of BCI performance predictors. We emphasize the potential of resting-state brain activity for BCI use, particularly for patient groups who could greatly benefit from BCIs, such as those in the locked-in (LIS) or complete locked-in state (CLIS) [8].

MATERIALS AND METHODS

This review follows the *PRISMA-ScR* guidelines [9]. We searched *Web of Science* and *IEEE Xplore* for experimental studies linking resting-state EEG to subsequent BCI performance (last search: October 16, 2025), using the search query term “(BCI OR brain-computer interface OR brain-machine interface OR direct neural interface OR thought-translation device OR speller) AND (correlate OR predictor) AND resting-state”, and restricted the search to the title, abstract, and keywords. The review protocol was not pre-registered. Inclusion was restricted to active BCI paradigms; passive BCIs were excluded. Resting-state was defined as a passive acquisition state, with the eyes open or closed [5]. To ensure signal stability [10], we excluded studies with recordings shorter than 30 seconds, e.g., pre-stimulus

activity, and (concatenated) between/inter-trial resting epochs. We also excluded studies that used *rest* merely as a control condition, as well as those that made resting-state recordings exclusively after BCI use or between BCI blocks. Records were divided between two authors (TS and RP) for screening (no overlap). Uncertain cases were discussed jointly. Data extraction included resting-state EEG feature(s), the resting-state protocol, the BCI paradigm, participant demographics, and BCI performance associations (see Tab. 1).

RESULTS

Database searches identified 372 records (*IEEE Xplore*: N = 86; *Web of Science*: N = 286). Seven papers were added manually. Four of these papers were known to the authors but were not retrieved by the screening strategy. The other three were found through reference checking. After screening and an eligibility assessment (see Fig. 1 for details), 23 studies were included in the review (see Tab. 1). One study that was initially added on behalf of the authors was deemed ineligible after a thorough assessment.

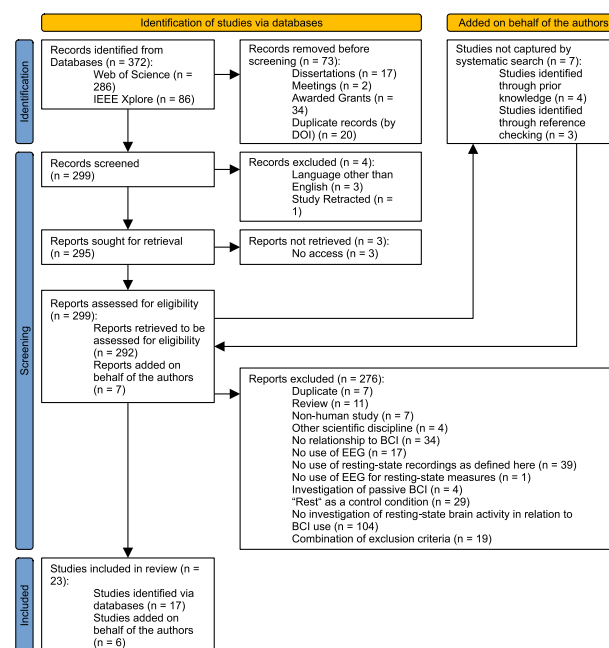


Figure 1: PRISMA flowchart of the study selection process, following the structure suggested for PRISMA reviews [11].

Overview

Table 1: Overview of reported associations between resting-state EEG features and BCI performance.

Ref.	Resting-State EEG Feature(s)	Resting-State Protocol	BCI Paradigm	Participants	Results (Association with BCI Performance)
[12]	<i>SMR predictor</i>	EO/EC (10x15s); task (EO): fix. moving shapes. Anal.: concatenated EO.	SMR (MI) BCI	N = 148/168 (H)	<i>SMR predictor</i> $\uparrow$ ($r = .53$; $r = .66$ without outlier).
[13]	RPL of canonical frequency bands (except δ) & <i>PPfactor</i>	"Non-task related state", EO; task: mind-wandering, think about nothing. Duration: NR.	MI BCI	N = 52 (H) + N = 9 (H), Data: [14]	High performer: low θ (4-8 Hz) & high α (8-13 Hz) RPL. <i>PPfactor</i> $\uparrow$ ($r = .59$; $r = .70$ without outlier).
[15]	<i>SMR predictor</i>	EO/EC (10x15s); task: relax. Anal.: concatenated EO.	SMR (MI) BCI	N = 80 (H)	<i>SMR predictor</i> $\uparrow$ ($r = .53$; $r = .61$ without outlier; ~28% variance explained).
[16]	PSD (δ , θ , α , β) in four ROI	5 min EO, task: fix.dot, stay relaxed. Anal.: central 4 min.	"Visuo-mental" BCI (3 protocols)	N = 5 ALS patients + N = 6 (H)	Stepwise MLR: Temporo-parietal δ (1-3 Hz) power identified as a negative predictor for performance. Model includes H & ALS patients.
[17]	Relative PSD & IAF	EO (5x6s); task: fix. non-flickering LEDs. Anal.: concatenated EO ("baseline resting state").	SSVEP BCI (4 classes)	N = 18 (H)	<i>ACL</i> : relative "thetaHigh" (6.5-7.5 Hz) $\uparrow$ ($r = .47$), relative "betaMid" (18-24Hz) $\downarrow$ ($r = -.53$ without outlier) at Oz. IAF $\uparrow$ ($r = .58$ without outlier). <i>Top. Freq.</i> : relative "betaMid" $\uparrow$ at F3 ($r = .48$).
[18]	Coupling strength in motor network (EffC via DCM)	1 min EO; task: gaze at center. 2 sessions on different days.	MI BCI	N = 54 (H), Data: [19]	High performer: stronger SMA $\rightarrow$ DLPFC coupling. SMA $\rightarrow$ DLPFC coupling $\uparrow$ (Session 1: $r = .54$; Session 2: $r = .42$) & predicts MI BCI performance.
[20]	Network efficiency (Graph Theory Metrics: CC, CPL, GE & LE) & <i>NEI</i> derived from FC/EffC	1 min EO; task (Dataset 2): gaze at center. Dataset 2: 2 sessions on different days.	MI BCI	N = 52 (H), Data: [21] + N = 54 (H), Data: [19]	High performer: higher network efficiency (eps. α -band: 8-13 Hz). <i>NEI</i> (COH/PLV) $\uparrow$ ($r = .37/.35$ without outlier) & predicts MI performance (RMSE: 13.31%/13.37%).
[22]	Spectral features (C3/C4): fmax, at fmax: PSD/noise (1/f fit, <i>SMR Predictor</i>)	EO/EC (10x15s); task: relax. Anal.: concatenated EO.	SMR (MI) BCI	N = 78/80 (H), Data: [15]	<i>SMR predictor</i> $\uparrow$ ($r = .53$; ~28% variance explained). Further corr. at C3/C4: noise/PSD fit, fmax. Corr. from average (C3 + C4) > single channels.
[23]	<i>Criticality-related measures</i> (1-40 Hz, scalp-wide average): PLE, LZC	EO/EC (139s each), task: fix. cross, stay relaxed. Anal.: EO (before BCI, 130s).	Visual P300 BCI speller.	N = 55 (H), Data: [24]	LZC $\uparrow$ ($r = .38$), PLE $\downarrow$ ($r = -.34$).
[25]	<i>Criticality-related measures</i> (1-40 Hz, scalp-wide average): PLE, LZC	LIS-ALS: 1-2 min, EO, no task. Anal.: central 110s. H: EO/EC (139s each), task: fix. cross, stay relaxed. Anal.: EO (before BCI, central 110s).	LIS-ALS: vibrotactile P300 BCI. H: visual P300 BCI speller.	N = 1 (LIS-ALS patient, N = 10 sessions), Data: [26] + N = 55 (H), Data: [24]	LIS-ALS (longitudinal): PLE $\uparrow$ ($r = .76$), LZC $\downarrow$ ($r = -.75$). H: $\emptyset$ (ceiling effects), but opposite trends.
[27]	PSD, functional connectivity (PLV) in four ROIs (Frontal, Central, Parietal, Occipital)	4 min EC, no task.	Visual P300 BCI speller	N = 8 (H)	Frontal δ (0.5-4 Hz) power $\downarrow$ ($r = -.94$), frontal δ connectivity $\downarrow$ ($r = -.77$ to $-.88$), α (8-12 Hz) connectivity $\downarrow$ (Frontal-Central; $r = -.73$), γ (30-45 Hz) connectivity $\uparrow$ (Frontal-Occipital; $r = .77$).
[28]	PSD, RPL	EO/EC (3 min each), task: fix. cross.	MI BCI (unilateral)	N = 26 (H)	α (8-13 Hz) PSD/RPL $\downarrow$ (max. correlation with PSD at AF8: $r = -.44$), β (13-30 Hz) PSD/RPL $\uparrow$ (max correlation with PSD at Fp1: $r = .65$). Significant corr.: EO > EC.
[29]	Weighted PDFs derived from PSD	EO/EC (10x15s); task: relax. Anal.: EO + EC.	SMR (MI) BCI	N = 80/83 (H), Data: [15]	High performer: visible peaks in μ - (8-12 Hz) and β - (16-22 Hz) bands in PDFs. Linear regression based on these clusters predicts performance ($r = .58$) using only C3, Cz, and C4.
[30]	At Oz: <i>Alpha predictor</i> & slope of the 1/f PSD component (referred to as "lambda")	EO/EC (2 min each, once before & once after BCI), task: fix. cross. Anal.: EO (before BCI).	c-VEP BCI	N = 26 (H)	$\emptyset$. <i>Alpha predictor</i> non-significant trend $\downarrow$ ($r = -.18$; $r = -.29$ without outlier), & slope of the 1/f PSD component non-significant trend $\uparrow$ ($r = .13$).
[31]	<i>SMR predictor</i> , <i>PPfactor</i> & spectral peak metrics (sum of the number of peaks & variance of the peak frequency) at C3/C4 extracted).	EO/EC (3 min each), task: relax, fix. cross. Anal.: EO (2 min extracted).	MI BCI	N = 85/87 (H), Data: [32]	<i>SMR predictor</i> $\uparrow$ ($\rho = .49$), variance of the μ -band peak frequency $\downarrow$ ($\rho = -.33$), number of peaks $\uparrow$ ($\rho = .24$), <i>PPfactor</i> $\emptyset$.
[33]	μ - & β -band predictor, spectral peak dynamics in μ - (5-15 Hz) & β -band (15-25 Hz) (C3/C4): peak width/frequency, num. of peaks, (temporal) variance of amplitude/frequencies of μ -band peak (width)	3 min EO, task: relax. Anal.: 2 min extracted.	MI BCI	N = 56/59 (H), Data: [34]	μ -band predictor $\uparrow$ ($r = .43$), mean (C3 & C4) μ -band peak width $\uparrow$ ($r = .27$), number of peaks $\uparrow$ ($r = .35$), temporal variance of the μ -band peak frequency $\downarrow$ ($r = -.48$). Combined features predict performance (MAE = 12.43%).
[35]	Relative IAB amplitude at Oz	EO/EC (1 min each), no task. Anal.: EO.	SSVEP (ten classes)	N = 33 (H)	IAB amplitude $\downarrow$ (SNR: $r = -.42$; classification accuracy: $r = -.58$).
[36]	RPL & non-linear metrics (LZC & PSE) (frequency-band- and channel-wise)	EO/EC (1 min each), no task. Anal.: EO + EC (individually).	MI BCI	N = 105/109 (H), Data: [37,38]	Neurofeedback-based downregulation of IAB improves performance. Strongest correlations in α -band (8-13 Hz; EO). At C4: RPL $\uparrow$ ($r = .50$), PSE $\downarrow$ ($r = -.53$), LZC $\downarrow$ ($r = -.46$). Combined features classify high/low performer (84.05% accuracy).
[39]	Normalized occipital "alphaLow" (7.5-9 Hz) & "alphaHigh" (9-12.5 Hz) band power	1 min, no task. Anal.: last 45s. No information if eyes open/closed.	SSVEP-based QWERTY speller (30 classes)	N = 26 (H)	Correlation depends on SSVEP-stimulation frequency. Low-frequency SSVEP: "alphaLow" $\downarrow$ ($r = -.62$), "alphaHigh" $\downarrow$ ($r = -.52$).
[40]	Coupling strength in DMN (EffC via DCM)	1 min EO, no task.	MI BCI	N = 52 (H), Data: [21]	High performer: weaker ILP-rLP coupling strength. ILP-rLP coupling strength $\downarrow$ ($r = -.37$).
[41]	Mean FC (COH) & network topology (CC & CPL) (five	2 min EC, no task. Anal.: seven	SSVEP experiment (BCI)	N = 12 (H)	Significant corr. for freq.-specific RS features (excl. 10 Hz) & averages across frequencies. Corr. between averaged RS features and averaged SNR: mean FC $\downarrow$ ($r = -.80$), CC $\downarrow$ ($r = -.78$), CPL $\uparrow$ ($r = .85$). Corr. with

	frequencies: 7.5, 10, 12, 15, 20 Hz).	concatenated artifact-free 10s epochs.	equivalent) (5 classes)	classification accuracy: SNR ↑ ($r = .67$), relative "alphaHigh" (9-12.5 Hz) ↓ ($r = -.64$). Best model (normalized CC and normalized reciprocal CPL): $R^2 = .76$.
[42]	Mean SE based on relative PSD (0.5-14 Hz), <i>SMR predictor</i> & <i>ATR predictor</i>	EO/EC (2 min each, order randomized), no task. Anal.: central 100s.	SMR (MI) N = 40 (H) (N = 26 re-tested in 2 nd session)	Mean SE ↑ (maximum correlations at C3, EC): Session 1: $r = .61$; Session 2: $r = .70$. Pooled dataset (N = 66 sessions): mean SE (C3, EC) ↑ ($r = .65$), outperforming <i>ATR predictor</i> (FC3, EC; $r = .51$) and <i>SMR predictor</i> (C4, EO; $r = .29$). LDA classifier (mean SE) predicts aptitude with 75% (Session 1) and 89% (Session 2) accuracy.
[43]	Mean FC, node degree, edge strength, network efficiency (CC, CPL, LE, GE) based on COH & PLI (4-14 Hz)	EO/EC (2 min each, order randomized), no task. Anal.: central 100s.	SMR (MI) N = 40 (H) (N = 26 re-tested in 2 nd session), Data: [42]	Network efficiency (higher CC, LE, GE; lower CPL) ↑, mean FC ↑, node degree ↑, edge strength ↑. COH-based metrics outperform PLI-based metrics. Session-1 RS network efficiency (COH) predicts Session-2 classification accuracy ($r = .61$, $RMSE = 12.2\%$).

Notes: **Ref.**: numerical reference. **Resting-State Protocol**: Anal. = analysis based on. **Participants**: H = healthy participants; ALS = amyotrophic lateral sclerosis; LIS = locked-in state. **Results**: ↑/↓ = positive/negative correlation; Ø = no correlation. All reported correlations refer to correlations with performance metrics. Highlighted in light green: identified through reference checking. Highlighted in light purple: identified through prior knowledge but not retrieved by screening strategy. **Abbreviations**: ACL: assisted closed-loop protocol; CC: clustering coefficient; CPL: characteristic path length; COH: coherence; DCM: dynamic causal modeling; DLPFC: dorsolateral prefrontal cortex; EO/EC: eyes open/closed; DMN: default-mode network; EffC: effective connectivity; fmax: frequency maximizing difference between modeled PSD and noise fit; FC: functional connectivity; GE: global efficiency; IAF/LAB: individual alpha frequency/band; LE: local efficiency; LLP: left lateral parietal area; LZC: Lempel-Ziv complexity; MAE: mean absolute error; MLR: multivariate linear regression; NEI: network efficiency indicator; NR: not reported; PDF: probability density function; PLE: power-law exponent; PLI/PLV: phase-locking index/value; PPfactor: performance potential factor; (P)SE: (power) spectral entropy; RMSE: root mean square error; rLP: right lateral parietal area; RPL: relative power level; ROI: sensor-based regions of interest; SMA: supplementary motor area; STR: strength; Top. Freq.: protocol using top frequencies.

Of 23 eligible studies (six conference proceedings [23,25,27,31,33,40], 17 journal articles; 2010-2025), 14 focused on motor imagery (MI) BCIs [12,13,15,18,20,22,28,29,31,33,36,40,42,43], four on steady-state visually evoked potential (SSVEP)-based BCIs [17,35,39,41], and three on P300-based BCIs [23,25,27]. Single studies investigated a BCI using code-modulated visual evoked potentials (c-VEP) [30], and "visuo-mental" BCI protocols combining a P300 speller and an arithmetic task [16]. Participants were predominantly healthy (N = 8 to 148); only two studies included patients with amyotrophic lateral sclerosis (ALS) [16,25] – one of whom was in LIS [25].

Resting-State EEG Predictors Detailed associations between resting-state EEG features and BCI performance are provided in Tab. 1.

For MI BCIs, the *SMR predictor* – reflecting the mean peak amplitude difference between a fitted power spectral density (PSD) and a fitted 1/f noise spectrum at Laplacian channels C3 and C4 [15] – showed consistent positive correlations with performance [12,15,22,31,42]. Another C3/C4-based metric, the *performance potential factor* (PPfactor) [13], showed less consistency; a recent replication attempt [31] failed to confirm the originally reported positive correlation. In general, channels C3 and C4 were repeatedly identified as optimal recording sites for correlations with subsequent BCI performance across EEG features [22,29,33,36,42]. For example, the number of peaks in the PSD and the dynamics of the μ -frequency band peaks have repeatedly been shown to correlate with MI BCI performance [31,33]. More generally, the prominent presence of μ - and β -frequency band activity in the PSD appears to be associated with MI BCI performance. For example, Suk et al. [29] were able to cluster high and low performers based on the visibility of peaks in the μ - and β -frequency bands in the subjects' probability density functions, and, recently, Sonntag et al. [28] showed positive correlations between absolute and relative β -frequency band power and accuracy.

Associations with resting-state α -frequency band activity appear to be paradigm-specific. For MI BCIs, high performance generally relates to high α -frequency band

relative power level (RPL) [13,36] alongside low θ -frequency band RPL [13], though the inverse pattern was also observed [28]. In contrast, for SSVEP-based BCIs, (parieto-)occipital resting-state α -frequency band activity consistently correlates negatively with subsequent performance [35,39,41]. Wan et al. [35] even demonstrated that decreasing the individual α -band amplitude with neurofeedback improved performance with a SSVEP-based BCI.

Negative associations were observed between resting-state δ -frequency band power and BCI performance. In healthy participants, it negatively correlated with subsequent P300-based speller performance [27]. Similarly, temporo-parietal resting-state δ -frequency band power was retained as a significant negative predictor in a multivariate linear regression model for predicting visuo-mental BCI performance [16]. The model was trained with data from ALS patients and healthy participants.

Paradigm-specific patterns also emerged for connectivity and network features. Across studies, higher network efficiency at rest was associated with better MI BCI performance. This was demonstrated in a 4-14 Hz frequency band [43] and the α -frequency band [20]. Conversely, for SSVEP-based BCIs, mean functional connectivity and greater network efficiency across frequencies were negatively correlated with the SSVEP signal-to-noise ratio [41].

Non-linear features yielded inconclusive results. Spectral entropy exhibited both positive [42] and negative correlations [36] with MI BCI performance – the latter accompanied by negative correlations between performance and Lempel-Ziv complexity (LZC) [36] – potentially depending on the frequency bands used and eye closure during rest. Regarding P300-based BCIs, scalp-wide average broadband (1-40 Hz) LZC at rest was positively correlated with subsequent BCI performance in healthy participants [23], but negatively in a LIS-ALS patient [25].

Similarly, the slope of the 1/f PSD component exhibited inverted patterns. While healthy participants performed better with a flatter slope (i.e., smaller power-law

exponents) [23], performance in a LIS-ALS patient increased with a steeper slope [25]. Thielen [30] investigated this feature with regard to c-VEP-based BCI use. The study yielded a positive – albeit non-significant – correlation at electrode Oz, suggesting that a steeper slope at rest is related to better BCI performance. Lastly, Sannelli et al. [22] reported a correlation with the 1/f “noise” at rest and BCI performance at electrode C3. However, the 1/f “noise” was investigated only at the frequency that maximized the difference between the fitted PSD and the fitted 1/f spectrum at the Laplacian electrodes C3 and C4. Therefore, a direct comparison to studies investigating the 1/f slope [23,25,30] is impossible.

DISCUSSION

While psychological [44–46] and neurophysiological [47] factors are known to influence BCI performance, resting-state EEG prior to BCI use has largely been neglected in patient-centered research. This scoping review demonstrates that resting-state EEG contains valuable information about subsequent BCI performance, yet, only two of the 23 included studies examined patients [16,25].

Despite most studies reporting correlations rather than *predictions*, the temporal precedence of the resting-state EEG before BCI operation supports their interpretation as candidate predictors.

Crucially, resting-state EEG predictors appear paradigm-specific. For MI BCIs, results are consistent for the *SMR predictor* [12,15,22,31,42], as well as for higher mean functional connectivity, and network efficiency [20,43]. Furthermore, μ - and β -frequency band activity at rest appears to be important for subsequent MI BCI performance [28,29,31,33], as is information from electrodes C3 and C4 [12,13,15,22,29,31,33,36,42].

Conversely, performance with SSVEP-based BCIs shows inverted connectivity and network feature patterns. Mean functional connectivity and network efficiency metrics were negatively correlated with performance [41]. More importantly, α -frequency band activity appears to be particularly relevant for SSVEP-based BCI use. Particularly at (parieto-)occipital electrodes, α -frequency band activity showed consistent negative correlations with performance across studies [35,39,41]. These results suggest that an *optimal* user state is specific to the BCI paradigm. Although this heterogeneity prevents a meta-analysis across all featured studies, the consistent association between the SMR predictor and MI BCI performance [12,15,22,31,42] may justify a focused meta-analysis on this subsample in the future.

The transferability of predictors to patient samples also appears to be feature-dependent. For example, Borgheai et al. [16] found that resting-state δ -frequency band power remained a negative predictor in a regression model for predicting BCI performance involving data from healthy participants and ALS patients. However, LZC and the slope of the 1/f PSD component exhibited

inverted correlation patterns in a LIS-ALS patient [25] compared to healthy participants [23]. We hypothesize that the *brain criticality hypothesis* [48] could reconcile these conflicting results, given that LZC and the 1/f slope are both considered *criticality-related measures* [49]. The brain criticality hypothesis proposes a nonlinear relationship between information processing capacity and proximity to the point of criticality [48,50]. While healthy participants with brain activity close to criticality [51] are likely to approach the point of criticality with a flatter 1/f slope and increasing LZC values, the LIS-ALS patient may demonstrate a shift to supercriticality [23,25] due to ALS-related hyperexcitability [52] and the connection between hyperexcitability and supercritical dynamics [53]. This could help to explain the inverted correlation patterns observed for healthy controls and the LIS-ALS patient [23,25], suggesting that BCI performance predictors may be transferable to clinical populations when the underlying pathophysiology of the clinical sample is considered.

The 1/f slope emerges as a promising marker of user state fluctuations. Not only does it track changes in the ratio of excitatory to inhibitory synaptic activity [54], which, for its part, is related to changes in criticality [53], but it appears to change globally across the scalp – reducing the need for high-density electrode coverage [55]. However, conflicting results between Settgast and Kübler [23] and Thielen [30] call for further investigation. It is unclear whether patterns observed for P300-based BCIs [23,25] can easily be transferred to another BCI paradigm and whether a single electrode can adequately reflect the global change in the 1/f slope across the scalp.

This review emphasizes that ignoring systematic patterns between resting-state EEG and BCI performance can result in explainable variance being overlooked. This seems to be particularly significant for potential end-user groups, such as ALS patients [56], and is likely to increase with the consciousness fluctuations observed in CLIS [57,58]. To focus efforts on features related to promising *Windows of Consciousness* [59], systematic investigations in potential end-user groups are essential. To facilitate future systematic reviews and potential meta-analyses, we strongly recommend recording resting-state brain activity before BCI use and systematically reporting the recording length, to avoid confusion with other resting epochs during the BCI use, as well as reporting the eye status (open vs. closed) during the recording.

CONCLUSION

This scoping review identified consistent, yet paradigm-specific, associations between resting-state EEG and BCI performance. One key finding indicates that potential predictors may not easily transfer to patient samples without considering the underlying pathophysiological processes. This overview can guide future studies investigating resting-state EEG predictors for BCI performance in potential end-user groups. It may also

provide essential information for detecting *Windows of Consciousness* that allow for BCI-based communication, even for patients in CLIS.

REFERENCES

- [1] Northoff G, Duncan NW, Hayes DJ. The brain and its resting state activity—Experimental and methodological implications. *Prog Neurobiol* 2010;92(4):593–600.
- [2] Northoff G, Qin P, Nakao T. Rest-stimulus interaction in the brain: a review. *Trends Neurosci* 2010;33(6):277–84.
- [3] Cross ZR, Chatburn A, Melberzs L, Temby P, Pomeroy D, Schlesewsky M, et al. Task-related, intrinsic oscillatory and aperiodic neural activity predict performance in naturalistic team-based training scenarios. *Sci Rep* 2022;12(1):16172.
- [4] Xu L, Feng J, Yu L. Avalanche criticality in individuals, fluid intelligence, and working memory. *Hum Brain Mapp* 2022;43(8):2534–53.
- [5] Finn ES. Is it time to put rest to rest? *Trends Cogn Sci* 2021;25(12):1021–32.
- [6] Han Y, Ziebell P, Riccio A, Halder S. Two sides of the same coin: adaptation of BCIs to internal states with user-centered design and electrophysiological features. *Brain-Comput Interfaces* 2022;9(2):102–14.
- [7] Nikulin V. Predicting task performance and brain responses with ongoing neural activity. 2022 10th Int. Winter Conf. Brain-Comput. Interface BCI, 2022, p. 1–2.
- [8] Kübler A, Mattia D, Rupp R, Tangermann M. Facing the challenge: Bringing brain–computer interfaces to end-users. *Artif Intell Med* 2013;59(2):55–60.
- [9] Tricco AC, Lillie E, Zarin W, O’Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018;169(7):467–73.
- [10] Gudmundsson S, Runarsson TP, Sigurdsson S, Eiriksdottir G, Johnsen K. Reliability of quantitative EEG features. *Clin Neurophysiol* 2007;118(10):2162–71.
- [11] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
- [12] Acqualagna L, Botrel L, Vidaurre C, Kübler A, Blankertz B. Large-Scale Assessment of a Fully Automatic Co-Adaptive Motor Imagery-Based Brain Computer Interface. *PLOS ONE* 2016;11(2):e0148886.
- [13] Ahn M, Cho H, Ahn S, Jun SC. High theta and low alpha powers may be indicative of BCI-illiteracy in motor imagery. *PLoS ONE* 2013;8(11).
- [14] Leeb R, Lee F, Keinrath C, Scherer R, Bischof H, Pfurtscheller G. Brain–Computer Communication: Motivation, Aim, and Impact of Exploring a Virtual Apartment. *IEEE Trans Neural Syst Rehabil Eng* 2007;15(4):473–82.
- [15] Blankertz B, Sannelli C, Halder S, Hammer EM, Kübler A, Müller K-R, et al. Neurophysiological predictor of SMR-based BCI performance. *NeuroImage* 2010;51(4):1303–9.
- [16] Borgheai SB, Zisk AH, McLinden J, McIntyre J, Sadjadi R, Shahriari Y. Multimodal pre-screening can predict BCI performance variability: A novel subject-specific experimental scheme. *Comput Biol Med* 2024;168:107658.
- [17] Fernandez-Vargas J, Pfaff HU, Rodriguez FB, Varona P. Assisted closed-loop optimization of SSVEP-BCI efficiency. *Front Neural Circuits* 2013;7.
- [18] Lee M, Yoon J-G, Lee S-W. Predicting Motor Imagery Performance From Resting-State EEG Using Dynamic Causal Modeling. *Front Hum Neurosci* 2020;14.
- [19] Lee M-H, Kwon O-Y, Kim Y-J, Kim H-K, Lee Y-E, Williamson J, et al. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience* 2019;8(5):giz002.
- [20] Park H, Jun SC. Connectivity study on resting-state EEG between motor imagery BCI-literate and BCI-illiterate groups. *J Neural Eng* 2024.
- [21] Cho H, Ahn M, Ahn S, Kwon M, Jun SC. EEG datasets for motor imagery brain–computer interface. *GigaScience* 2017;6(7):gix034.
- [22] Sannelli C, Vidaurre C, Müller K-R, Blankertz B. A large scale screening study with a SMR-based BCI: Categorization of BCI users and differences in their SMR activity. *PLOS ONE* 2019;14(1):e0207351.
- [23] Settgast T, Kübler A. Resting-State Brain Criticality and Performance with P300-based BCIs. *Proc. 9th Graz Brain-Comput. Interface Conf. 2024, Graz, Austria: Verlag der Technischen Universität Graz; 2024, p. 162–7.*
- [24] Won K, Kwon M, Ahn M, Jun SC. EEG Dataset for RSVP and P300 Speller Brain-Computer Interfaces. *Sci Data* 2022;9(1):388.
- [25] Settgast T, Zilio F, Kübler A, Northoff G. Correlation between Neurophysiological Measures of Consciousness and BCI Performance in a Locked-in Patient. 2023 11th Int. Winter Conf. Brain-Comput. Interface BCI, 2023, p. 1–6.
- [26] Ziebell P, Pfeiffer M, Tröger W, Halder S, Giesler T, Kübler A, et al. Persistent tactile P3 in long-term brain-computer interface home use by a patient in locked-in state 2024.
- [27] Shin G-H, Lee M, Kim H-J, Lee S-W. Prediction of Event Related Potential Speller Performance Using Resting-State EEG. 2020 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC, Montreal, QC, Canada: IEEE; 2020, p. 2973–6.
- [28] Sonntag J, Yu L, Wang X, Schack T. Neurophysiological predictors of deep learning based unilateral upper limb motor imagery classification. *Front Hum Neurosci* 2025;19.
- [29] Suk H-I, Fazli S, Mehnert J, Müller K-R, Lee S-W. Predicting BCI Subject Performance Using Probabilistic Spatio-Temporal Filters. *PLOS ONE* 2014;9(2):e87056.
- [30] Thielen J. Addressing BCI inefficiency in c-VEP-based BCIs: A comprehensive study of neurophysiological predictors, binary stimulus sequences, and user comfort. *Biomed Phys Eng Express* 2025;11(4):045017.
- [31] Trocellier D, N’Kaoua B, Lotte F. Validating neurophysiological predictors of BCI performance on a

- large open source dataset. GBCIC 2024 - 9th Graz Brain-Comput. Interface Conf. 2024, Graz, Austria: 2024.
- [32] Dreyer P, Roc A, Pillette L, Rimbert S, Lotte F. A large EEG database with users' profile information for motor imagery brain-computer interface research. *Sci Data* 2023;10(1):580.
- [33] Tzdaka E, Benaroch C, Jeunet C, Lotte F. Assessing The Relevance Of Neurophysiological Patterns To Predict Motor Imagery-based BCI Users' Performance. 2020 IEEE Int. Conf. Syst. Man Cybern. SMC, 2020, p. 2490–5.
- [34] Roc A, Pillette L, N'Kaoua B, Lotte F. Would Motor-Imagery based BCI user training benefit from more women experimenters? GBCIC2019 - 8th Graz Brain-Comput. Interface Conf. 2019, Graz, Austria: 2019.
- [35] Wan F, da Cruz JN, Nan W, Wong CM, Vai MI, Rosa A. Alpha neurofeedback training improves SSVEP-based BCI performance. *J Neural Eng* 2016;13(3):036019.
- [36] Wang K, Tian F, Xu M, Zhang S, Xu L, Ming D. Resting-State EEG in Alpha Rhythm May Be Indicative of the Performance of Motor Imagery-Based Brain-Computer Interface. *Entropy* 2022;24(11):1556.
- [37] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* 2000;101(23):e215–20.
- [38] Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans Biomed Eng* 2004;51(6):1034–43.
- [39] Won D-O, Hwang H-J, Dähne S, Müller K-R, Lee S-W. Effect of higher frequency on the classification of steady-state visual evoked potentials. *J Neural Eng* 2015;13(1):016014.
- [40] Yoon J-G, Lee M. Effective Correlates of Motor Imagery Performance based on Default Mode Network in Resting-State. 2020 8th Int. Winter Conf. Brain-Comput. Interface BCI, 2020, p. 1–5.
- [41] Zhang Y, Xu P, Guo D, Yao D. Prediction of SSVEP-based BCI performance by the resting-state EEG network. *J Neural Eng* 2013;10(6):066017.
- [42] Zhang R, Xu P, Chen R, Li F, Guo L, Li P, et al. Predicting Inter-session Performance of SMR-Based Brain-Computer Interface Using the Spectral Entropy of Resting-State EEG. *Brain Topogr* 2015;28(5):680–90.
- [43] Zhang R, Yao D, Valdés-Sosa PA, Li F, Li P, Zhang T, et al. Efficient resting-state EEG network facilitates motor imagery performance. *J Neural Eng* 2015;12(6):066024.
- [44] Geronimo A, Simmons Z, Schiff SJ. Performance predictors of brain-computer interfaces in patients with amyotrophic lateral sclerosis. *J Neural Eng* 2016;13(2).
- [45] Riccio A, Simone L, Schettini F, Pizzimenti A, Inghilleri M, Olivetti Belardinelli M, et al. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Front Hum Neurosci* 2013;7.
- [46] Riccio A, Schettini F, Simone L, Pizzimenti A, Inghilleri M, Olivetti-Belardinelli M, et al. On the Relationship Between Attention Processing and P300-Based Brain Computer Interface Control in Amyotrophic Lateral Sclerosis. *Front Hum Neurosci* 2018;12(May):1–10.
- [47] Halder S, Ruf CA, Furdea A, Pasqualotto E, De Massari D, van der Heiden L, et al. Prediction of P300 BCI Aptitude in Severe Motor Impairment. *PLoS ONE* 2013;8(10):e76148.
- [48] O'Byrne J, Jerbi K. How critical is brain criticality? *Trends Neurosci* 2022;45(11):820–37.
- [49] Maschke C, O'Byrne J, Colombo MA, Boly M, Gosseries O, Laureys S, et al. Critical dynamics in spontaneous EEG predict anesthetic-induced loss of consciousness and perturbational complexity. *Commun Biol* 2024;7(1):1–14.
- [50] Müller PM, Miron G, Holtkamp M, Meisel C. Critical dynamics predicts cognitive performance and provides a common framework for heterogeneous mechanisms impacting cognition. *Proc Natl Acad Sci* 2025;122(14):e2417117122.
- [51] Meisel C, Olbrich E, Shriki O, Achermann P. Fading Signatures of Critical Brain Dynamics during Sustained Wakefulness in Humans. *J Neurosci* 2013;33(44):17363–72.
- [52] Gunes ZI, Kan VWY, Jiang S, Logunov E, Ye X, Liebscher S. Cortical Hyperexcitability in the Driver's Seat in ALS. *Clin Transl Neurosci* 2022;6(1):5.
- [53] Javed E, Suárez-Méndez I, Susi G, Román JV, Palva JM, Maestú F, et al. A Shift Toward Supercritical Brain Dynamics Predicts Alzheimer's Disease Progression. *J Neurosci* 2025;45(9).
- [54] Gao R, Peterson EJ, Voytek B. Inferring synaptic excitation/inhibition balance from field potentials. *NeuroImage* 2017;158:70–8.
- [55] Hebron H, Stephan AM, Cataldi J, Siclari F. Aperiodic EEG Activity Provides a Linear, Bidirectional, and Spatially Uniform Marker of Subjective and Objective Vigilance in Humans, Both Within and Across States 2025:2025.08.30.673229.
- [56] Marchetti M, Priftis K. Brain-computer interfaces in amyotrophic lateral sclerosis: A metanalysis. *Clin Neurophysiol* 2015;126(6):1255–63.
- [57] Adama S, Bogdan M. Assessing consciousness in patients with locked-in syndrome using their EEG. *Front Neurosci* 2025;19.
- [58] Zilio F, Gomez-Pilar J, Chaudhary U, Fogel S, Fomina T, Synofzik M, et al. Altered brain dynamics index levels of arousal in complete locked-in syndrome. *Commun Biol* 2023;6(1):1–19.
- [59] Kübler A. The history of BCI: From a vision for the future to real support for personhood in people with locked-in syndrome. *Neuroethics* 2020;13(2):163–80.

Comparison of sequence, stimulus and decoding algorithm combinations in c-VEP-based BCIs

A. Martín-Fernández^{1,2}, V. Martínez-Cagigal^{1,2,3}, E. Santamaría-Vázquez^{1,2,3}, R. Hornero^{1,2,3}

¹Biomedical Engineering Group, University of Valladolid, ETSIT, Paseo de Belen, 15, 47011, Valladolid, Spain

²Centro de Investigacion Biomedica en Red de Bioingenieria, Biomateriales y Nanomedicina (CIBER-BBN), Valladolid, Spain

³Valladolid Health Research Institute (IBioVALL), C. Rondilla Sta. Teresa, 47010, Valladolid, Spain

E-mail: ana.martin.fernandez23@uva.es

ABSTRACT: Brain-computer interfaces (BCIs) enable users to control external devices using brain activity. Among exogenous paradigms, code-modulated visual evoked potentials (c-VEP) stand out. These are neural responses elicited by flickering visual stimuli encoded with pseudorandom sequences, allowing the system to identify the target attended by the user. Optimizing c-VEP performance requires understanding the combined effects of encoding sequences, stimulus patterns, and decoding algorithms. In this study, 26 healthy participants evaluated a 16-target speller with 6 c-VEP configurations, combining 2 sequences (m-sequences, burst codes) with 3 stimulus patterns (plain, grating, checkerboard) using 4 decoding algorithms. Results showed that burst codes with a checkerboard pattern and a bit-wise reconstruction decoder offered the best balance between accuracy (99.76%) and comfort (eyestrain score 4.81/10). However, several combinations also yielded reliable performance, highlighting that optimal design may depend on the specific application needs. This work emphasizes the importance of jointly considering sequence type, stimulus pattern, and decoding algorithm in c-VEP research.

INTRODUCTION

Brain-computer interfaces (BCIs) can be described as systems that enable individuals to interact with the environment relying solely on their brain activity [1]. By establishing a direct communication pathway between the brain and an external application, BCIs translate recorded neural signals into executable commands through signal processing and intention decoding algorithms. Over the past decades, substantial research efforts have advanced the development of these systems, demonstrating their feasibility across a wide range of applications [1].

Despite relying on a common operational framework, BCIs may exhibit considerable variability depending on the intended use. One aspect common to most BCI systems is the method used to acquire brain activity, which is typically electroencephalography (EEG), due to its non-invasive nature, relatively low cost, and high temporal

resolution compared to other techniques [2]. In contrast, the paradigm employed to decode user intentions is highly application-dependent. Rehabilitation-oriented systems typically rely on endogenous paradigms, in which users learn to voluntarily modulate their brain activity [1]. Conversely, communication and control systems more often employ exogenous paradigms, where external stimuli are presented to elicit specific, stimulus-driven neural responses [1].

Among the exogenous approaches that have gained considerable attention in the literature are code-modulated visual evoked potentials (c-VEP) [3]. These signals arise as natural brain responses to visual stimulation in which brief light flashes are presented according to a pseudorandom sequence. By analyzing the EEG activity the system can determine which target the user is focusing on and execute the task associated with it [3]. Although impressive performance has been demonstrated, with reported accuracies exceeding 90%, the main challenges of c-VEP-based BCIs lie in maintaining such performance while reducing calibration and selection times [4], as well as ensuring an adequate user experience by minimizing visual fatigue induced by prolonged stimulation [5, 6]. Research efforts aimed at optimizing these systems have primarily focused on three main aspects: encoding sequence, stimulus pattern, and decoding algorithms.

Regarding the encoding sequences used to codify the visual stimulation, it is necessary to employ sets of codes with minimal cross-correlation, as greater dissimilarity between sequences facilitates accurate target classification. Several families of codes have been proposed, such as m-sequences [3] or burst codes [5]. Concerning stimulus patterns, the classical binary approach represents the 0-bit as black and the 1-bit as white. However, due to the visual discomfort induced by high-contrast flashes, alternative designs have been proposed, incorporating edges and contours instead of plain patterns. These designs include checkerboard [6] and grating-like stimuli [5]. Finally, with respect to decoding algorithms, two main processing pipelines are commonly followed: template matching (TM) [3, 7] and bit-wise reconstruction (BWR)

[5, 8]. The TM approach averages and spatially filters the user's EEG responses to each stimulation code to generate representative templates. During online operation, the incoming EEG signal is compared against these templates. On the other hand, the BWR approach trains a classifier to predict whether a given EEG epoch corresponds to a 0-bit or a 1-bit stimulation. The reconstructed bit sequence is then compared with the original codes to identify the intended command.

The increasing number of alternatives in the design of c-VEP-based BCI systems highlights the need to understand not only which encoding sequences, stimulus patterns, or decoding algorithms perform best individually under specific conditions, but also which combinations of these three elements yield the best overall performance. Previous studies have primarily focused on pairwise comparisons, typically evaluating a novel proposal against a standard approach. However, no study has analyzed the performance of multiple configurations while jointly considering encoding sequences, stimulus patterns, and decoding algorithms within a unified comparison framework. The main objective of the present study is to examine both performance and visual fatigue across all possible combinations derived from a selected subset of sequences, stimulus patterns, and decoding algorithms. This work represents a novel contribution by introducing previously unexplored c-VEP configurations and by providing a comprehensive, multi-perspective evaluation framework for c-VEP-based BCI systems.

MATERIALS AND METHODS

Participants and equipment: Twenty-six healthy volunteers (15 males and 11 females; mean age 24.88 ± 4.96 years) with normal or corrected-to-normal vision participated in the study. EEG signals were recorded using a g.USBamp amplifier (g.tec Medical Engineering GmbH, Austria) at a sampling frequency of 256 Hz. Sixteen active electrodes (F3, Fz, F4, C3, Cz, C4, CPz, P3, Pz, P4, PO7, POz, PO8, Oz, I1, and I2) were positioned according to the International 10-10 System. The ground electrode was placed at AFz, and the reference electrode at the right earlobe (A2). Data acquisition and signal processing were performed on a PC equipped with an Intel Core i7-13700 processor, 32 GB RAM, and an NVIDIA GeForce RTX 3050 GPU. Visual stimulation was presented on a Full HD LED monitor with a refresh rate of 120 Hz.

System configuration: The study was implemented using MEDUSA[®], a software ecosystem for the development and deployment of BCI experiments [9]. The different combinations of encoding sequences, stimulus patterns, and decoding algorithms were configured and evaluated using the c-VEP Speller application (version 5.0.0), which is publicly available at medusabci.com.

The configuration parameters that remained constant throughout the experiment included the number of targets, fixed at 16; the layout, defined as a 4×4 matrix;

and the background, which consisted of a uniform bluish color. EEG signals were consistently preprocessed with a 1-60 Hz bandpass filter and a 50 Hz notch filter. In contrast, the encoding sequence, stimulus pattern, and decoding algorithm were varied according to the experimental condition. To conduct the comparative analysis, a subset of configurations for these three elements was selected based on their relevance in the literature and practical feasibility.

The selected sequences were binary m-sequences and burst codes (see Fig. 1). M-sequences represent the standard approach in c-VEP, whereas burst codes have recently emerged as an alternative to increase VEP amplitude and reduce visual fatigue [5]. M-sequences are characterized by excellent autocorrelation properties, exhibiting a maximum autocorrelation at zero lag and near-zero autocorrelation for shifted versions of the sequence. This property allows the system to be trained using a single base sequence, while additional commands are generated through temporal shifts of that sequence [3]. In this study, a 63-bit m-sequence was generated using a 6th-order base-2 polynomial implemented through a linear feedback shift register. For the 16-target configuration, each command was generated by circularly shifting the base sequence, resulting in an average lag of 3.94 bits between commands. Burst codes consist of short aperiodic visual flashes presented at a slow stimulation rate. Burst onsets derived from a master sequence were pseudo-randomly assigned to the different codes in batches, with the assignment order reshuffled after each distribution cycle until all onsets were allocated [5]. In this study, a set of 16 burst codes was generated, each with a fixed length of 80 bits and containing 3 single-bit bursts. This resulted in a minimum inter-burst interval of 50 ms and a maximum interval of 480 ms, with no temporal overlap between bursts of different codes.

The evaluated stimulus patterns comprised plain, grating, and checkerboard (see Fig. 1). As with the encoding sequences, the plain pattern represents the classical approach, whereas the grating and checkerboard designs correspond to more recent proposals aimed at improving user experience [3]. Plain stimulation encodes each bit of the binary sequence using a single color: bit-0 is represented by black and bit-1 by white, thereby producing maximum contrast [3]. Grating stimulation is implemented using 75 small, randomly oriented Gabor patches consisting of a central white bar surrounded by two parallel black bars on a mid-gray background. In this case, bit-0 corresponds to “off” state in which only the mid-gray background is displayed, whereas bit-1 corresponds to “on” state in which the Gabor patches appear superimposed on the background [5]. Checkerboard stimulation follows a similar on/off logic. However, bit-0 is represented by black background, and bit-1 by the appearance of a 16×16 black-and-white checkerboard pattern [6].

The decoding approaches considered in this study included two TM and two BWR pipelines. The first TM approach is the standard one, here referred as event-related

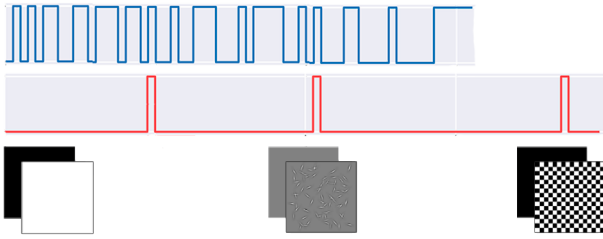


Figure 1: Subset of sequences and stimuli. Sequences (top to bottom): m-sequence and burst code. Stimuli (right to left): plain, grating, checkerboard.

potential template matching (TM-ERP) [3]. For each command, a template of the EEG response is generated. To obtain this template, trials of stimulation are averaged. Then, a canonical correlation analysis (CCA) is performed to maximize the correlation between the averaged and the single-trial signals. The first CCA component is used as a spatial filter [3]. The second TM approach is Reconvolution (TM-Reconvolution) [7]. Here, instead of averaging across trials, EEG response is first decomposed into individual event responses of 300 ms, which capture the brain’s response to elementary visual events such as contrast changes. Then, full responses for each command are reconstructed by convolving these event responses according to the temporal structure of the sequences. The CCA step is subsequently applied to compute a spatial filter, which, when applied to the signals, produces the final template for each command [7]. For the BWR approaches, the first method is Riemann bit-wise reconstruction (BWR-Riemann) [5]. This is a machine-learning approach based on Riemannian geometry. First, xDAWN spatial filtering is applied to enhance the evoked response, after that for each EEG epoch covariance matrices are computed. These matrices are projected into the tangent space of the Riemannian manifold to yield vector features. The features are input to a linear discriminant analysis (LDA) classifier to estimate the probability that a given EEG epoch belongs to a bit-0 or bit-1. The predicted sequence is then correlated to all sequences [5]. The remaining BWR algorithm is the EEG-Inception bit-wise reconstruction (BWR-EEGInception) [8]. A deep learning pipeline based on the EEG-Inception model, a convolutional neural network (CNN) whose effectiveness in c-VEP decoding has been previously demonstrated [8]. EEG signals are downsampled and segmented into short epochs time-locked to each bit. These epochs are processed by the EEG-Inception network, which extracts spatiotemporal features to classify each epoch as a bit-0 or bit-1. The predicted sequence is then correlated to all the sequences [8].

Experimental protocol: To conduct the comparative analyses addressed in this study, an experimental protocol was designed to evaluate 6 distinct c-VEP configurations. These configurations resulted from combining each of the two proposed pseudorandom sequences (m-sequences and burst codes) with each of the three stimulus patterns (plain, grating, and checkerboard). Regarding

the decoding algorithms, all conditions were evaluated online using a single method (TM-ERP for m-sequence conditions and BWR-Riemann for burst code conditions). The remaining decoding algorithms were evaluated offline to avoid excessively increasing the duration of the experimental session.

Each participant completed a single experimental session in which all proposed c-VEP conditions were evaluated, with the order of condition presentation randomized to minimize potential order effects. For each condition, participants completed two calibration runs followed by one online run. Each calibration run consisted of 80 cycles (i.e., complete sequence repetitions). In the m-sequence conditions, the same target was presented throughout, as a single base sequence and its shifted versions were used to encode all commands. In the burst code conditions, the target letter changed every 10 cycles to ensure that all 16 targets were presented, since each command is associated with a distinct code. The online run consisted of 16 selections, one for each target, with 8 cycles per selection. After completing each condition, participants rated their perceived visual fatigue using a 0-10 Likert scale, where 0 indicated no fatigue and 10 indicated extreme fatigue.

Evaluation metrics: The evaluation of the experimental conditions in this study was based on three key aspects: system performance, EEG responses, and user experience. System performance was assessed in terms of accuracy, calculated as the number of correct selections divided by the total number of selections. To characterize EEG responses, calibration data were analyzed. For the m-sequence conditions, the average EEG response across complete cycle presentations was computed. In contrast, for the burst code conditions, each of the 16 commands is associated with a distinct sequence, and responses to individual events exhibit considerably less temporal overlap. Therefore, ERPs were computed time-locked to each burst onset (i.e., bit-1 onset). Finally, user experience was evaluated through the subjective visual fatigue ratings provided by participants in the questionnaires completed during the experimental protocol.

To assess differences between experimental conditions, statistical analyses were conducted on both accuracy and visual fatigue metrics. Since the data were paired and did not follow a normal distribution, pairwise comparisons between conditions were performed using the Wilcoxon signed-rank test. To account for multiple comparisons and control the false discovery rate (FDR), the p -values were adjusted using the Benjamini-Hochberg procedure.

RESULTS

The system performance was computed separately for each of the decoding algorithms selected for comparison. Results were averaged across participants, yielding an accuracy curve for each condition and algorithm. These curves are shown in Fig. 2. The time point closest to 4.2 s was used for statistical comparisons, as this corresponds to the maximum duration of data shared across all condi-

Table 1: Subjective visual fatigue ratings reported by participants for each experimental condition.

Condition	Mean	SD	CLD
burst-grat	2.35	1.96	a
mseq-grat	3.19	1.94	b
burst-check	4.81	1.90	c
burst-plain	4.96	1.99	c
mseq-check	6.00	2.00	d
mseq-plain	7.42	1.96	e

SD: Standard Deviation

CLD: Compact Letter Display

tions, limited by the length of the m-sequence conditions. The corrected p -values from these statistical analyses are also shown in Fig. 2, where the central matrices display the results of all pairwise comparisons.

EEG responses computed for each condition were grand-averaged across participants at Oz, and the resulting mean signals are shown in Fig. 3.

Finally, regarding user experience, the results are presented in Tab. 1. Visual fatigue scores provided by the participants were averaged for each condition and are reported together with their standard deviation (SD). The last column of the table shows the results of the pairwise statistical comparisons. These results are presented using the Compact Letter Display (CLD) format, with a significance level of corrected p -value < 0.05 . According to this format, any two conditions assigned the same CLD letter do not differ statistically significantly, whereas any two conditions labeled with different letters exhibit statistically significant differences [10].

DISCUSSION

System performance: Looking at the accuracy curves in Fig. 2, a high degree of variability can be observed depending on the combination of the three factors considered (sequence, stimulus pattern, and decoding algorithm). The statistical analysis provides great insight into this variability, as numerous statistically significant differences are found among different sequence and stimulus combinations evaluated with the same decoding algorithm. Additionally, certain conditions defined by specific sequences and stimulus patterns are not consistent across different decoding algorithms. In other words, its performance depends strongly on the decoding method applied. These results suggest that the choice of all three factors has a considerable impact on system performance and should therefore be carefully considered. Importantly, not only do these factors influence performance independently, but their interaction must also be jointly taken into account to optimize the configuration of c-VEP-based BCIs.

Focusing on the influence of the pseudorandom sequence choice, it cannot be clearly stated whether m-sequences or burst codes exhibit superior performance, as the results also depend on the stimulus pattern. However, m-sequences tend to achieve higher accuracies overall. Except for the checkerboard stimulus pattern, m-sequences

show higher accuracy curves than their corresponding burst-code conditions across all decoding algorithms.

Regarding stimulus patterns, no single design can be considered universally superior, as performance also depends on the sequence and decoding algorithm. Checkerboard patterns achieve the best overall results when paired with burst codes, whereas plain patterns may outperform checkerboard in combination with m-sequences. One consistent finding across all decoding algorithms is that grating patterns produce the lowest accuracies, regardless of the sequence type. This outcome may be influenced by the monitor refresh rate; previous studies analyzed the performance at 60 Hz [5], whereas the present study used 120 Hz.

Regarding the decoding algorithms, BWR-Riemann achieved the highest and most consistent performance, outperforming the other methods. It was the only algorithm to reach accuracies above 90% for both plain and checkerboard conditions. Moreover, despite the lower performance observed for the grating conditions, BWR-Riemann still outperformed the other algorithms. The remaining decoding methods showed effective performance but just for specific combinations of sequence and stimulus pattern.

Overall, the highest performance was achieved with the combination of burst codes and the checkerboard pattern, which consistently outperformed other conditions across all decoding algorithms, reaching nearly 100% grand-averaged accuracy in less than 4 s of decoding. However, it is important to note that several other combinations also yielded effective performance. The key takeaway is that system optimization should consider the joint effects of all elements, designing the configuration coherently based on the combined impact of sequence, stimulus pattern, and decoding algorithm.

EEG responses: To analyze EEG responses, the m-sequence and burst code conditions need to be considered separately, due to the inherent differences in the temporal structure and characteristics of the two sequence types. For the m-sequence conditions, EEG responses obtained from the calibration data (upper plot in Fig. 3) show that the grating and checkerboard patterns elicit similar neural responses, both in amplitude and waveform morphology, compared to the plain stimulation. In contrast, the plain stimulation produces a stronger neural response in terms of amplitude, but also shows greater variability across participants, as reflected by the wider confidence interval showed in the graph as the shaded region.

For the burst code conditions, EEG responses (lower plot in Fig. 3) resemble classical VEPs, due to the longer intervals between bursts and consequently less overlapping stimulation. In this context, the plain condition resembles a flash VEP, whereas the checkerboard and grating conditions resemble onset/offset VEPs [11]. In the plain condition, the positive peak around 120 ms likely corresponds to the P2 component of the flash VEP [11]. In the checkerboard condition, the negative peak near 120 ms and a positive peak around 150 ms may correspond to the

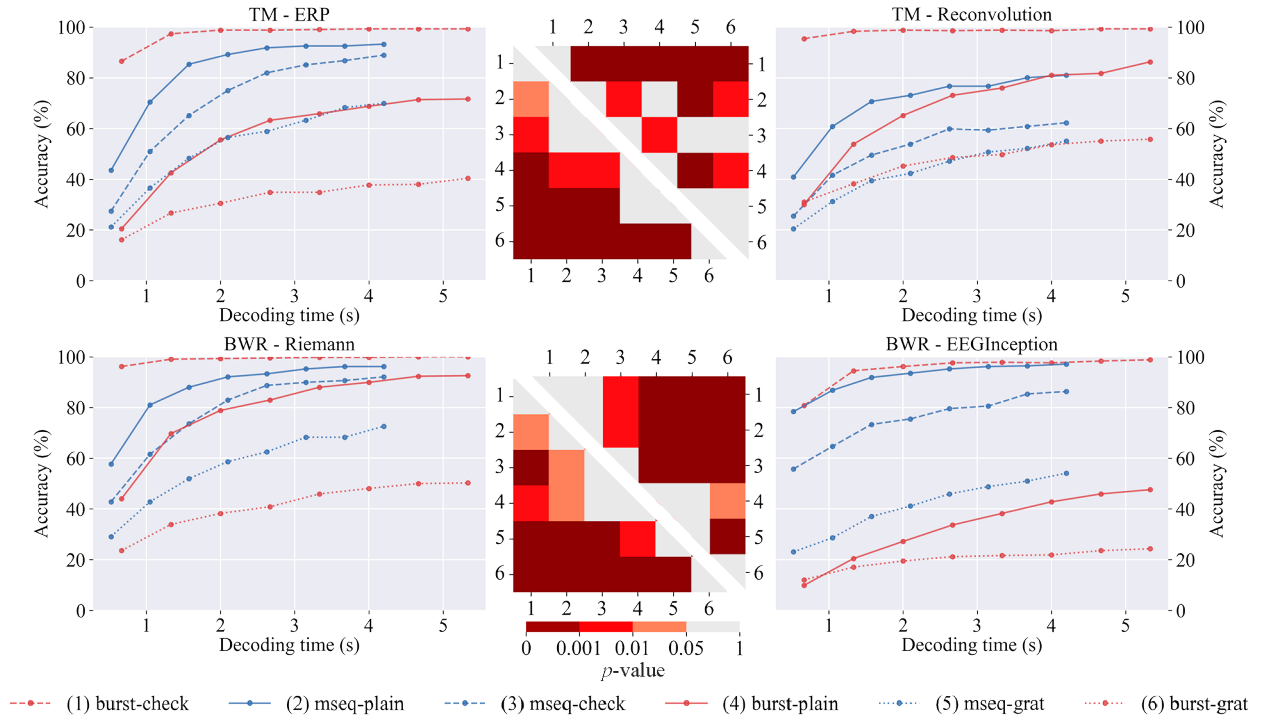


Figure 2: Grand-average accuracy curves over decoding time for each condition and decoding algorithm. The central matrices show corrected p -values from pairwise statistical comparisons. Each matrix is divided by the main diagonal, with the lower triangle corresponding the left-hand decoding algorithms and the upper triangle to those for the right-hand. Conditions are labeled from 1 to 6, where 1 denotes the condition with the highest grand-average accuracy across participants and decoding algorithms, and 6 the opposite.

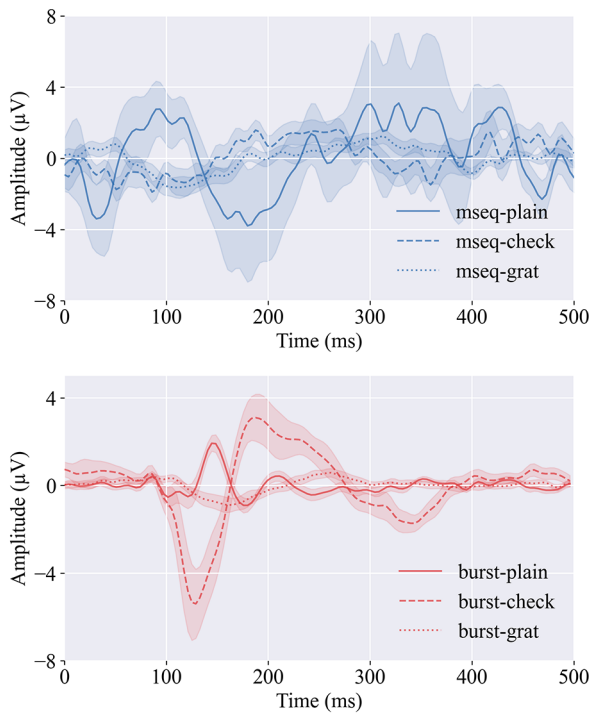


Figure 3: Grand-average EEG responses across users at Oz for each experimental condition using calibration data. Shaded areas represent the 95% confidence interval.

C2 and C3 components of the onset/offset VEP [11]. For the grating condition, these components are also present but they exhibit markedly lower amplitude and a latency delayed by approximately 20 ms. These characteristics could explain the observed differences in system performance: the delayed and weaker response in the grating condition may account for its lower accuracy, while the pronounced amplitude and consistent morphology of the checkerboard may contribute to its superior performance.

User experience: The subjective visual fatigue ratings reported by participants during the experimental protocol reveal a clear ranking of conditions, as shown in Tab. 1. Moreover, almost every step in this ranking corresponds to statistically significant differences. The only exception is the comparison between the plain and checkerboard burst conditions, for which no statistically significant differences were observed.

The conditions that induced the lowest levels of visual fatigue were those using the grating stimulus pattern, suggesting that the stimulus design may be the factor with the greatest impact on user experience. Intermediate fatigue levels were observed for the burst code conditions with plain and checkerboard patterns, whereas the most fatiguing conditions corresponded to the same stimulus patterns combined with m-sequences. Overall, these results indicate that burst codes are generally preferred over m-sequences in terms of sequence type. Regarding stimulus patterns, grating is the most preferred, followed by checkerboard, and finally by plain, which is perceived as

the most fatiguing.

When considering both system performance and user experience, a clear trade-off emerges. The conditions that yielded the lowest performance were those preferred by users in terms of visual fatigue. Therefore, achieving an optimal balance between these two key aspects of c-VEP-based BCIs becomes essential. In this context, the burst code condition with the checkerboard pattern appears to offer the best solution. It provides the highest system performance while remaining among the least fatiguing conditions, only behind the grating conditions, which, despite being more comfortable, showed poor performance.

Limitations and future research: This study offers important contributions to research on c-VEP-based BCIs; however, several limitations should be considered. A larger and more heterogeneous cohort would improve statistical power and generalization of the results. In addition, visual fatigue was assessed through subjective ratings; incorporating EEG biomarkers or eye-tracking metrics, would provide a more objective evaluation.

Additional limitations related to the comparative design must be acknowledged. Given the large number of possible combinations of sequence type, stimulus pattern, and decoding algorithm, only a subset of conditions was evaluated, potentially excluding others. Moreover, differences between sequences led to unequal amounts of online and calibration data; while decoding performance was analyzed as a function of decoding time, its dependence on calibration time remains unstudied.

CONCLUSION

The present research provides a comprehensive comparison of multiple c-VEP-based BCI configurations, combining different pseudorandom sequences, stimulus patterns, and decoding algorithms. Results indicate that the best balance between system performance and user experience is achieved with burst codes combined with the checkerboard pattern and the BWR-Riemann decoding algorithm, yielding a mean accuracy of 99.76% and a mean fatigue score of 4.81 ± 1.90 . However, findings also showed that a wide range of parameter combinations can yield reliable performance, and that the optimal choice may depend on the specific research or application requirements. Importantly, performance does not depend on any single factor alone but on the interaction between sequence type, stimulus pattern, and decoding algorithm. When a particular sequence is chosen, the most suitable stimulus pattern or compatible decoding algorithm may vary considerably. This interaction-driven behavior underscores the importance of considering these parameters jointly rather than independently when designing c-VEP-based BCI systems.

ACKNOWLEDGMENTS

This work was supported by 0124_EUROAGE_MAS_4_E project, co-financed

by the EU through the Interreg VI-A Spain-Portugal Programme (POCTEP) 2021-2027, and by the VA140P24 project, funded by the Regional Government of Castilla y León (Consejería de Educación) and EU-ERDF. Support has also been provided by the Centre for Biomedical Research in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN, CB19/01/00012) through the Carlos III Health Institute, co-financed with ERDF funds. Ana Martín-Fernández is a beneficiary of a PIF-UVa grant.

REFERENCES

- [1] Wolpaw J, Wolpaw EW. Brain-computer interfaces: Principles and practice. Oxford University Press (2012).
- [2] Min BK, Marzelli MJ, Yoo SS. Neuroimaging-based approaches in the brain-computer interface. Trends in Biotechnology. 2010;28(11):552–560.
- [3] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): A literature review. Journal of Neural Engineering. 2021;18(6):061002.
- [4] Zheng L, Dong Y, Tian S, Pei W, Gao X, Wang Y. A calibration-free c-VEP based BCI employing narrow-band random sequences. Journal of Neural Engineering. 2024;21(2):026023.
- [5] Dehais F, Castillos KC, Ladouce S, Clisson P. Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry EEG code-VEP BCI. Journal of Neural Engineering. 2024;21(6):066023.
- [6] Fernández-Rodríguez Á, Martínez-Cagigal V, Santamaría-Vázquez E, Ron-Angevin R, Hornero R. Influence of spatial frequency in visual stimuli for cVEP-based BCIs: Evaluation of performance and user experience. Frontiers in Human Neuroscience. 2023;17:1288438.
- [7] Thielen J, Broek Pvd, Farquhar J, Desain P. Broad-band visually evoked potentials: Re(con)volution in brain-computer interfacing. PLOS ONE. 2015;10(7):e0133797.
- [8] Santamaría-Vázquez E, Martínez-Cagigal V, Hornero R. Bit-wise reconstruction of non-binary visual stimulation patterns from EEG using deep learning: A promising alternative for user-friendly high-speed c-VEP-based BCIs. In: Advances in Computational Intelligence. Switzerland, 2023, 603–614.
- [9] Santamaría-Vázquez E et al. MEDUSA©: A novel python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. Computer Methods and Programs in Biomedicine. 2023;230:107357.
- [10] Piepho HP. An algorithm for a letter-based representation of all-pairwise comparisons. Journal of Computational and Graphical Statistics. 2004;13(2):456–466.
- [11] Odom JV et al. ISCEV standard for clinical visual evoked potentials: (2016 update). Documenta Ophthalmologica. 2016;133(1):1–9.

Wearable brain-computer interfaces and their applications: advancements from the ARHeMlab research group

Antonio Esposito¹, Nicola Moccaldi¹, Ludovica Gargiulo², Luigi Duraccio¹, Egidio De Benedetto¹,
Enza Galasso³, Lucrezia Di Marino¹, Leopoldo Angrisani¹, Pasquale Arpaia^{1,4}

¹Department of Electrical Engineering and Information Technology (DIETI), Università degli Studi di Napoli Federico II, Naples, Italy

²Institute of industrial technologies and automation, National Council of Research (STIIMA-CNR), Lecco, Italy

³Department of Chemical, Materials and Industrial Production Engineering (DICMaPI), Università degli Studi di Napoli Federico II, Naples, Italy

⁴Interdepartmental Research Center on Management and Innovation in Healthcare (CIRMIS), Università degli Studi di Napoli Federico II, Naples, Italy

E-mail: pasquale.arpaia@unina.it

ABSTRACT: This work presents recent advancements in the development of wearable brain-computer interfaces (BCIs) by the ARHeMlab research group, focusing on the period from 2021 to 2025. The activities address the key challenges of translating BCIs from controlled laboratory settings to real-world, daily-life scenarios. The research integrates wearable and portable EEG hardware with custom signal processing pipelines and XR environments to enhance signal quality and user engagement. Experimental contributions span reactive, passive, and active BCI paradigms, including SSVEP-based interfaces with adaptive XR stimulation, cognitive load and distraction monitoring using low-channel EEG, and motor imagery control with multimodal neurofeedback. The proposed solutions are validated through application-specific studies, supporting the feasibility of practical, scalable, and ecologically valid BCI systems. No original research is presented in this work but the abovementioned works are reviewed.

INTRODUCTION

Brain-Computer Interfaces (BCIs) offer the potential to establish a direct communication pathway between the brain and external devices, thus disclosing novel applications in neurorehabilitation, assistive technologies, mental workload monitoring, and user engagement enhancement [1]. By interpreting neural signals, which are typically acquired through electroencephalography (EEG), BCIs can translate user intent into actionable outputs, offering an alternative control strategy for individuals with severe motor impairments and enriching human-machine interaction [2], [3].

Despite their promises, most BCI systems are often restricted to controlled laboratory environments due to cumbersome equipment, time-consuming setup procedures, and limited wearability [4]. These constraints

significantly hinder the adoption of BCIs in everyday life and long-term use cases, such as home-based (neuro)rehabilitation, adaptive gaming, and continuous mental state monitoring. To bridge this gap, there is a growing demand for wearable BCI solutions, capable of delivering reliable performance while prioritizing user comfort, ease of use, and real-time operation [5], [6].

Recent literature highlights an increase in the development of compact, low-density EEG systems, dry-sensor technologies, and wireless platforms integrated with extended reality (XR) [7], [8]. These advances are complemented by machine learning techniques for robust signal classification, as well as multimodal feedback strategies to enhance user engagement and signal quality [9, 10]. At the same time, challenges such as reduced signal-to-noise ratio, calibration time, artifact susceptibility, and inter-user variability remain open areas of investigation, particularly for mobile and self-paced BCI systems [11, 12].

Within this framework, the Augmented Reality for Health Monitoring laboratory (ARHeMlab) research group contributes to the advancement of wearable BCIs, focusing on the metrological validation of consumer-grade components for BCI, the development of neurofeedback protocols based on dry sensor technology, and the integration of BCIs into immersive and adaptive environments. For instance, previous works involved engagement monitoring in educational and rehabilitative settings, rehabilitation protocols with multimodal sensory feedback, and real-time interaction. Emphasis is placed on uncertainty-aware signal acquisition, signal processing pipelines tailored to low-channel systems, and user-centred experimental design.

This paper thus gives an overview of the key advancements from the ARHeMlab group in contributing to the progression of wearable BCI research by tackling three

core aspects: (i) signal acquisition through wearable and portable EEG setups, (ii) processing strategies suitable for low-channel setting and possibly real-time classification, (iii) sensory stimulation and/or feedback designed for natural and engaging user interaction. Findings are supported by experimental validation and application-specific insights, illustrating a cohesive trajectory toward daily-life BCI deployment. It is worth remarking that only research from the ARHeMlab research group have been reviewed, by particularly contextualizing the findings in the field of wearable BCI systems.

WEARABLE BCI ARCHITECTURES

This section concerns the approaches to building wearable BCI systems by adopting consumer grade EEG hardware, developing processing pipeline tailored to few acquisition channels, and providing sensory stimulation and/or feedback mainly through XR. The reference architecture is reported in Fig. 1 for the sake of clarity.

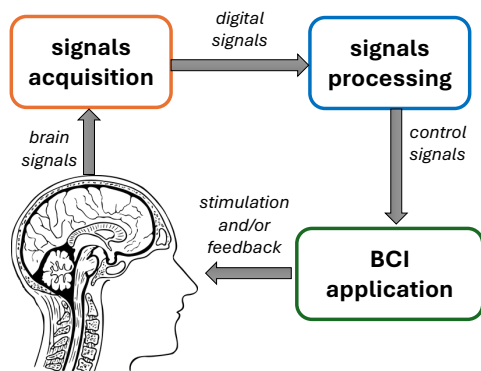


Figure 1: Essential block diagram of a wearable brain-computer interface.

Signal acquisition: To ensure usability in real-world contexts, the signal acquisition strategy for the developed BCI systems emphasizes practical wearability without compromising essential signal quality. Proposed architectures involve three main pillars: a reduced number of channels, the use of suitable commercial devices, and a strong focus on metrological aspects to address performance limitations.

In contrast to traditional EEG systems that rely on high-density montages, proposed designs prioritize a low number of electrodes (typically 1 to 8), enabling faster setup times, reduced discomfort, and enhanced mobility [13–15]. Electrode placement is task-specific: occipital positions are preferred for Steady-State Visually Evoked Potentials (SSVEPs), sensorimotor areas are targeted for Motor Imagery (MI)-based applications, while the frontal area is privileged for cognitive load assessment [16]. Despite the inherent trade-off in spatial resolution, the studies have demonstrated that minimal setups can still bring to meaningful classification accuracy in various contexts, including XR-enhanced interaction, motor re-

habilitation, and engagement monitoring [13, 15, 17–19]. This is partly due to the temporal resolution, where sampling rates of the chosen devices span from 125 Sa/s to 1024 Sa/s.

To promote scalability and lower costs, commercial-grade wireless EEG systems were adopted. The chosen instrumentation often features dry electrodes. Acquisition devices, selected also for their ergonomic design, were EEG-SMT by Olimex, Helmate by abmedica, Emotiv Epoch, FlexEEG by Neuroconceive, and Unicorn Hybrid Black by g.tec. Safety standards and real-time data acquisition capabilities were also taken into account. These systems facilitate the BCI usage even where user movement and portability are foreseen [15], and apart from the Helmate case, they support integration with wearable XR interfaces [18, 20].

Recognizing the limitations introduced by wearable EEG systems, metrological characterizations were conducted for the selected equipment. Evaluated performances were related to acquisition latencies, sampling stability, linearity errors, and amplitude uncertainties [20, 21]. Particular attention was paid to the impact of wireless communication and electrode-skin interface, both of which can affect the detection of time-locked phenomena and real-time usage.

The discussed approach to signal acquisition aims to the identification of bottlenecks to performance while enabling wearable and user-friendly solutions for continuous daily-life applications. By tackling the mentioned aspects, disclosed out-of-the-lab scenarios. Nonetheless, optimizing signal-to-noise ratio also requires channel selection and artifact removal by involving tailored signal processing pipelines, which are ultimately devoted to a proper classification of the acquired brain signals.

Signal processing:

Artifacts removal is a fundamental step to detect neural activity inside the signals acquired via wearable and portable systems. Given the adoption of low-density EEG setups, well-known techniques like Independent Component Analysis, Principal Component Analysis, and Artifact Subspace Reconstruction (ASR), were firstly investigated. ASR demonstrated superior performance in preserving EEG integrity when at least four electrodes were available [22]. Instead, for conditions involving fewer sensors, a hybrid technique integrating ASR with multivariate empirical mode decomposition was developed to deal with ocular, blink, and muscular artifacts [23]. Currently, techniques relying on Riemannian geometry are also under investigation.

In reduced EEG channels sets, carefully choosing the ones to install is also crucial to optimize the available brain signals information. As anticipated above, electrodes placement can be done with a-priori knowledge on brain areas. Nonetheless, data-driven selection was investigated as well [13]. Sequential feature selection approaches were implemented in conjunction with different classifiers (k-NN, SVM, LDA, RF, and ANN), and a leave-one-subject-out cross-validation scheme was

adopted to choose subject-independent channels [24]. While previous works had already supported the placement around sensorimotor area for MI, electrodes in the frontal and parietal areas consistently emerged as the most informative for cognitive load-related EEG analysis in more recent studies. Such evidences are in accordance with multiple studies found in literature [10, 13].

Multiple classical machine learning algorithms were adopted for feature selection and classification. In active and passive paradigms, feature extraction was largely conducted using by Filter Bank Common Spatial Pattern, which applies Chebyshev band-pass filters followed by spatial projection to maximize inter-class variance [19, 25]. The selected features were classified through SVM, LDA, or Naive Bayesian Parzen Window, by considering computational constraints especially in real-time operation [26]. On the other side, studies for reactive BCI demonstrated that even simple classifiers like k-NN can outperform canonical methods such as Canonical Correlation Analysis (CCA) when optimized feature sets are used, providing a compelling case for ML in low-complexity wearable systems [18, 27].

Within the passive paradigm, the Sequential Feature Selection (SFS) algorithm was employed to identify optimised subsets of EEG features in the feature selection stage. SFS is a wrapper-based method that, given a target classifier, iteratively selects the most informative features, maximizing the classifier's performance by adding them one at a time (*forward selection*). SFS helps reduce overfitting and enhances model generalisation by tailoring feature selection to the specific classifier architecture [24].

In parallel to classical ML, deep learning models were evaluated for their capability to automatically learn feature representations. These architectures were particularly effective in scenarios involving emotion classification or mental state detection, leveraging both spatial and spectral information in an end-to-end fashion [28]. When embedded in real-time pipelines, simplified versions of FBCSP with minimal filter sets are sought to be combined with deep neural classifiers to achieve real-time BCI control [26]. Finally, Convolutional Variational Autoencoder were also investigated to boost artifacts removal [29].

Given the high inter-subject variability of EEG signals, domain adaptation (DA) and transfer learning methods are crucial for achieving generalizable models. Feature-based DA techniques such as Subspace Alignment and Transfer Component Analysis were investigated to align source and target data distributions prior to classification [30]. In the deep learning context, Domain-Adversarial Neural Networks (DANN) were employed to jointly optimize class separation and domain invariance. The use of unsupervised DA—leveraging unlabeled target data during training—proved particularly suitable for wearable BCIs, where session calibration time is often limited. Comparative ablation studies confirmed the performance gains offered by DA, with certain normalization

strategies (e.g., Stratified Normalization) in some cases outperforming traditional DA methods alone.

Stimulation and feedback:

XR has been increasingly integrated to enhance both sensory stimulation and sensory feedback in BCI applications. This technology is devoted to creating realistic and adaptive environments that foster user engagement and facilitate neuroplasticity.

XR devices such as the Epson Moverio BT-350, Oculus Rift S, and Microsoft HoloLens 1 were employed to deliver targeted visual stimulation in reactive paradigms [18, 31]. Each device offers different capabilities in terms of field of view, refresh rate, and see-through technology. For instance, the Moverio and HoloLens are optical see-through devices, while the Oculus Rift S was adapted into a video see-through system. These headsets enabled the presentation of flickering visual stimuli for SSVEP elicitation at specific frequencies (e.g., 8–15 Hz), implemented using Unity3D.

In terms of feedback, XR systems delivered multimodal neurofeedback by combining visual and haptic modalities. Visual feedback was typically rendered as the motion of a virtual object (e.g., a bar, a ball or game avatar) within a Unity-based virtual environment [15, 25, 26]. The object's behaviour—such as direction and velocity—was dynamically modulated based on real-time EEG classification outputs, where the predicted class dictated movement direction and the classification confidence influenced intensity or speed. For haptic feedback, wearable vibrotactile technology like the bHaptics TactSuit X40 was employed. Vibrations were mapped to the user's torso based on the classified EEG signals, with direction and intensity corresponding to BCI output [15, 25].

XR feedback was also integrated into adaptive gaming environments, where BCI control allowed users to navigate avatars through virtual tasks such as collecting coins [26]. These scenarios included progressively challenging levels with changing environmental conditions (mainly increased speed). Feedback was tightly coupled with performance metrics such as score and time, offering real-time adjustments to maintain an engaging but manageable difficulty level.

APPLICATIONS AND ACHIEVEMENTS

This section reports the results achieved with the described architectures. These are grouped by adopted paradigm. The Institutions' Ethical Review Boards approved all experimental procedures involving human subjects for the respective works recalled in the following.

Reactive paradigms:

In the development of reactive BCIs, recent research activities in ARHeMLab have focused on SSVEP signals, with particular emphasis on the generation of visual stimuli through XR technology as an alternative to traditional computer screen (CS)-based setups. In fact, the recent advancements in XR technology, particularly in terms of pixel resolution and field of view (FoV), have made

XR head-mounted displays (HMDs) a viable alternative for presenting visual stimuli, thereby enhancing system portability and immersivity and extending the applicability of SSVEP-based BCIs beyond controlled laboratory settings [14].

However, the use of XR-based systems introduces several challenges that can impair accurate target detection. Firstly, visual stimuli rendered in XR typically exhibit a degree of transparency to allow for superimposition onto the physical environment. This leads to a reduced contrast compared to CS setups, where the background is generally black and the stimuli are fully opaque. The consequent decrease in the signal-to-noise ratio (SNR) significantly affects system performance, particularly under ambient illumination levels exceeding 600 lux [32]. Another major issue relates to the hardware limitations inherent to XR HMDs, which often suffer from frame rate instability relative to their nominal specifications [18, 33]. Such fluctuations can lead to undesired shifts in the frequency of the flickering stimuli, thereby reducing the accuracy of target classification. As shown in Fig. 2, given an XR device with a nominal refresh rate of 30 Hz and programmed to generate visual stimulation at 15 Hz, a frame rate error resulting in 32 frames per second (fps) produces an SSVEP frequency of 16 Hz, thus compromising the system's reliability. Therefore, measuring frame rate stability is essential and requires careful optimization of the software used to deliver the stimuli.

To address these limitations, several mitigation strategies have been developed. In [18], a classification algorithm capable of adapting to the actual frame rate of the AR device was proposed, leading to significant performance improvements over an otherwise identical XR-based system lacking this self-adaptive capability. Furthermore, in [27], the use of machine learning-based classifiers was investigated as an alternative to conventional techniques such as CCA. The results demonstrated clear enhancements in classification accuracy through the application of machine learning methods.

Overall, the incorporation of frame rate error awareness into stimulus generation, coupled with the implementation of self-adaptive classification algorithms, including those based on Machine Learning, has enabled the development of highly wearable XR-based SSVEP BCIs. These systems can operate with a reduced number of EEG channels, including single-channel configurations, while achieving performance levels comparable to state-of-the-art CS-based systems.

Passive paradigms:

Within the context of BCIs, passive paradigms enable continuous monitoring of users' cognitive states without requiring an intentional interaction. These paradigms rely on EEG signals to detect conditions such as fatigue, stress, or distraction, and offer promising applications in both clinical and operational domains. In this perspective, the study in [24] investigated cognitive load in neurosurgeons engaged in fine motor tasks. Six trainee neurosurgeons were monitored using an eight-channel EEG

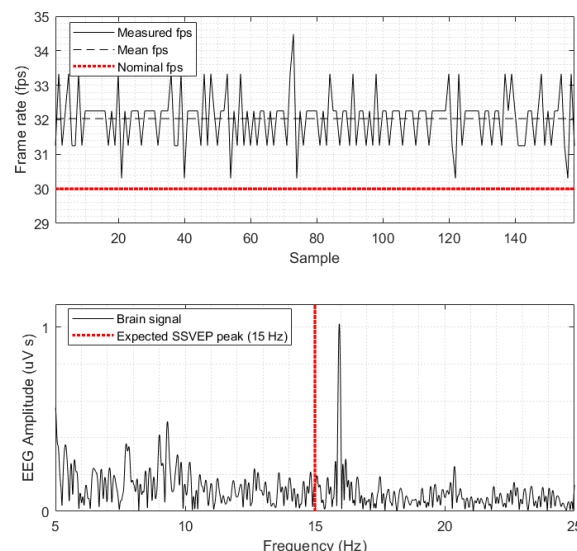


Figure 2: Example of fps variation in an XR HMD used to generate SSVEP stimuli.

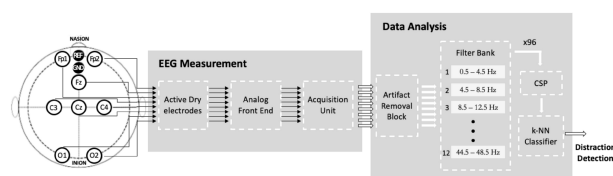


Figure 3: The proposed distraction-detection method (CSP: Common Spatial Pattern algorithm).

system while performing the Purdue Pegboard Test at two difficulty levels. EEG analysis, supported by SFS and supervised classifiers, identified relative alpha power in Fz, O1, and O2 as key markers of mental workload, achieving up to 100 % accuracy in detecting cognitive load levels. The employed lightweight, wireless system allows for real-time monitoring of cognitive fatigue, a common condition in neurosurgical practice.

In [34], a wearable EEG device-based method for detecting distraction during motor rehabilitation exercises was proposed (Fig. 3). An average classification accuracy of 92.8 % in distinguishing attentive from distracted states was achieved by combining a 12-band filter bank with the Common Spatial Pattern (CSP) algorithm. The EEG device's high wearability makes it suitable for clinical implementation, providing therapists with real-time information to manage and enhance patient engagement.

Both studies demonstrate how passive BCIs, implemented through wearable EEG technologies, can monitor cognitive states during motor activity. While the former focuses on preventing mental overload in high-stakes surgical settings, the latter addresses patient engagement in rehabilitation contexts. These contributions underscore the potential of passive paradigms to support adaptive interventions across several real-world healthcare environments.

Active paradigms:

In exploring wearable BCIs based on motor imagery for daily-life contexts, initial efforts focused on evaluating signal processing and classification strategies for motor imagery detection as they were proposed and validated in scientific literature. Both knowledge-based and deep learning approaches demonstrated comparable performance in the 83 % to 93 % range. Despite requiring feature engineering, traditional methods like CSP combined with LASSO and SVM offered low-latency and efficient solutions especially for binary classification. In contrast, deep learning strategies provided superior performance in multi-class problems by directly learning representations from raw EEG data. Despite the results, data augmentation and transfer learning are under exploration to mitigate data scarcity and improve model generalization.

Building upon this, the next phase involved incorporating neurofeedback to enhance user engagement and measurement quality. A wearable closed-loop BCI was developed using wet and then dry EEG sensors, as well as a vibrotactile chest-based feedback system. Visual, haptic, and combined feedback modalities were tested under a synchronous paradigm. Notably, multimodal feedback tended to improve classification accuracy by 5–12 % with respect to the absence of feedback (statistical significance confirmed by permutation test with $\alpha = 5\%$). Moreover, dry sensors facilitated usage but introduced artifacts, particularly in long-haired users or when vibration was activated for the haptic feedback. Still, user feedback via system usability and user experience questionnaires indicated overall system usability and comfort, despite increased perceived effort (NASA-TLX) in the neurofeedback group.

To enhance natural interaction and reduce dependence on external triggers, the study progressed to implementing a self-paced BCI framework. A gamified virtual reality environment allowed participants to control an avatar using continuous MI commands without explicit cues. Among 23 participants, a $54\% \pm 15\%$ composite classification accuracy (3 balanced classes) was achieved during system calibration. Yet, most users were able to properly control the avatar through reaching the targets as reconstructed in Fig. 4. Future work will target reducing calibration time and improving feature separability through refined models and transfer learning approaches, so as to keep developing scalable and practical BCI applications for real-world scenarios.

CONCLUSIONS

This contribution described ARHeMlab's research activities from 2021 to 2025 focusing on the deployment of BCI systems suitable for use outside the laboratory. By integrating wearable and portable EEG with multimodal XR-based feedback and machine learning-enhanced signal processing, the group addressed major limitations in wearability, usability, and robustness. The results validate the feasibility of real-time, low-channel

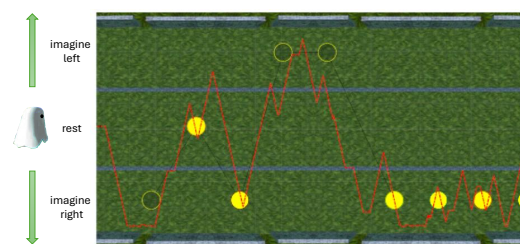


Figure 4: Reconstructed path (red line) of the ghost avatar controlled by a wearable active BCI. Collected targets are in yellow, missed ones in gray.

BCI implementations across reactive, passive, and active paradigms. Future work will focus on reducing calibration times, improving cross-subject generalization through transfer learning, and optimizing signal acquisition under unconstrained conditions, paving the way for widespread application in healthcare, entertainment, and industrial sectors.

Further recent technologies are currently under exploration by the ARHeMlab group working on the wearable BCI domain. These include novel XR visors (e.g., Hololens 2 and Meta Quest 3) with the addition of vibrotactile feedback, as well as advanced signal processing techniques, such as task-related component analysis, filter-bank CCA, data augmentation, and adversarial neural networks. On the other hand, the inter-subject and intra-subject performance will need deeper investigations by extending the reviewed results to different sensory feedbacks and by greater sample sizes.

REFERENCES

- [1] Saha S et al. Progress in brain computer interface: Challenges and opportunities. *Frontiers in systems neuroscience*. 2021;15:578875.
- [2] Ramadan RA, Vasilakos AV. Brain computer interface: Control signals review. *Neurocomputing*. 2017;223:26–44.
- [3] Ding Y, Udompanyawit C, Zhang Y, He B. Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nature Communications*. 2025;16(1):1–20.
- [4] Minguillon J, Lopez-Gordo MA, Pelayo F. Trends in eeg-bci for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*. 2017;31:407–418.
- [5] Hairston WD et al. Usability of four commercially-oriented eeg systems. *Journal of neural engineering*. 2014;11(4):046018.
- [6] Käthner I et al. A multifunctional brain-computer interface intended for home use: An evaluation with healthy participants and potential end users with dry and gel-based electrodes. *Frontiers in Neuroscience*. 2017;11:286.
- [7] Zhu HY, Hieu NQ, Hoang DT, Nguyen DN, Lin CT. A human-centric metaverse enabled by brain-computer

interface: A survey. *IEEE Communications Surveys & Tutorials*. 2024.

[8] Harvie DS, Kelly J, Kluver J, Deen M, Spitzer E, Coppieters MW. A randomized controlled pilot study examining immediate effects of embodying a virtual reality superhero in people with chronic low back pain. *Disability and Rehabilitation: Assistive Technology*. 2024;19(3):851–858.

[9] Abiri R, Borhani S, Sellers EW, Jiang Y, Zhao X. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of neural engineering*. 2019;16(1):011001.

[10] Arpaia P, Esposito A, Natalizio A, Parvis M. How to successfully classify eeg in motor imagery bci: A metrological analysis of the state of the art. *Journal of Neural Engineering*. 2022;19(3).

[11] Mason SG, Birch GE. A brain-controlled switch for asynchronous control applications. *IEEE Transactions on Biomedical Engineering*. 2000;47(10):1297–1307.

[12] Choi J, Kim KT, Jeong JH, Kim L, Lee SJ, Kim H. Developing a motor imagery-based real-time asynchronous hybrid bci controller for a lower-limb exoskeleton. *Sensors*. 2020;20(24):7309.

[13] Apicella A, Arpaia P, Isgro F, Mastrati G, Moccaldi N. A survey on eeg-based solutions for emotion recognition with a low number of channels. *IEEE Access*. 2022;10:117411 – 117428.

[14] Angrisani L, Arpaia P, De Benedetto E, Duraccio L, Regio FL, Tedesco A. Wearable brain-computer interfaces based on steady-state visually evoked potentials and augmented reality: A review. *IEEE Sensors Journal*. 2023;23(15):16501 – 16514.

[15] Arpaia P et al. Paving the way for motor imagery-based tele-rehabilitation through a fully wearable bci system. *Sensors*. 2023;23(13).

[16] Friedman NP, Robbins TW. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*. 2022;47(1):72–89.

[17] Apicella A, Arpaia P, Giugliano S, Mastrati G, Moccaldi N. High-wearable eeg-based transducer for engagement detection in pediatric rehabilitation. *Brain-Computer Interfaces*. 2022;9(3):129 – 139.

[18] Arpaia P, De Benedetto E, De Paolis L, D’Errico G, Donato N, Duraccio L. Performance enhancement of wearable instrumentation for ar-based ssvep bci. *Measurement: Journal of the International Measurement Confederation*. 2022;196.

[19] Apicella A, Arpaia P, Frosolone M, Improta G, Moccaldi N, Pollastro A. Eeg-based measurement system for monitoring student engagement in learning 4.0. *Scientific Reports*. 2022;12(1).

[20] Arpaia P, Callegaro L, Cultrera A, Esposito A, Ortolano M. Metrological characterization of consumer-grade equipment for wearable brain-computer interfaces and extended reality. *IEEE Transactions on Instrumentation and Measurement*. 2022;71.

[21] Arpaia P, Gargiulo L, Moccaldi N, Nalin M, Picciafuoco M. A critical note to iec 80601-19 standard for eeg

wireless transducers about trigger uncertainty in event-related potential measurement. *IEEE Transactions on Instrumentation and Measurement*. 2025;74.

[22] Cataldo A et al. A method for optimizing the artifact subspace reconstruction performance in low-density eeg. *IEEE Sensors Journal*. 2022;22(21):21257–21265.

[23] Arpaia P, Esposito A, Natalizio A, Parvis M, Pesola M. Low-density eeg correction with multivariate decomposition and subspace reconstruction. *IEEE Sensors Journal*. 2023;23(19):23621 – 23628.

[24] Arpaia P et al. Specific feature selection in wearable eeg-based transducers for monitoring high cognitive load in neurosurgeons. *Computer Standards and Interfaces*. 2025;92.

[25] Arpaia P, Coyle D, Donnarumma F, Esposito A, Natalizio A, Parvis M. Visual and haptic feedback in detecting motor imagery within a wearable brain–computer interface. *Measurement: Journal of the International Measurement Confederation*. 2023;206.

[26] Arpaia P, Esposito A, Galasso E, Galdieri F, Natalizio A. A wearable brain-computer interface to play an endless runner game by self-paced motor imagery. *Journal of Neural Engineering*. 2025;22(2).

[27] Apicella A et al. Enhancement of ssveps classification in bci-based wearable instrumentation through machine learning techniques. *IEEE Sensors Journal*. 2022;22(9):9087 – 9094.

[28] Lotte F et al. A review of classification algorithms for eeg-based brain–computer interfaces: A 10 year update. *Journal of neural engineering*. 2018;15(3):031005.

[29] Criscuolo S, Apicella A, Prevete R, Longo L. Interpreting the latent space of a convolutional variational autoencoder for semi-automated eye blink artefact detection in eeg signals. *Computer Standards & Interfaces*. 2025;92:103897.

[30] Apicella A et al. Domain adaptation for fear of heights classification in a vr environment based on eeg and ecg. *Information Systems Frontiers*. 2025;27(1):139 – 154.

[31] Arpaia P, De Benedetto E, De Paolis L, D’errico G, Donato N, Duraccio L. Performance and usability evaluation of an extended reality platform to monitor patient’s health during surgical procedures. *Sensors*. 2022;22(10).

[32] Zhang R et al. Improving ar-ssvep recognition accuracy under high ambient brightness through iterative learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2023;31:1796–1806.

[33] Ke Y, Liu P, An X, Song X, Ming D. An online SSVEP-BCI system in an optical see-through augmented reality environment. *Journal of neural engineering*. 2020;17(1):016066.

[34] Apicella A, Arpaia P, Frosolone M, Moccaldi N. High-wearable eeg-based distraction detection in motor rehabilitation. *Scientific Reports*. 2021;11(1):5297.

Evaluating Baseline Correction Techniques for EEG Time-Frequency Feature Classification

Manuel Hons¹, Silvia Erika Kober¹

¹ Department of Psychology, University of Graz, Graz, Austria

E-mail: manuel.hons@uni-graz.at

ABSTRACT: EEG-based classification is sensitive to preprocessing procedures. Although baseline correction is widely applied in EEG research, its impact on classification performance remains underexplored.

We compared four baseline strategies: (I) amplitude-domain correction, (II) normalized power-domain correction, (III) combined amplitude- and power-domain correction, and (IV) no baseline correction. Analyses were conducted using both an empirical speech imagination dataset and simulated data.

Across both datasets, power-domain baseline correction produced significantly lower classification accuracies than amplitude-domain correction and no baseline correction. The replication of this pattern in simulated data indicates that these differences reflect fundamental signal-processing properties rather than dataset-specific confounds.

These results suggest that baseline correction applied after nonlinear power transformation can distort feature distributions, whereas amplitude-domain correction removes additive offsets prior to nonlinear expansion, thereby preserving more stable variance structures.

INTRODUCTION

Electroencephalogram (EEG) signal processing comprises a multi-stage preprocessing pipeline that can substantially affect downstream machine learning classification performance [1]. Among these stages, baseline correction procedures as well as their concrete positions in the preprocessing sequence received limited attention, despite the potential for nonlinear interactions that may alter feature distributions [1,2]. Different baseline correction procedures have been shown to produce substantially different outcomes [2,3], with some authors even reporting the presence of illusory interactions arising from the usage of inadequate baseline correction techniques [4]. Machine learning algorithms are highly sensitive to preprocessing order and feature normalization, thus, inappropriate sequencing can reduce signal-to-noise ratios and impair model training [5].

One standard feature type commonly used in EEG-BCI research are time-frequency features, which are typically extracted by means of short-time Fourier, Hilbert or wavelet transforms. The resulting estimated band-specific power features are then fed into classifiers (e.g., linear discriminant analysis or support vector

machines) to map features to labels [6,7]. Although certain neuroscience subfields, such as event-related potential (ERP) research, commonly adhere to relatively standardized processing pipelines [8], the decision to apply trial-based baseline correction remains a matter of ongoing scientific debate, even within this domain [9,10]. While some authors advocate the usage of baseline correction [1,10] and state that baseline correction can complement high-pass filtering [11], others doubt the usefulness or even warn of adverse results when used in conjunction with high-pass filters [9]. This ambiguity can also be found in the time-frequency feature realm of EEG research, where some authors endorse the application of trial-based baseline correction [2,3,12,13], while others report no usage of baseline correction techniques [14,15]. Variation also exists with respect to the mode of baseline correction: some have adopted the ERP ‘standard’, performing subtractive pre-stimulus baseline correction in the amplitude-domain [12,13,16,17] followed by power transformation, while others have adopted single-trial spectral baseline correction techniques, which are applied after power transformation of the EEG signal ([2,18–20]). The latter represent techniques for time-frequency feature extraction which were derived from established spectral metrics with integrated baseline correction, such as ERSPs [21] and event-related (de-) synchronization (ERDS) [22]. These measures were originally designed to incorporate baseline correction based on signals averaged across multiple trials but have since been adopted for single-trial application [2,3]. Ultimately, clarifying the optimal arrangement and methods used in the preprocessing pipeline is essential to maximize classification accuracy and to enhance the robustness and generalizability of EEG-based systems across subjects and recording conditions [1–3].

In the current study, we explored the effects that four trial-wise baseline strategies had on classification performance based on EEG-based imagined speech data: I. Subtractive baseline correction based on amplitude signals followed by power transformation. II. Baseline-normalized baseline correction in the power domain [2]. III. A combination of baseline-correction in the amplitude- and the power domain. This version represents a mixture of I and a non-normalized version of II [3]. IV: No baseline correction was applied. One empirical dataset was utilized, made available by Hons and colleagues [12,23]. Secondly, we explored the same

mechanisms based on simulated data designed to reflect EEG data. We used a normalized approach in II instead of a simple baseline subtraction to keep the number of statistical contrasts to a minimum and thereby not further reduce the power of the analysis.

MATERIALS AND METHODS

Sample The publicly available empirical dataset [23] was collected by the authors and comprised data of 22 participants between the ages of 19 and 37 ($M = 25.41$, $SD = 4.02$). All participants were native German speakers, showed no hearing problems and had normal or corrected-to-normal vision.

Experimental design The dataset was recorded during speech imagination: Participants were instructed to imagine speaking one of four different phonemes /a/, /i/, /b/ and /k/. The dataset consisted of 5 second trials, with 1 second baseline intervals. Each block consisted of 20 trials representing only one phoneme. Ultimately, a total of 80 trials were collected for each phoneme.

Data simulation We simulated data of 22 subjects consisting of 10 channels for each frequency band of interest, which were alpha (8-12 Hz), beta (12-30 Hz) and gamma (30-70 Hz). For each of the four classes, 80 trials were created. First, we simulated one sine wave per frequency band by means of the formula in (1). The specific frequency f was drawn for each subject from a uniform distribution limited by the minimal and maximal band frequency.

$$y(t) = A \sin(2\pi ft + \phi) \quad (1)$$

Here, A denotes the amplitude, f denotes frequency, t denotes time points, and ϕ denotes phase. Phase values were randomly drawn from a uniform distribution from 0 to 2π . Subsequently, normalized colored noise was added, such that the power of the noise was proportional to the inverse of the frequency (2), with P denoting power and f denoting frequency.

$$P(f) \propto 1/f \quad (2)$$

To introduce class-specific activity that enables downstream classification of the signals, anchor amplitudes of the produced sine waves systematically differed between classes (Tab. 1). These anchor amplitudes defined the means of a normal distribution with standard deviation of 0.4 from which the actual amplitudes were drawn for each subject independently. The rationale behind this approach was to maintain a level of variation across subjects both in the frequency and amplitude domain.

Table 1: Anchor amplitudes for each frequency band and class

Dataset	Class 1	Class 2	Class 3	Class 4
Alpha	2.2	2.4	2.6	2.8
Beta	1.2	1.4	1.6	1.8
Gamma	0.2	0.4	0.6	0.8

EEG preprocessing The data were processed in Python 3.12 by means of the mne package (version 1.7) [24]. The initial part of the pipeline was only applied to the empirical dataset: the data were re-referenced to the bilateral mastoid averages and were then corrected for ocular artifacts by means of Gratton-Coles regression [25]. Subsequently, the data were visually inspected to mark trials corrupted by motor, technical or residual ocular artifacts for removal. From this point onward, the preprocessing pipeline was applied to the simulated data as well. The data were filtered for the frequency bands of interest, which were alpha (8-12 Hz), beta (12-30 Hz) and gamma (30-70 Hz). Subsequently, the filtered data were epoched. From here, the data were subjected to four different preprocessing subsequences: I. Baseline correction by subtraction of the average baseline value in the amplitude domain $B_{(f)A}$ from all data points t of the trial data $A_{(t,f)}$, and power-transformed (3).

$$P_{(t,f)I} = (A_{(t,f)} - B_{(f)A})^2 \quad (3)$$

II. The power-transformed trial data $P_{(t,f)}$ were baseline corrected by subtraction of the average power value of the baseline interval $B_{(f)P}$ from all data points t , and normalized by the standard deviation of the baseline interval $SD(B_{(f)P})$ [2].

$$P_{(t,f)II} = (P_{(t,f)} - B_{(f)P}) / SD(B_{(f)P}) \quad (4)$$

III. A combination of non-normalized baseline-correction in the amplitude- and the power domain [3]. The initial step is identical to subsequence I. The power-based signal $P_{(t,f)}$ that we derived from employing subsequence I was then baseline corrected again in the power domain by subtracting the mean power-value of the baseline $B_{(f)P}$.

$$P_{(t,f)III} = P_{(t,f)I} - B_{(f)P} \quad (5)$$

IV. The trial data was power transformed without applying baseline correction. Irrespective of the subsequence performed, this resulted in a feature matrix of shape $m \times n$, with m being the number of trials and n being the number of channels multiplied by the number of frequency bands.

Table 2: Estimated marginal means of all factorial combinations

Dataset	Baseline correction mode	Estimated marginal mean	SE
Empirical	Amplitude	72.01	2.53
Empirical	Power (normalized)	25.60	1.53
Empirical	Combination	28.23	1.31
Empirical	No baseline correction	71.92	2.58
Simulated	Amplitude	72.44	1.22
Simulated	Power (normalized)	23.93	1.21
Simulated	Combination	26.63	1.21
Simulated	No baseline correction	72.51	1.19

Feature extraction and classification pipeline Final feature matrices consisted of trial-wise averaged band power values. The extracted feature matrices were split into training and test set by a ratio of 80/20 for each subject independently. The data were z-transformed based on a scaler fitted on the training data. For classification a Multilayer Perceptron (MLP) with two hidden layers of sizes 200 and 20 and the corresponding ReLU activation function was employed. The stochastic optimization algorithm ‘Adam’ [26] was used and the regularization parameter was set to 10^{-4} .

Statistical analysis To test whether the differences between the different baseline correction modes were significant, repeated measures ANOVA was performed for each of the two datasets. Since the sphericity assumption has been violated in both cases, p-values were Greenhouse-Geisser corrected.

Table 3: Statistical results for the contrasts of interest

Dataset	Contrast	df	t-ratio	p
Empirical	Amplitude - Power (normalized)	21	19.12	<0.001
Empirical	Amplitude - Combination	21	16.44	<0.001
Empirical	Amplitude - No baseline correction	21	0.25	0.80
Empirical	Power (normalized) - Combination	21	-1.57	0.26
Empirical	Power (normalized) - No baseline correction	21	-18.47	<0.001
Empirical	Combination - No baseline correction	21	-16.11	<0.001
Simulated	Amplitude - Power (normalized)	21	26.40	<0.001
Simulated	Amplitude - Combination	21	26.95	<0.001
Simulated	Amplitude - No baseline correction	21	-0.33	0.75
Simulated	Power (normalized) - Combination	21	-1.50	0.30
Simulated	Power (normalized) - No baseline correction	21	-26.84	<0.001
Simulated	Combination - No baseline correction	21	-26.64	<0.001

RESULTS

Tab. 2 displays the estimated marginal means of all possible combinations of the tested baseline correction techniques. There was a significant main effect of baseline correction mode for both the empirical ($F_{3, 63} = 275.03$, $p < 0.001$; Fig. 1A) and the simulated dataset ($F_{1, 63} = 564.42$, $p < 0.001$; Fig. 1B). Regarding both datasets, post-hoc tests (Tab. 3) revealed that amplitude-based baseline correction resulted in significantly higher classification accuracies than normalized power-based

baseline correction (Empirical: $t\text{-ratio}_{21} = 19.12$, $p < 0.001$, $d = 4.08$; Simulated: $t\text{-ratio}_{21} = 26.40$, $p < 0.001$, $d = 5.63$) and the combination of amplitude- and power-based baseline correction (Empirical: $t\text{-ratio}_{21} = 16.44$, $p < 0.001$, $d = 3.51$; Simulated: $t\text{-ratio}_{21} = 26.95$, $p < 0.001$, $d = 5.75$). Similarly, applying no baseline correction yielded higher downstream classification accuracy than normalized power-based baseline correction (Empirical: $t\text{-ratio}_{21} = 18.47$, $p < 0.001$, $d = 3.94$; Simulated: $t\text{-ratio}_{21} = 26.84$, $p < 0.001$, $d = 5.72$) and the combination of amplitude- and power-based baseline correction (Empirical: $t\text{-ratio}_{21} = 16.11$, $p < 0.001$, $d = 3.43$; Simulated: $t\text{-ratio}_{21} = 26.64$, $p < 0.001$, $d = 5.68$). All remaining contrasts were insignificant at the significance level $\alpha = 0.05$.

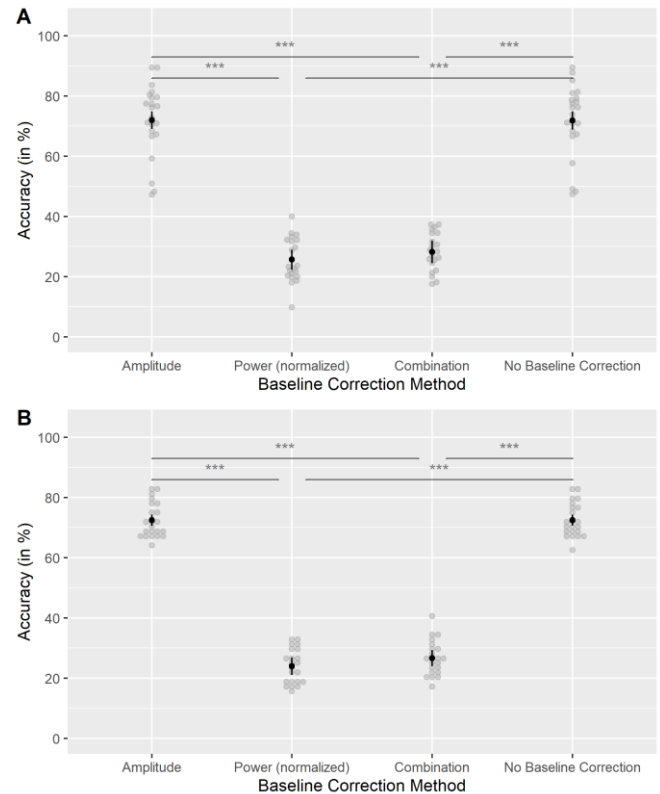


Figure 1: Classification performance of all factorial combinations for **A** the empirical and **B** the simulated dataset. Error margins denote standard errors.

DISCUSSION

The present study examined how different methods of subtractive baseline correction influence classification performance with regards to empirical EEG-based imagined speech data as well as simulated data. Across both datasets, amplitude-based baseline correction yielded higher classification accuracies than alternative approaches in the power-domain. Notably, this superiority extended not only to normalized power-domain baseline correction, but also to the combination amplitude- and power-based correction. When no baseline correction was performed, a similar superiority over baseline corrections in the power domain was

observed. These findings indicate that the method of baseline normalization constitutes a critical factor of downstream decoding performance.

Nonlinear transformation as a mechanistic explanation A principled explanation for these results lies in the nonlinear transformation of the EEG signal from the amplitude to the power domain. Power is derived from the squared analytic amplitude, thus, any additive offset present at the amplitude level is nonlinearly amplified after squaring. The observation that power-based baseline correction is prone to biases is partially coherent with previous reports: Grandchamp and Delorme [2] found that trial-based normalized subtractive baseline correction in the power-domain led to a positive bias in the averaged ERSP maps compared to alternative procedures, meaning they are disproportionally affected by outliers. Gyurkovics and colleagues [4], on the other hand, showed that divisive baseline correction methods produced illusory interactions in a simulated factorial design, while subtractive baseline correction in the power domain did not. Note that both author groups did not compare their findings to amplitude domain baseline correction.

Baseline correction after power transformation causes residual amplitude-level offsets to be transformed into multiplicative distortions in the power domain. Subtracting a mean baseline from already squared signals does not reverse this nonlinear expansion and potentially culminates in biased feature estimation. In contrast, subtractive correction in the amplitude domain removes additive offsets before nonlinear transformation occurs. As a result, the subsequent squaring operation acts on a signal that is already centred relative to baseline, preserving proportional differences between task-related oscillatory components. The importance of preprocessing order has already been underlined by prior research [1]. The consistently lower performance observed for the combined amplitude–power correction further suggests that baseline correction in the amplitude domain does not stabilize the subsequent power-domain baseline correction. This finding does not cohere with earlier research, in which the application of combined trial-based, subtractive amplitude- and power-domain baseline-correction was shown to be superior to alternative measures, producing no bias in resulting ERD/S maps compared to a trial-based percentage approach [3]. Our results suggest that this reported superiority is not simply transferable to classification tasks.

Non-stationary within session influences EEG recordings are further characterized by slow drifts, 1/f background activity [4], and fluctuations related to vigilance, fatigue [27], and electrode impedance [28]. While slow drifts can be addressed by employing high-pass filters, other signal fluctuations may be more complex and require alternative correction measures, such as baseline correction. When baseline correction is applied exclusively in the power domain, these non-stationary influences may already have been nonlinearly

scaled. Consequently, subtractive power-domain correction may insufficiently remove session-specific variation or may alter the variance structure in a way that increases heterogeneity across trials. This is particularly relevant in BCI contexts, where classifiers operate on single-trial data and are sensitive to distributional instability. The simulated dataset provides additional insight into this mechanism. Because the simulated signals consisted of controlled sinusoidal components combined with 1/f noise, the superiority of amplitude-based correction cannot be attributed to unknown cognitive or physiological confounds. Instead, the replication of the empirical pattern in simulated data supports the interpretation that the effect is driven by fundamental signal-processing properties rather than dataset-specific artifacts. Finally, despite the observation that baseline correction in the amplitude-domain did not exhibit better classification performance compared to no baseline correction, the potential influence of non-stationary, within-session confounds, such as changes in mental fatigue and electrode signal quality, may represent a principle-driven argument for the application of baseline correction rather than the omission thereof. Specifically, if intra-session drifts and influences are pronounced and no baseline correction is performed, artificially enhanced classification accuracy might arise [29], impairing model validity and generalizability. This might be especially relevant when data show pronounced levels of noise. Delorme [29], for instance, reported that referencing measures and baseline correction methods decreased the number of significant trials needed for significant results. Furthermore, Xu and colleagues [30], found that mere temporal autocorrelation in the EEG signal sufficed to successfully classify arbitrarily assigned, block-wise class labels in watermelons. Contrary to popular belief that autocorrelations in the EEG are only present in low-frequency components of the signal, the authors demonstrated that artificial decoding performance could be achieved in all frequency bands and despite employing a high-pass filter.

Methodological implications and reporting standards The findings contribute to the ongoing lack of consensus regarding preprocessing standards in time–frequency EEG classification research. While ERP research has converged on relatively standardized amplitude-domain baseline procedures, spectral feature extraction in BCI research remains heterogeneous. The present results suggest that adopting amplitude-domain baseline correction prior to power transformation may represent a principled and empirically supported strategy for oscillatory feature-based decoding.

The issue of reporting transparency is of equal importance. Given that many EEG classification studies do not explicitly describe baseline handling, it is difficult to assess reproducibility or compare results across studies. The large effect sizes observed here demonstrate that baseline strategy alone can substantially influence performance metrics. Therefore, detailed documentation of preprocessing order and

baseline correction measures should be considered essential for methodological rigor. If no baseline correction was applied, this should be explicitly stated. The rationale for omitting baseline correction should also be provided, for example, if a sufficiently strong high-pass filter when deliberately used to adequately remove slow drifts and DC offsets.

Limitations and future directions Several limitations should be acknowledged. First, only subtractive baseline correction methods were examined. Alternative approaches—such as divisive normalization [2], decibel conversion [21], regression-based drift removal [11], or adaptive baseline modelling [31]—may interact differently with nonlinear transformations. Second, only one classifier architecture was used. Although multilayer perceptrons are sensitive to distributional properties, future studies should test whether similar effects emerge for linear discriminant analysis, support vector machines, or convolutional neural networks. Third, the paradigm focused on imagined speech; generalization to motor imagery, visual paradigms, or resting-state classification requires further investigation. Future research should additionally examine cross-session transfer performance and online BCI implementations to determine whether amplitude-based baseline correction enhances robustness under realistic deployment conditions.

CONCLUSION

In summary, the present findings show that subtractive baseline correction in the amplitude domain, as well as the omission of baseline correction, resulted in superior downstream classification performance compared to baseline correction applied in the power domain or a combined approach. Under conditions of non-stationary intra-session fluctuations - such as mental fatigue or changes in electrode impedance - amplitude-domain baseline correction may be preferable to omitting baseline correction entirely, as it may enhance model validity and generalizability. Overall, these results emphasize the necessity of accounting for nonlinear signal transformations when designing preprocessing pipelines and demonstrate that preprocessing order is a critical determinant of robustness, validity, and interpretability in EEG-based machine learning applications.

ACKNOWLEDGMENTS

The authors acknowledge the support by the Field of Excellence COLIBRI (Complexity of Life in Basic Research and Innovation, University of Graz, Austria).

REFERENCES

[1] Kessler R, Enge A, Skeide MA. How EEG preprocessing shapes decoding performance. *Commun Biol.* 2025;8(1):1039. doi:10.1038/s42003-025-08464-3

[2] Grandchamp R, Delorme A. Single-Trial Normalization for Event-Related Spectral Decomposition Reduces Sensitivity to Noisy Trials. *Front Psychol.* 2011;2. doi:10.3389/fpsyg.2011.00236

[3] Hu L, Xiao P, Zhang ZG, Mouraux A, Iannetti GD. Single-trial time–frequency analysis of electrocortical signals: Baseline correction and beyond. *NeuroImage.* 2014;84:876–887. doi:10.1016/j.neuroimage.2013.09.055

[4] Gyurkovics M, Clements GM, Low KA, Fabiani M, Gratton G. The impact of 1/f activity and baseline correction on the results and interpretation of time-frequency analyses of EEG/MEG data: A cautionary tale. *NeuroImage.* 2021;237:118192. doi:10.1016/j.neuroimage.2021.118192

[5] Bigdely-Shamlo N, Mullen T, Kothe C, Su KM, Robbins KA. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front Neuroinformatics.* 2015;9. doi:10.3389/fninf.2015.00016

[6] Värbu K, Muhammad N, Muhammad Y. Past, Present, and Future of EEG-Based BCI Applications. *Sensors.* 2022;22(9):3331. doi:10.3390/s22093331

[7] Vaid S, Singh P, Kaur C. *EEG Signal Analysis for BCI Interface: A Review.* 2015:147. doi:10.1109/ACCT.2015.72

[8] Jacobsen NSJ, Kristanto D, Welp S, Inceler YC, Debener S. Preprocessing choices for P3 analyses with mobile EEG: A systematic literature review and interactive exploration. *Psychophysiology.* 2025;62(1):e14743. doi:10.1111/psyp.14743

[9] Maess B, Schröger E, Widmann A. High-pass filters and baseline correction in M/EEG analysis. Commentary on: “How inappropriate high-pass filters can produce artefacts and incorrect conclusions in ERP studies of language and cognition.” *J Neurosci Methods.* 2016;266:164–165. doi:10.1016/j.jneumeth.2015.12.003

[10] Tanner D, Norton JJS, Morgan-Short K, Luck SJ. On high-pass filter artifacts (they’re real) and baseline correction (it’s a good idea) in ERP/ERMF analysis. *J Neurosci Methods.* 2016;266:166–170. doi:10.1016/j.jneumeth.2016.01.002

[11] Alday PM. How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology.* 2019;56(12):e13451. doi:10.1111/psyp.13451

[12] Hons M, Kober SE, Wriessnegger SC, Wood G. Hybrid EEG-fNIRS phoneme classification based on imagined and perceived speech. *Front Neuroergonomics.* 2026;7. doi:10.3389/fnrgo.2026.1696865

[13] Huang J, Li G, Zhang Q, Yu Q, Li T. Adaptive Time–Frequency Segment Optimization for Motor Imagery Classification. *Sensors.*

- 2024;24(5):1678. doi:10.3390/s24051678
- [14] Herman P, Prasad G, McGinnity TM, Coyle D. Comparative Analysis of Spectral Approaches to Feature Extraction for EEG-Based Motor Imagery Classification. *IEEE Trans Neural Syst Rehabil Eng.* 2008;16(4):317-326. doi:10.1109/TNSRE.2008.926694
- [15] Uyulan C, Erguzel TT. Analysis of Time — Frequency EEG Feature Extraction Methods for Mental Task Classification. *Int J Comput Intell Syst.* 2017;10(1):1280-1288. doi:10.2991/ijcis.10.1.87
- [16] Lee SH, Lee M, Lee SW. Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication. *IEEE Trans Neural Syst Rehabil Eng.* 2020;28(12):2647-2659. doi:10.1109/TNSRE.2020.3040289
- [17] Kaya E, Saritas I. Motor Imagery BCI Classification with Frequency and Time-Frequency Features by Using Different Dimensions of the Feature Space Using Autoencoders. *PLUSBASE Akad Organ VE DANISMANLIK LTDSTI.* Published online December 29, 2022;1. doi:10.58190/imiens.2022.2
- [18] Koelstra S, Muhl C, Soleymani M, et al. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Trans Affect Comput.* 2012;3(1):18-31. doi:10.1109/T-AFFC.2011.15
- [19] Ahmed MdZI, Sinha N, Ghaderpour E, Phadikar S, Ghosh R. A Novel Baseline Removal Paradigm for Subject-Independent Features in Emotion Classification Using EEG. *Bioengineering.* 2023;10(1):54. doi:10.3390/bioengineering10010054
- [20] Cho H, Ahn M, Ahn S, Kwon M, Jun SC. EEG datasets for motor imagery brain–computer interface. *GigaScience.* 2017;6(7):gix034. doi:10.1093/gigascience/gix034
- [21] Makeig S. Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones. *Electroencephalogr Clin Neurophysiol.* 1993;86(4):283-293. doi:10.1016/0013-4694(93)90110-H
- [22] Pfurtscheller G, Lopes Da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol.* 1999;110(11):1842-1857. doi:10.1016/S1388-2457(99)00141-8
- [23] Hons M. Hybrid EEG-fNIRS phoneme classification based on imagined and perceived speech. Published online September 2024. doi:10.17605/OSF.IO/F7RQ2
- [24] Gramfort A, Luessi M, Larson E, et al. MEG and EEG data analysis with MNE-Python. *Front Neurosci.* 2013;7. doi:10.3389/fnins.2013.00267
- [25] Gratton G, Coles MGH, Donchin E. A new method for off-line removal of ocular artifact. *Electroencephalogr Clin Neurophysiol.* 1983;55(4):468-484. doi:10.1016/0013-4694(83)90135-9
- [26] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv.* Preprint posted online January 30, 2017;arXiv:1412.6980. doi:10.48550/arXiv.1412.6980
- [27] Tran Y, Craig A, Craig R, Chai R, Nguyen H. The influence of mental fatigue on brain activity: Evidence from a systematic review with meta-analyses. *Psychophysiology.* 2020;57(5):e13554. doi:10.1111/psyp.13554
- [28] da Silva Souto CF, Pätzold W, Scanlon JEM, Winneke AH, Debener S, Wolf KI. Evaluation of Pre-Applied Conductive Materials in Electrode Grids for Longterm EEG Recording. *Sensors.* 2025;25(22):6810. doi:10.3390/s25226810
- [29] Delorme A. EEG is better left alone. *Sci Rep.* 2023;13(1):2372. doi:10.1038/s41598-023-27528-0
- [30] Xu X, Wang B, Xiao B, et al. Beware of Overestimated Decoding Performance Arising from Temporal Autocorrelations in Electroencephalogram Signals. *Biomed Signal Process Control.* 2026;113:108783. doi:10.1016/j.bspc.2025.108783
- [31] Raghu S, Sriraam N, Gommer ED, et al. Cross-database evaluation of EEG based epileptic seizures detection driven by adaptive median feature baseline correction. *Clin Neurophysiol.* 2020;131(7):1567-1578. doi:10.1016/j.clinph.2020.03.033

Riemannian Geometry-Preserving Variational Autoencoder for MI-BCI Data Augmentation

Viktorija Polaka¹, Ivo Pascal De Jong¹, Andreea Ioana Sburlea¹

¹Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands

E-mail: victoria_polaka@proton.me

ABSTRACT: This paper addresses the challenge of generating synthetic electroencephalogram (EEG) covariance matrices for motor imagery brain-computer interface (MI-BCI) applications. *Objective.* We aim to develop a generative model capable of producing high-fidelity synthetic covariance matrices while preserving their symmetric positive-definite nature. *Approach.* We propose a Riemannian geometry-preserving variational autoencoder (RGP-VAE) integrating geometric mappings with a composite loss function combining Riemannian distance, tangent space reconstruction accuracy and generative diversity. *Results.* The model generates valid, representative EEG covariance matrices, while learning a subject-invariant latent space. Synthetic data proves practically useful for MI-BCI, with its impact depending on the paired classifier. *Contribution.* This work introduces and validates the RGP-VAE as a geometry-preserving generative model for EEG covariance matrices, highlighting its potential for signal privacy, scalability and data augmentation.

INTRODUCTION

While Riemannian geometry-based classifiers currently dominate MI-BCI competitions, their advancement towards mainstream applications is hindered by data scarcity and inter-subject variability, which necessitates lengthy calibration sessions [1–4]. Deep learning alternatives have yet to surpass these geometric pipelines, possibly explained by the limited availability of subject-level data [3]. To overcome these limitations, we propose a novel data augmentation framework tailored to the specific geometric properties of EEG covariance matrices, which are symmetric positive-definite (SPD), i.e., symmetric with strictly positive eigenvalues.

Previous work exploring data augmentation directly on the SPD manifold by geometrically interpolating between existing covariance matrices of the same class has successfully boosted BCI classification accuracy in data-scarce scenarios for SSVEP and ERP tasks [5]. However, this approach is fundamentally limited to the convex hull of the original data and thus cannot generate plausible variations that exist in unexplored regions of the manifold. A variational autoencoder (VAE) [6], which can learn a latent representation of a manifold [7], may offer an alternative to overcome this convex-hull limitation

and generate diverse yet plausible synthetic samples that extend beyond the convex hull.

However, standard VAEs assume Euclidean geometry, creating a conflict when working with the Riemannian SPD manifold structure of EEG covariance matrices; applying standard Euclidean operations on this curved manifold causes geometric distortions (e.g., the "swelling effect") [5]. We address this by proposing the Riemannian geometry-preserving VAE (RGP-VAE) designed to preserve geometric integrity, utilizing parallel transport [8] to align data and thus enable the model to learn subject-invariant features. The focus is specifically on the challenging and practically relevant problem of cross-subject generalization, with the aim to reduce the need for extensive calibration [1]. Accordingly, this paper aims to: (1) establish if a Riemannian geometry-preserving VAE can generate valid synthetic EEG covariance matrices, and (2) evaluate whether this synthetic data improves cross-subject MI-BCI performance.

METHODS

Data and Preprocessing: We use the dataset from Faller et al. [9], containing 13-channel EEG recordings from 12 subjects performing a two-class motor imagery task (right hand versus both feet). The data loading procedure, using the "Mother of All BCI Benchmarks" framework [10], resulted in a total of 5572 trials across the 12 subjects (398 or 597 trials per individual).

The EEG trials are bandpass filtered (8–30 Hz) to capture sensorimotor rhythms. The raw voltage signals were then scaled to microvolts (10^6). To address non-stationarity, exponential moving standardization (EMS) [11] is applied (decay factor = 0.01, initial block size = 1000, epsilon = $1e-6$). EMS can be seen applied to data for training deep learning models as these models tend to be sensitive to the input scale [12]. Finally, trials are converted into spatial covariance matrices in $\mathbb{R}^{13 \times 13}$ using the oracle approximating shrinkage estimator [13], yielding well-conditioned SPD matrices.

To address inter-subject EEG variability, which manifests as geometric differences in their location on the Riemannian manifold, parallel transport [8] was applied. This technique geometrically transports matrices from each subject reference mean to a global (class) reference mean via a congruence transformation.

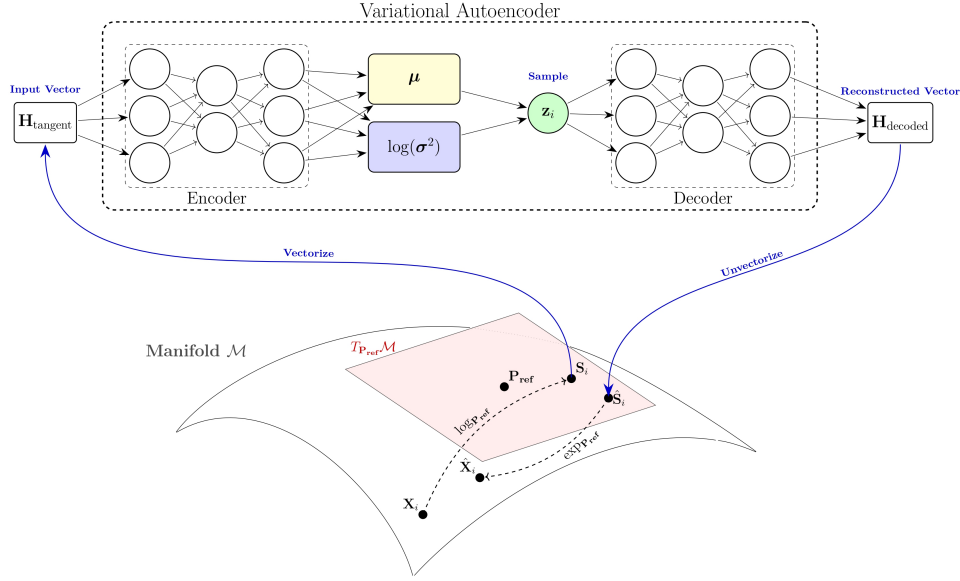


Figure 1: An overview of the proposed RGP-VAE, illustrating the integration of a standard VAE with geometric operations on the SPD manifold. An input SPD matrix $\mathbf{X}_i$ is first projected onto the tangent space at a reference point $\mathbf{P}_{\text{ref}}$ using the logarithmic map $\log_{\mathbf{P}_{\text{ref}}}$ (Eq. 3). This tangent representation $\mathbf{S}_i$ is then vectorized to serve as the encoder input $\mathbf{H}_{\text{tangent}}$. The encoder maps this input to a latent distribution parameterized by μ and $\log(\sigma^2)$, from which a latent vector $\mathbf{z}_i$ is sampled and passed to the decoder to produce the reconstructed vector $\mathbf{H}_{\text{decoded}}$. The vector is unvectorized back into a tangent space representation $\hat{\mathbf{S}}_i$, which is finally mapped back onto the SPD manifold via the exponential map ($\exp_{\mathbf{P}_{\text{ref}}}$) (Eq. 4) to produce the reconstructed SPD matrix $\hat{\mathbf{X}}_i$.

Model Architecture: Conceptually building on prior work on Riemannian variational autoencoders for manifold-valued data [14], the modified VAE (Fig. 1) learns a latent representation $\mathbf{z}$ from SPD matrices by bridging the curved manifold $\mathcal{M}$ and the Euclidean space required by neural networks. The manifold of symmetric positive-definite matrices is defined as $\mathcal{M} = \{\mathbf{X} \in \mathbb{R}^{N \times N} \mid \mathbf{X} = \mathbf{X}^\top, \mathbf{X} \succ 0\}$, where N is the number of EEG channels and $\mathbf{X} \succ 0$ indicates that the matrix is composed of strictly positive eigenvalues [2]. The proposed architecture relies on a class-specific reference point $\mathbf{P}_{\text{ref}}$, calculated as the Riemannian Fréchet mean of the training class. Unlike arithmetic mean, which may not yield a valid SPD matrix, the Fréchet Mean guarantees a valid point on the manifold by minimizing the sum of squared Riemannian distances to all other matrices in a set $\{\mathbf{X}_i\}_{i=1}^{M_{\text{set}}}$.

$$\mathbf{G} = \arg \min_{\mathbf{P} \in \mathcal{M}} \sum_{i=1}^{M_{\text{set}}} d_r^2(\mathbf{P}, \mathbf{X}_i) \quad (1)$$

where $\mathbf{P}$ is the candidate SPD matrix over which the minimization occurs and $d_r(\mathbf{P}, \mathbf{X}_i)$ is the affine-invariant Riemannian metric (AIRM)[5, 15, 16] which defines the distance between two SPD matrices as:

$$d_r(\mathbf{P}_1, \mathbf{P}_2) = \|\log(\mathbf{P}_1^{-1/2} \mathbf{P}_2 \mathbf{P}_1^{-1/2})\|_F \quad (2)$$

The model processes a batch of aligned SPD covariance matrices $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_B\}$, with batch size $B = 128$, selected as a balance between computational load and the requirements for the diversity loss. Each $\mathbf{X}_i$ is projected to the tangent space—a local Euclidean approximation—at the class-specific reference point $\mathbf{P}_{\text{ref}}$ using the loga-

arithmic map [1, 2]:

$$\mathbf{S}_i = \log_{\mathbf{P}_{\text{ref}}}(\mathbf{X}_i) = \mathbf{P}_{\text{ref}}^{1/2} \log\left(\mathbf{P}_{\text{ref}}^{-1/2} \mathbf{X}_i \mathbf{P}_{\text{ref}}^{-1/2}\right) \mathbf{P}_{\text{ref}}^{1/2}. \quad (3)$$

This is implemented via batched whitening followed by the matrix logarithm to support numerical stability by centring operations around the identity.

The resulting batch $\mathcal{S} = \{\mathbf{S}_1, \dots, \mathbf{S}_B\}$ consists of symmetric matrices in the tangent space at $\mathbf{P}_{\text{ref}}$ and each $\mathbf{S}_i$ is vectorized by using only the upper-triangular elements forming the batch of vectors $\mathbf{H}_{\text{tangent}} \in \mathbb{R}^{B \times D_{\text{spd}}}$ (with $D_{\text{spd}} = N(N+1)/2$) as input to the encoder.

The encoder maps this batch of vectors to the parameters $(\mathbf{M}, \log \sigma^2)$ of the latent distribution, where $\mathbf{M} = \{\mu_1, \dots, \mu_B\}$ and $\log \sigma^2 = \{\log \sigma_1^2, \dots, \log \sigma_B^2\}$. The encoder consists of five sequential blocks (linear $\rightarrow$ batch normalization [17] $\rightarrow$ LeakyReLU [18]) with dimensions $D_{\text{spd}} \rightarrow 32 \rightarrow 64 \rightarrow 16 \rightarrow 32 \rightarrow 64$, followed by two separate linear projections to produce $\mu_i, \log \sigma_i^2 \in \mathbb{R}^{D_{\text{lat}}}$ where $D_{\text{lat}} = 64$. Batch normalization stabilizes training [17], while LeakyReLU activations preserve representational capacity by preventing inactive neurons [18]. The batch of latent vectors $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_B\} \in \mathbb{R}^{B \times D_{\text{lat}}}$ is sampled via the reparameterization trick: $\mathbf{z}_i = \mu_i + \varepsilon_i \odot \exp(0.5 \cdot \log \sigma_i^2)$, where $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, allowing gradients to flow back through $\mathbf{M}$ and $\log \Sigma^2$ during training.

The decoder MLP mirrors the encoder structure, mapping the batch of latent vectors $\mathbf{Z}$ back to a batch of decoded vectors $\mathbf{H}_{\text{decoded}} \in \mathbb{R}^{B \times D_{\text{spd}}}$, which is subsequently unvectorized into a batch of symmetric matrices $\hat{\mathbf{S}}' \in \mathbb{R}^{B \times N \times N}$. The decoder output $\hat{\mathbf{S}}'_i$ is explicitly re-symmetrized via $\hat{\mathbf{S}}''_i = (\hat{\mathbf{S}}'_i + (\hat{\mathbf{S}}'_i)^T)/2$ to eliminate any asymmetries. To return to the manifold, we apply the Exponential Map to

each matrix:

$$\hat{\mathbf{X}}_i = \exp_{\mathbf{P}_{\text{ref}}}(\hat{\mathbf{S}}_i) = \mathbf{P}_{\text{ref}}^{1/2} \exp\left(\mathbf{P}_{\text{ref}}^{-1/2} \hat{\mathbf{S}}_i \mathbf{P}_{\text{ref}}^{-1/2}\right) \mathbf{P}_{\text{ref}}^{1/2} \quad (4)$$

Numerical instability caused by floating-point arithmetic can violate the strict SPD constraints, therefore validity is enforced throughout the model architecture and parallel transport. During the matrix exponential computation, eigenvalues are conditionally scaled (threshold $T = 20$) to prevent overflow: if $\lambda_{\max} > T$, all eigenvalues are scaled by $T/\lambda_{\max}$. Throughout all geometric operations, we maintain a numerical threshold $\varepsilon = 10^{-6}$. If the minimum eigenvalue $\lambda_{\min}$ of any intermediate or output matrix falls below ε , we add $(\varepsilon - \lambda_{\min})\mathbf{I}$ to shift all eigenvalues above the threshold, ensuring positive-definiteness.

Training and Optimization: The network is optimized using a loss function L_{total} balancing reconstruction accuracy, latent space regularization, and diversity:

$$L_{\text{total}} = (L_{\text{manifold}} + L_{\text{tangent}}) + \beta L_{\text{KL}} + \gamma L_{\text{diversity}} \quad (5)$$

The reconstruction term combines L_{manifold} which enforces geometric fidelity using the AIRM distance (Eq. 2):

$$L_{\text{manifold}} = \frac{1}{B} \sum_{i=1}^B d_r(\mathbf{X}_i, \hat{\mathbf{X}}_i) \quad (6)$$

and L_{tangent} , which minimized the normalized Euclidean error between original and decoded tangent vectors:

$$L_{\text{tangent}} = \frac{1}{B} \sum_{i=1}^B \frac{\sum_j (h_{\text{decoded},i,j} - h_{\text{tangent},i,j})^2}{\sum_j h_{\text{tangent},i,j}^2 + \varepsilon} \quad (7)$$

where $\varepsilon = 10^{-6}$ for numerical stability. Meanwhile, latent space regularization is achieved through KL divergence L_{KL} toward a standard Gaussian prior:

$$L_{\text{KL}} = \frac{1}{B} \sum_{i=1}^B \left[-0.5 \sum_{k=1}^{D_{\text{lat}}} (1 + \log \sigma_{i,k}^2 - \mu_{i,k}^2 - \exp(\log \sigma_{i,k}^2)) \right] \quad (8)$$

We apply KL cost annealing [19], linearly increasing β from 0.0001 to 0.2 during training to prevent posterior collapse while maintaining reconstruction fidelity.

The diversity loss $L_{\text{diversity}}$ encourages sample diversity by maximizing the geometric volume of generated tangent vectors. Since the determinant of a covariance matrix quantifies the generalized variance (i.e., the volume spanned by data points), maximizing it promotes wider spatial coverage in the tangent space. The loss minimizes the negative log-determinant of the batch covariance of decoded tangent space vectors $\mathbf{H}_{\text{decoded}} \in \mathbb{R}^{B \times D_{\text{spd}}}$:

$$L_{\text{diversity}} = -\log \det(\text{Cov}(\mathbf{H}_{\text{decoded}}^T) + \varepsilon_{\text{cov}} \mathbf{I}) \quad (9)$$

with $\varepsilon_{\text{cov}} = 10^{-6}$ for numerical stability, weighted by an empirically determined $\gamma = 0.035$.

The AdamW optimizer [20] is employed for a fixed 100 epochs with an empirically found learning rate of 1×10^{-4} and a weight decay parameter of 1×10^{-6} . Training is further regularized with gradient clipping (max

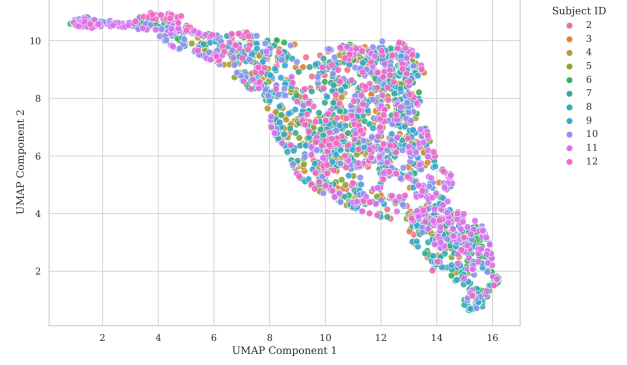


Figure 2: 2D UMAP visualization of the latent space of the RGP-VAE for right-hand movement data. Points are colored by Subject ID; their significant overlap indicates the learning of a subject-invariant representation.

norm=1.0) and learning rate reduction (factor=0.5) after 20 epochs of stagnation.

Data Generation and Evaluation Protocol: leave-one-subject-out cross-validation (LOSO-CV) is employed; in each fold, class-specific RGP-VAEs are trained on aligned data from $N - 1$ subjects to test generalization to unseen individuals. Two synthetic generation strategies are evaluated: Posterior sampling encodes each training matrix $\mathbf{X}_i$, samples $\mathbf{z}_i$ via reparameterization, and decodes to create variations preserving core characteristics of each sample (1:5 real-to-synthetic ratio). Prior sampling draws $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ directly to generate novel samples beyond the training convex hull (5000 per class).

Three Riemannian classifiers—minimum distance to mean (MDM), k-nearest neighbours (KNN; $k = 5$), and support vector classifier (SVC; $C = 1.0$)—are trained and evaluated on held-out test subjects under three conditions: (1) baseline using only the original training data, (2) augmented with synthetic data, and (3) synthetic-only training to assess standalone quality. Balanced accuracy, averaged across all folds, serves as the primary metric.

Synthetic data quality is assessed by verifying SPD properties (symmetry and positive-definiteness), comparing statistical variance (element-wise and global) between real and synthetic matrices, and measuring geometric spread via mean pair-wise Riemannian distances within each class. A scrambled-label diagnostic test confirms that performance degrades to chance level, indicating no spurious correlations.

Source code and additional information can be found at <https://641e16.github.io/RGP-VAE/>.

RESULTS

We validate the proposed RGP-VAE through an assessment of the generated synthetic data fidelity, addressing the fundamental question of whether the model can produce valid and realistic covariance matrices, and with a comparison against a standard VAE approach. Our final analysis quantifies the impact of this data on cross-

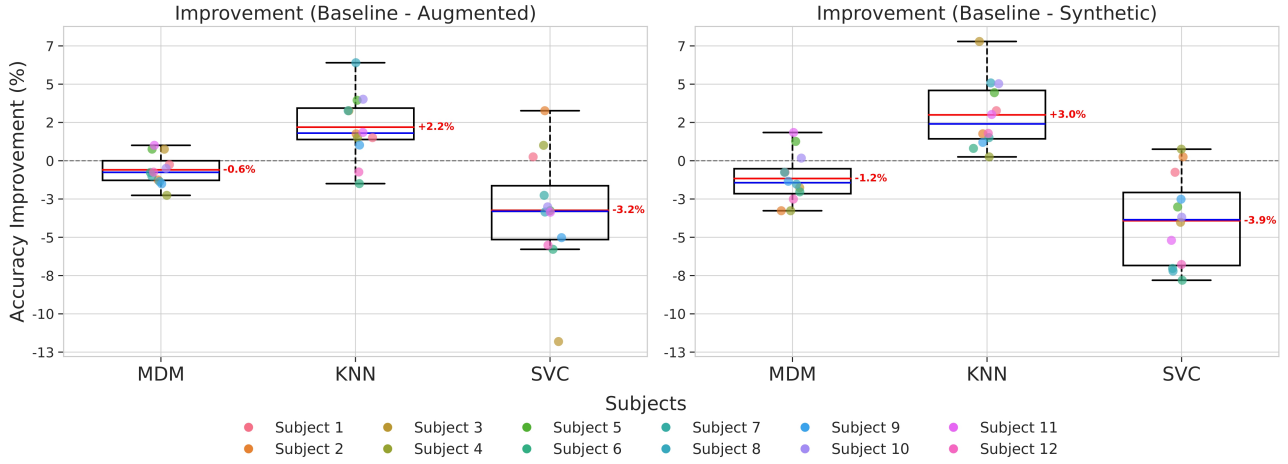


Figure 3: Distribution of accuracy improvement for each classifier using the prior generator. The plot shows the percentage point difference between the ‘Augmented’ and ‘Synthetic-Only’ conditions relative to the ‘Baseline’ across all subjects. The red line signifies the mean whilst the blue line is the median.

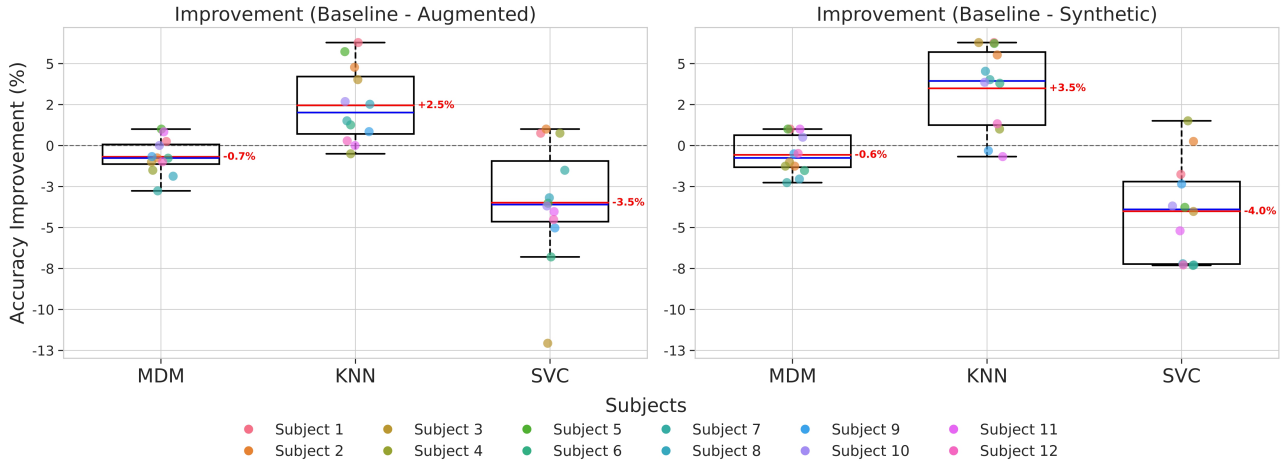


Figure 4: Distribution of accuracy improvement for each classifier using the posterior generator, showing similar trends to the prior generator but with more pronounced fluctuations.

subject classification performance to determine the practical value of the proposed method.

Latent Space Structure: UMAP [21] visualization (Fig. 2) reveals that latent codes organize into a unified structure where subjects are heavily intermingled rather than clustered by individual. This suggests the model learned a largely subject-invariant representation—a critical property enabled by parallel transport alignment—implying generated samples will reflect generalized task patterns rather than subject-specific details.

Fidelity Assessment: Across all folds, 100% of synthetic matrices from both prior and posterior generators passed symmetry and positive-definiteness verification checks, confirming the effectiveness of the architecture’s geometric constraints and numerical stabilisation steps. Tab. 1 compares the statistical variance and geometric spread of synthetic data relative to the original. The chosen γ maintained statistical variance close to the original. To address the lower geometric diversity of synthetic samples, the noise vector is scaled by $\epsilon_i = 2.2$ during

generation, increasing the mean intra-class Riemannian distance to ≈ 1.95 , closely matching the original data’s spread (2.03) without distorting statistical properties.

Cross-Subject Classification Performance: The impact of data augmentation was evaluated by comparing classification accuracies under different augmentation conditions using Wilcoxon signed-rank tests with Bonferroni correction ($p < 0.0083$). As detailed in Tab. 2, data augmentation produced divergent effects. For the KNN classifier, augmentation consistently and significantly improved performance. Posterior-based synthetic-only training yielded the largest gain (+3.49%, $p = 0.002$), while augmented training provided +2.45% ($p = 0.002$). Prior generation produced similar but slightly smaller significant benefits (+3.00% synthetic-only, $p < 0.001$; +2.19% augmented, $p = 0.003$). In contrast, SVC performance significantly degraded with augmentation (up to -4.01%, $p = 0.002$), while MDM remained largely unaffected. Figs. 3 and 4 illustrate subject-wise distributions, revealing high variability: KNN augmentation yielded

Table 1: Fidelity analysis of synthetic data averaged across 12 folds. The table compares the statistical variance ratio and the mean intra-class Riemannian distance, showing that the synthetic data distribution is valid.

Generator	Statistical Variance		Geometric Diversity	
	Original	Synthetic (Ratio)	Original	Synthetic
Prior	0.208	0.221 (1.061)	2.032	1.946
Posterior	0.208	0.221 (1.063)	2.032	1.918

Table 2: Average balanced accuracy (%) across 12 subjects for all training conditions and generators with corresponding p-values.

Generator	Classifier	Baseline	Augmented Scenario			Synthetic-Only Scenario		
		Acc. (%)	Acc. (%)	Improvement	p-value	Acc. (%)	Improvement	p-value
Prior	MDM	59.52 ± 5.52	58.92 ± 5.40	-0.59%	0.092	58.36 ± 5.03	-1.16%	0.043
	KNN	53.19 ± 4.00	55.38 ± 4.17	+2.19%	0.003	56.19 ± 4.19	+3.00%	< 0.001
	SVC	60.67 ± 5.33	57.43 ± 6.32	-3.24%	0.016	56.75 ± 6.37	-3.92%	0.002
Posterior	MDM	59.52 ± 5.52	58.83 ± 5.29	-0.69%	0.092	58.95 ± 5.51	-0.57%	0.151
	KNN	53.19 ± 4.00	55.64 ± 4.13	+2.45%	0.002	56.68 ± 4.06	+3.49%	0.002
	SVC	60.67 ± 5.33	57.18 ± 6.57	-3.48%	0.007	56.66 ± 6.25	-4.01%	0.002

gains up to +7.8% for subject no.3 (prior generation, synthetic only condition). A scrambled label test confirmed classifiers learned meaningful features, yielding chance-level average accuracy on scrambled data (prior: MDM = 50.2%, KNN = 49.9%, SVC = 52.3%; posterior: MDM = 46.6%, KNN = 50.5%, SVC = 47.2%).

Standard VAE Comparison: To validate the Riemannian framework, we compared the proposed RGP-VAE against a standard Euclidean VAE. The standard VAE failed to generate valid data, with > 40% of outputs in every fold violating positive-definiteness. Furthermore, augmenting with the valid portion of its data significantly degraded MDM performance (−9.49%, $p < 0.001$) and offered no statistically significant benefit to KNN or SVC. This confirms that the proposed architecture’s geometric constraints are essential for generating valid and useful SPD matrices in this domain.

DISCUSSION

This study investigated whether the proposed RGP-VAE could generate high-fidelity EEG covariance matrices to improve cross-subject MI-BCI classification.

Generative Fidelity and Validity: A primary contribution of this work is confirming that the RGP-VAE framework inherently generates valid SPD matrices—a non-trivial task where standard Euclidean VAEs failed (producing > 40% invalid matrices). This success is attributable to the underlying Riemannian geometry that enforces the SPD constraint by design. While Parallel Transport enabled a subject-invariant latent space for cross-subject generalization, computing the target-subject reference mean still requires a minimal calibration step prior to online deployment.

While valid, the synthetic data exhibited a slightly elevated statistical variance (ratio ≈ 1.06) but reduced geometric diversity (ratio ≈ 0.95). With the chosen parameters ($\gamma = 0.035$, $\epsilon_i = 2.2$), the model generated more pro-

totypical samples concentrated near the class geometric means rather than spanning the full outlier range of real-world data.

Classifier-Dependent Utility: The impact of the synthetic data was highly divergent, revealing that data augmentation utility is not universal but classifier-dependent. Augmentation yielded statistically significant improvements for the KNN classifier, with posterior sampling boosting performance up to +3.49% ($p = 0.002$). KNN likely benefits because the prototypical synthetic samples densify the class manifolds, creating more dense and reliable local neighbourhoods for distance-based classification. Conversely, performance significantly degraded for the SVC (up to −4.01%, $p = 0.002$). The reduced diversity of synthetic data likely caused the SVC to learn decision boundaries too narrowly fitted around class centres, reducing generalization to boundary-case real samples. Meanwhile, performance for the MDM classifier remained stable—a positive result compared to the standard VAE, which caused a massive degradation (−9.49% under posterior, synthetic-only condition). Unlike the naive Euclidean approach that failed to even generate valid SPD matrices, the RGP-VAE preserved the SPD validity and successfully learnt Riemannian class means. Beyond immediate classification impacts, synthesizing this data holds broader practical value; it provides a mechanism to test pipeline scalability, mitigates data scarcity—possibly for data hungry models—and enables privacy protection by avoiding raw signal sharing.

Future Research Directions: This study provides a foundational proof of concept that opens avenues for future research. Building on these findings, future work may explore advanced manifold sampling techniques, such as Riemannian Hamiltonian VAEs to capture complex latent distributions more faithfully [22]. Additionally, as demonstrated by vEEGNet [23], integrating the RGP-VAE’s geometric constraints and subject-invariance with discriminative frameworks could potentially yield

latent spaces that are simultaneously geometrically valid, subject-invariant and class-discriminative. Finally, while relying solely on training metrics and geometric spread to determine parameters successfully prevented data leakage, future work would benefit from a systematic hyperparameter search.

CONCLUSION

This paper developed and validated a novel Riemannian Geometry-Preserving VAE (RGP-VAE) for generating synthetic EEG covariance matrices in the challenging cross-subject MI-BCI context. The RGP-VAE is not only capable of consistently generating valid SPD matrices—overcoming the limitations of standard VAEs—but also closely matches the original data diversity. The high-fidelity synthetic data can maintain or even significantly improve classification performance for specific classifiers. However, divergent classifier results highlight generative capabilities on the SPD manifold do not guarantee universal downstream improvements.

REFERENCES

- [1] Congedo M, Barachant A, Bhatia R. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*. 2017;4(3):155–174.
- [2] Yger F, Berar M, Lotte F. Riemannian approaches in brain-computer interfaces: A review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2017;25(10):1753–1762.
- [3] Chevallier S et al. The largest eeg-based bci reproducibility study for open science: The moabb benchmark. 2024. arXiv: 2404.15319 [eess.SP].
- [4] Blankertz B, Dornhege G, Krauledat M, Müller KR, Curio G. The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*. 2007;37(2):539–550.
- [5] Kalunga E, Chevallier S, Barthélemy Q. Data augmentation in Riemannian space for Brain-Computer Interfaces. In: *STAMILINS 2015 proceedings*. Lille, France, Jun. 2015.
- [6] Kingma DP, Welling M. Auto-encoding variational bayes. 2013. arXiv: 1312.6114 [stat.ML]. [Online]. Available: <https://arxiv.org/abs/1312.6114>.
- [7] Shao H, Kumar A, Fletcher PT. The riemannian geometry of deep generative models. *CoRR*. 2017;abs/1711.08014.
- [8] Yair O, Ben-Chen M, Talmon R. Parallel transport on the cone manifold of SPD matrices for domain adaptation. *IEEE Transactions on Signal Processing*. 2019;67(7):1797–1811.
- [9] Faller J, Vidaurre C, Solis-Escalante T, Neuper C, Scherer R. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2012;20(3):313–319.
- [10] Aristimunha B et al. *Mother of all BCI Benchmarks*. Version 1.0.0. 2023. [Online]. Available: <https://github.com/NeuroTechX/moabb>.
- [11] Schirrmeister RT et al. Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human EEG. *CoRR*. 2017;abs/1703.05051.
- [12] Zhu H, Forenzo D, He B. On the deep learning models for EEG-based brain-computer interface using motor imagery. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2022:1–1.
- [13] Chen Y, Wiesel A, Eldar YC, Hero AO. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*. 2010;58(10).
- [14] Miolane N, Holmes SP. Learning weighted submanifolds with variational autoencoders and riemannian variational autoencoders. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019:14491–14499.
- [15] Fletcher PT, Lu C, Pizer SM, Joshi S. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*. 2004;23(8):995–1005.
- [16] Moakher M. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*. 2005;26(3):735–747.
- [17] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*. 2015;abs/1502.03167.
- [18] Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: *Proceedings of the 30th International Conference on Machine Learning*. 2013.
- [19] Bowman SR, Vilnis L, Vinyals O, Dai A, Jozefowicz R, Bengio S. Generating sentences from a continuous space. In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics: Berlin, Germany, Aug. 2016, 10–21.
- [20] Loshchilov I, Hutter F. Fixing weight decay regularization in adam. *CoRR*. 2017;abs/1711.05101.
- [21] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*. 2018;3(29):861.
- [22] Chadebec C, Thibaud-Sutre E, Burgos N, Allasonnière S. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;45(3):2879–2896.
- [23] Zancanaro A, Cisotto G, Zoppis I, Manzoni SL. Veegnet: Learning latent representations to reconstruct eeg raw data via variational autoencoders. In: *Information and Communication Technologies for Ageing Well and e-Health*. Springer Nature Switzerland: Cham, 2024, 114–129.

A Feasibility Study of a Motor Imagery–Based Brain–Computer Interface for Post-Stroke Patients

V. Martínez-Cagigal^{1,2,3}, S. Moreno-Calderón^{1,3}, A. Cisnal⁴, S. Pérez-Velasco^{1,2}, A. Martín-Fernández^{1,2,3}, E. Santamaría-Vázquez^{1,2,3}, J. C. Fraile⁴, J. Pérez-Turiel⁴, R. Hornero^{1,2,3}

¹ Biomedical Engineering Group (GIB), University of Valladolid, Valladolid, Spain

² Network Biomedical Research Center for Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Valladolid, Spain

³ Valladolid Health Research Institute (IBioVALL), Valladolid, Spain

⁴ Institute of Advanced Production Technologies (ITAP), University of Valladolid, Valladolid, Spain

E-mail: victor.martinez.cagigal@uva.es

ABSTRACT: Stroke is a major contributor to chronic disability, and many survivors continue to experience pronounced upper-limb weakness despite standard rehabilitation. By converting motor intent into real-time sensory feedback, brain–computer interfaces (BCIs) could help re-engage the sensorimotor loop. However, findings across clinical studies remain mixed, especially for robot-assisted approaches. In this feasibility study, we investigated whether a noninvasive motor-imagery (MI) BCI providing concurrent multimodal feedback (visual avatar and robot-assisted hand motion) yields additional improvements beyond usual care. Sixteen patients with a first-ever stroke were assigned to conventional therapy alone (CG, $n = 6$) or to the same therapy supplemented with 23 one-hour BCI sessions (EG, $n = 10$). Online EEG from eight sensorimotor channels was used to classify left- versus right-hand kinesthetic MI, and motor, cognitive, and functional tests were evaluated before and after the intervention. The EG showed significant within-group gains in voluntary motor control and coordination of the affected hand, episodic memory and visuoconstructive ability, as well as functional measures such as perceived general health and independence. These preliminary findings suggest that BCI-based therapies may enhance post-stroke rehabilitation.

INTRODUCTION

Stroke remains one of the most pressing public health challenges worldwide, ranking as the second leading cause of death and a leading cause of long-term disability [1]. Its incidence and prevalence continue to rise, particularly in Europe, where stroke represents the primary cause of acquired disability. Clinical outcomes vary according to lesion characteristics, but a substantial proportion of survivors experience persistent motor and cognitive impairments that compromise activities of daily living. Upper-limb paresis is especially prevalent, affecting up to 80% of patients in the acute stage and persisting in a significant fraction beyond six months [2].

Although interventions such as constraint-induced movement therapy, conventional physiotherapy, mirror therapy, and robot-assisted rehabilitation are routinely implemented, their effectiveness remains limited [3]. In this context, BCIs have emerged as a promising therapeutic strategy. By decoding movement intention from neural activity, typically via electroencephalography (EEG), and coupling it with contingent sensory feedback, BCI systems establish a closed sensorimotor loop that could facilitate functional recovery [2].

Over the past decade, multiple randomized controlled trials (RCTs) have evaluated BCI-based interventions for post-stroke rehabilitation [4, 5]. Collectively, these studies report that BCI-assisted therapy yields greater improvements in upper-limb motor function, commonly assessed with the Fugl–Meyer Assessment–Upper Extremity (FMA-UE), than conventional therapy alone [4, 5]. Functional electrical stimulation (FES)-based BCIs appear to produce the largest effect sizes, while robotic-assisted BCIs demonstrate significant motor gains but do not consistently outperform robotic therapy delivered without neural control [4]. Treatment effects tend to be more pronounced in subacute patients than in chronic populations, underscoring the importance of early intervention. Although motor imagery (MI) remains the most widely implemented paradigm, action observation has also shown promising, and in some cases superior, outcomes. Meta-analytic evidence further suggests that optimal protocols typically involve sessions lasting 20–90 minutes, administered several times per week over approximately three to four weeks, and that improvements may reflect genuine neuroplastic changes rather than just compensatory mechanisms [4, 5].

Despite these encouraging findings, important methodological limitations persist. Meta-analyses consistently highlight substantial heterogeneity across studies in terms of patient characteristics, time since stroke onset, feedback modalities, signal processing approaches, number of sessions, and pre-post measures. Sample sizes are frequently small, limiting statistical power and precluding

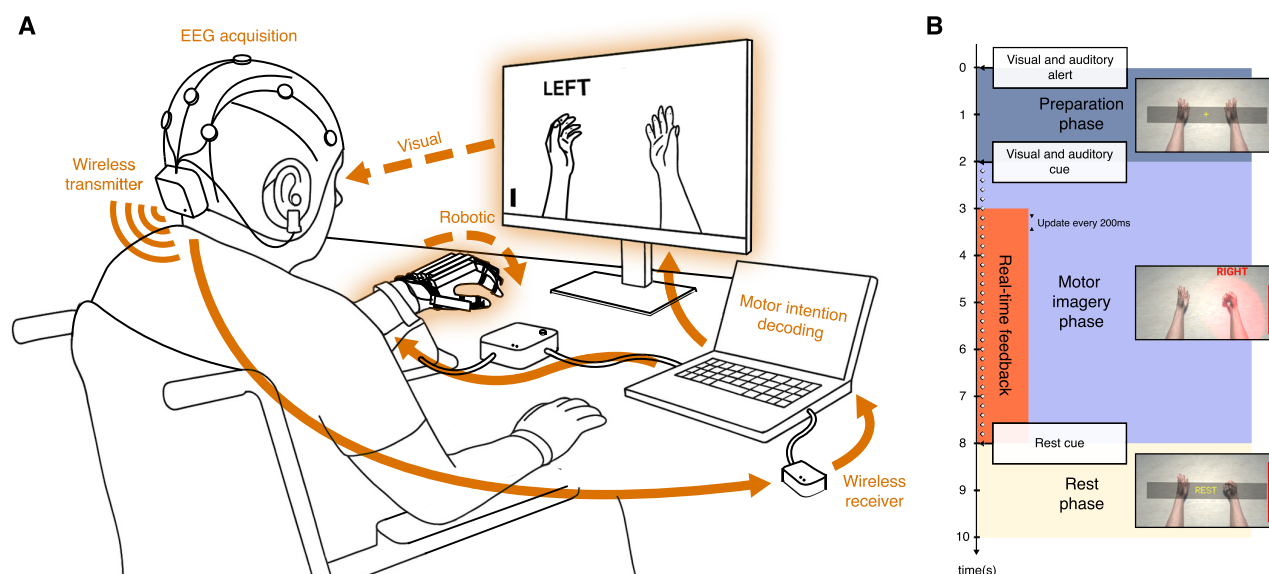


Figure 1: (A) General overview of the BCI system. EEG signals are acquired using portable hardware and transmitted wirelessly in real-time to a laptop, where the user's motor intentions are decoded. The system delivers multimodal feedback (dashed lines) comprising (1) visual feedback in the form of an avatar hand movement and (2) concurrent robotic assistance for flexion or extension movements applied to the paretic hand. (B) The structure of a MI trial, composed of (1) a preparation phase of 2 s, (2) a MI phase of 6 s (initiated by cue presentation and followed by continuous decoding of epochs every 200 ms), and (3) a rest phase of 2 s.

robust subgroup analyses [4, 5]. Thus, further studies are required to clarify the clinical value of specific BCI configurations and training paradigms. The present study addresses these gaps by evaluating a novel MI-based BCI system that delivers simultaneous multimodal feedback (embodied visual avatars with robotic-assisted movement) for stroke patients. By assessing motor, cognitive, and functional outcomes following a structured intervention protocol, this work seeks to provide more evidence regarding the therapeutic potential of MI-based BCI rehabilitation after stroke.

MATERIALS AND METHODS

Sixteen patients (mean age 66.63 ± 11.42 years; 8 females) with first-ever stroke and upper-limb hemiparesis were recruited from the Benito Menni Hospital Center (Valladolid, Spain) after providing written informed consent. Demographic data are shown in Table 1. The majority were in the early subacute phase (7 days to 3 months post-stroke, $n = 13$), with two in the late subacute phase and one in the chronic stage; both ischemic ($n = 11$) and hemorrhagic ($n = 5$) etiologies were represented, with heterogeneous lesion locations and variable laterality of impairment. The study was approved by the local ethics committee (reference: CASVE-NM-22-617) and conducted in accordance with the Declaration of Helsinki, with all data anonymized prior to analysis. Inclusion criteria comprised age ≥ 30 years, neuroimaging-confirmed first-ever stroke, and at least 7 days post-onset, whereas exclusion criteria included severe cognitive or behavioral disturbances, impaired consciousness, high seizure risk, major visual deficits, severe spasticity, in-

tense pain during mobilization, and other medical or dermatological conditions interfering with participation.

The overall system and trial structure are illustrated in Figure 1. A monitor positioned in front of the participant, who was seated at a table with both upper limbs supported, presented a cue-based MI paradigm, while the paretic hand was secured to a robotic exoskeleton mounted on an inclined support to ensure anatomical alignment and mechanical stability. During each trial, randomized visual cues instructed the participant to perform kinesthetic imagery of left- or right-hand movements during trials of 6 s. The system decoded participants' motor intentions in real-time and translated them into contingent multimodal feedback: the inferred movement was rendered by virtual avatar hands and, when the paretic hand was targeted, the robotic device delivered assisted wrist and finger flexion. This closed-loop setup therefore linked motor intention with congruent visual and proprioceptive feedback.

EEG data were acquired using a wireless system (g.Nautilus PRO, g.Tec, Austria) with active electrodes positioned over the sensorimotor cortex (FC5, FC6, C1, C2, C3, C4, CP5, and CP6; referenced to the earlobe and grounded at AFz). Signal acquisition, online processing, and feedback visualization were implemented in MEDUSA[®] v2023 (medusabci.com) [6]. The processing pipeline included band-pass filtering (8–30 Hz), epoching into 2 s windows, baseline normalization using the 2 s interval preceding each window, and spatial feature extraction via common spatial patterns (CSP). Log-variance features from four CSP projections were classified using a regularized linear discriminant analysis (rLDA) with automatic shrinkage, trained during calibration runs.

Table 1: Participants' demographic data

	ID	Age	Sex	Diagnosis	AH	Days post-stroke and phase	
EG	PE01	59	F	Hemorrhagic stroke involving the left midbrain and pons	R	42	Early subacute
	PE02	62	F	Acute ischemic stroke involving the right fronto-parieto-temporal regions	L	27	Early subacute
	PE03	60	F	Ischemic stroke involving the R-MCA territory	L	417	Chronic
	PE04	46	F	Hemorrhagic stroke involving the right basal ganglia	L	51	Early subacute
	PE05	69	M	Subacute ischemic stroke involving the R-MCA territory	L	34	Early subacute
	PE06	61	M	Hemorrhagic stroke involving the left thalamus and left midbrain	R	48	Early subacute
	PE07	49	M	Ischemic stroke involving the L-MCA, L-PCA, R-SCA, L-PICA, and R-PICA	R	20	Early subacute
	PE08	78	M	Ischemic cerebral infarction involving the right PCA territory	L	43	Early subacute
	PE09	84	F	Ischemic cerebral infarction involving the R-MCA territory	L	27	Early subacute
	PE10	79	F	Ischemic cerebral infarction involving the R-MCA territory	L	13	Early subacute
CG	Mean	64.70	6F,			72.20	
	STD	12.67	4M		7L, 3R	121.77	
	PC01	66	M	Ischemic stroke involving the right carotid artery territory	L	110	Late subacute
	PC02	63	F	Ischemic lesions in the right cerebellar hemisphere and basilar artery aneurysm	L	51	Early subacute
	PC03	68	M	Ischemic infarction involving the right hemipons in the PPA territory	L	110	Late subacute
	PC04	88	M	Intracerebral hemorrhage involving the left capsulolenticular region	R	73	Early subacute
	PC05	68	M	Intracerebral hemorrhage involving the left capsulolenticular region	R	83	Early subacute
	PC06	66	F	Acute bihemispheric ischemic cerebral infarctions (mainly left hemisphere)	R	38	Early subacute
	Mean	69.83	2F,			77.50	
	STD	9.09	4M		3L, 3R	29.75	

AH: affected hand, CG: control group, EG: experimental group, F: female, M: male, R: right, L: left, MCA: middle cerebral artery, PCA: posterior cerebral artery, SCA: superior cerebellar artery, PICA: posterior inferior cerebellar artery, PPA: paramedian pontine artery.

Participants were allocated to a control group (CG, $n = 6$) receiving conventional rehabilitation or to an experimental group (EG, $n = 10$) receiving additional BCI-based therapy.

Conventional rehabilitation: It followed the hospital's standard multidisciplinary stroke program and included individualized physiotherapy, occupational therapy, neuropsychological intervention, and speech/swallowing therapy when clinically indicated. Physiotherapy focused on motor recovery, balance, gait, tone management, and upper-limb mobilization, whereas occupational therapy targeted activities of daily living and upper-limb dexterity.

BCI-based therapy: The BCI intervention consisted of 23 one-hour sessions delivered five times per week. Initial sessions (3) provided visual feedback only for familiarization, whereas subsequent sessions (20) combined visual and robotic feedback. To reduce frustration and facilitate engagement during early exposure to the system, calibration runs were performed using positive feedback, whereby the virtual hand closed independently of classifier performance. This familiarization strategy was intended to help participants understand the task and progressively develop control over the MI paradigm before receiving fully contingent feedback. Each session included structured training and testing runs within a cue-based MI paradigm, with continuous real-time feedback provided during imagery periods.

In the first session, the calibration phase comprised 4 runs (40 trials total; 20 left-hand and 20 right-hand MI trials). In subsequent sessions, calibration consisted of 2 runs (20 trials total), which were combined with the 120 trials acquired during the previous session to retrain the decoding model. To mitigate EEG non-stationarities and preserve stable online performance, the CSP+rLDA pipeline was

recalibrated after 40 and 80 trials during each session. No additional cross-validation or hyperparameter tuning procedures were applied, since the objective of the study was to evaluate feasibility under real-time clinical conditions using a fixed decoding pipeline. During online operation, overlapping 2 s epochs were processed every 200 ms to produce probabilistic estimates of left- versus right-hand MI, which were used to drive continuous real-time feedback.

Motor, cognitive and functional assessments: Pre- and post-intervention assessments encompassed standardized measures of motor function, cognitive performance, and functional independence administered by certified clinicians. Motor function was assessed by FMA [7], Box and Block Test (BBT) [8] and Range of Motion (ROM) goniometry for wrist and fingers [9]. Cognitive tests comprised the Free and Cued Selective Reminding Test (FCSRT) [10] and the Rey-Osterrieth Complex Figure (ROCF) [10]. Functional status was assessed by the 36-Item Short Form Health Survey (SF-36) [11], as well as the Functional Independence Measure (FIM) [12].

RESULTS AND DISCUSSION

Table 2 summarizes the statistical outcomes of the motor, cognitive, and functional assessments at pre- and post-intervention in both the CG and EG. For cognitive measures, z-scores were computed using normative data adjusted for age, education, and sex from Spanish populations [10]. Within-group pre-post differences were assessed with one-sided Wilcoxon signed-rank tests, whereas between-group comparisons of change scores were performed using one-sided Mann-Whitney U tests. All p -values were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate proce-

Table 2: Statistical results for motor, cognitive and functional tests.

Type	Test	Measure	μ_{CG1}	μ_{CG2}	p_{CG}	μ_{EG1}	μ_{EG2}	p_{EG}	Δp
Motor	FMA	↑ Upper Extremity (score)	11.167	14.000	0.362	17.100	22.600	0.041	0.814
		↑ Wrist (score)	2.500	2.833	0.630	2.900	4.100	0.186	0.814
		↑ Hand (score)	3.667	4.833	0.362	4.800	6.900	0.041	0.814
		↑ Coordination / Speed (score)	1.167	1.667	0.372	1.700	2.200	0.125	0.814
		↑ Global (score)	20.167	29.167	0.302	25.000	33.600	0.032	0.814
	BBT	↑ Affected Hand (no. blocks)	11.167	13.333	0.372	6.700	11.900	0.053	0.814
		↑ Non affected Hand (no. blocks)	29.833	34.333	0.302	29.100	32.500	0.047	0.814
	ROM	↑ Palmar Flexion (°)	17.500	27.500	0.362	21.000	32.500	0.041	0.814
		↑ Dorsal Flexion (°)	7.500	12.500	0.362	10.500	15.500	0.125	0.814
		↑ Radial Deviation (°)	5.833	8.333	0.372	6.500	8.500	0.186	0.814
		↑ Ulnar Deviation (°)	4.167	8.333	0.372	8.500	11.000	0.186	0.814
		↑ Index MCP Flexion (°)	23.333	45.833	0.362	40.500	53.500	0.082	0.816
		↑ Index MCP Extension (°)	10.833	15.833	0.372	12.000	15.000	0.186	0.814
		↑ Thumb Abduction (°)	25.833	32.500	0.372	24.000	32.000	0.082	0.814
		↑ Thumb Adduction (°)	0.000	5.000	0.630	3.000	6.500	0.617	0.814
		↑ Thumb MCP Flexion (°)	10.000	19.167	0.362	23.500	32.000	0.186	0.814
		↓ Thumb Opposition (cm)	0.667	0.917	1.000	1.700	1.700	0.698	0.814
Cognitive	FCSRT	↑ Immediate Free Recall 1 (z-score)	-0.492	-0.417	0.590	-0.816	0.697	0.032	0.461
		↑ Immediate Free Recall Total (z-score)	-0.825	-0.075	0.302	-0.190	0.597	0.053	0.814
		↑ Total Recall (z-score)	-0.760	0.168	0.372	-0.438	0.936	0.053	0.814
		↑ Delayed Free Recall (z-score)	-0.917	-0.175	0.372	-0.087	1.412	0.032	0.814
		↑ Delayed Total Recall (z-score)	0.007	0.760	0.372	-0.044	1.520	0.032	0.814
	ROCF	↑ Copy (z-score)	-0.994	-0.600	0.453	-1.609	-0.200	0.032	0.814
		↓ Time (z-score)	-0.594	-1.134	0.302	-0.955	-0.676	0.769	0.947
Functional	SF-36	↑ Physical Function (score)	1.667	10.833	0.302	22.000	30.000	0.053	0.814
		↑ Role (score)	0.000	0.000	1.000	12.500	27.500	0.186	0.814
		↑ Role Emotional (score)	16.667	27.667	0.372	23.300	43.300	0.330	0.814
		↑ Vitality (score)	35.000	50.833	0.302	54.000	60.000	0.281	0.923
		↑ Mental Health (score)	47.333	56.000	0.302	66.800	68.800	0.290	0.837
		↑ Social Function (score)	31.500	46.000	0.302	53.850	72.700	0.076	0.814
		↑ Bodily Pain (score)	70.000	80.333	0.372	74.700	89.400	0.125	0.814
		↑ General Health (score)	50.000	54.167	0.362	46.000	57.500	0.041	0.814
	FIM	↑ Total (score)	62.500	80.333	0.302	82.900	96.700	0.032	0.881

MCP: metacarpophalangeal; FMA: Fugl–Meyer Assessment; BBT: Box and Block Test; ROM: Range of Motion goniometry; FCSRT: Free and Cued Selective Reminding Test; ROCF: Rey–Osterrieth Complex Figure Copy Test; SF-36: 36-Item Short Form Health Survey; FIM: Functional Independence Measure. Columns μ_{CG1} and μ_{EG1} denote pre-assessment means, and μ_{CG2} and μ_{EG2} post-assessment means for control (CG) and experimental (EG) groups, respectively. p_{CG} and p_{EG} are within-group pre–post comparison p -values, and Δp is the p -value for between-group comparison of change scores ($CG_2 - CG_1$) vs. ($EG_2 - EG_1$). Arrows indicate the expected direction of improvement (↑ increase, ↓ decrease). P -values were adjusted column-wise using the Benjamini–Hochberg false discovery rate procedure. One-sided Wilcoxon signed-rank tests were used for within-group comparisons, and a one-sided Mann–Whitney U test for the between-group comparison.

ture. The hypothesized direction of improvement for each metric is also indicated.

Overall, the analyses indicated significant pre–post gains in the EG across several subscales, whereas the CG did not exhibit significant changes. Because both groups received conventional rehabilitation and only the EG completed the additional BCI-based intervention, the observed within-group improvements in the EG may suggest a potential additive effect of the proposed BCI system beyond usual care. However, the absence of statistically significant between-group differences, together with the unequal and limited sample sizes, precludes drawing firm conclusions regarding superiority of the BCI intervention. In addition, several measures also showed numerical improvements in the CG, suggesting that spontaneous recovery, conventional rehabilitation, or limited statistical power may also contribute to the observed effects.

Motor effects: Regarding motor improvements, EG

demonstrated significant motor improvements across global FMA ($\Delta = 8.60$, $p = 0.032$) and FMA-UE ($\Delta = 5.50$, $p = 0.041$), indicating enhanced voluntary motor control and coordination; as well as improvements in the FMA hand subscale ($\Delta = 2.10$, $p = 0.041$), suggesting better distal dexterity and grasp–release capacity. A significant increase in palmar flexion ($\Delta = 11.50^\circ$, $p = 0.041$) further supports functional hand positioning. Interestingly, BBT performance improved significantly for the unaffected limb, with a positive trend for the affected side. This pattern may reflect task familiarization effects, compensatory increases in use of the unaffected limb during daily activities, and/or limited statistical power. The magnitude and direction of the FMA-UE effect were consistent with prior meta-analytic findings on BCI-based rehabilitation, including robotic-based [4, 5] and FES-based paradigms [13]. Notably, the observed standardized effect size was comparable to those reported for FES-based BCIs, despite the use of combined robotic

and visual feedback in the present protocol [13]. Collectively, these findings could point to motor benefits associated with the proposed BCI-based intervention.

Cognitive effects: Significant neuropsychological improvements were also observed in the EG, particularly in verbal episodic memory and visuoconstructive abilities. On the FCSRT, participants demonstrated enhanced immediate learning and uncued retrieval ($\Delta = 1.51$, $p = 0.032$), greater delayed spontaneous retrieval ($\Delta = 1.50$, $p = 0.032$), and higher delayed total recall ($\Delta = 1.56$, $p = 0.032$); indicating more efficient encoding, retention, and retrieval processes. Performance on the ROCF copy condition ($\Delta = 1.41$, $p = 0.032$) also improved significantly, suggesting better visuospatial organization and planning strategies. Comprehensive cognitive assessment is rarely incorporated into RCTs of MI-based BCI interventions in stroke, and such outcomes are largely absent from existing meta-analyses, limiting direct comparisons. Although analogous MI-based neurofeedback studies in elderly populations have documented cognitive gains across multiple domains [14, 15], these effects have sometimes been observed in both experimental and control conditions [15], raising the possibility of practice effects or non-specific participant engagement. In contrast, the confinement of cognitive improvements to the EG in the present cohort suggests a potential intervention-specific mechanism, while emphasizing the need for larger, adequately powered trials to establish the robustness and specificity of these findings.

Functional Effects: Finally, the EG demonstrated significant gains in functional outcomes, including improved health-related quality of life and greater independence in activities of daily living. Specifically, SF-36 general health scores increased significantly ($\Delta = 11.50$, $p = 0.041$), reflecting enhanced self-perceived well-being, although comparable data are largely absent from prior BCI-based stroke rehabilitation studies. FIM scores also improved significantly ($\Delta = 13.80$, $p = 0.032$), indicating reduced assistance requirements in daily functioning. The magnitude of FIM change was consistent with previous MI-based BCI interventions incorporating robotic feedback [16], while existing meta-analytic evidence reports considerable heterogeneity in functional outcome measures [4], which may obscure pooled effects. Overall, the observed improvements support enhancement in everyday functional autonomy.

Strengths and limitations: Recent commercial rehabilitation systems such as recoveriX [17] demonstrated the clinical potential of MI-based BCI paradigms combined with FES for promoting active muscular activation after stroke. In comparison, the present approach employed multimodal robotic and visual feedback, which may provide a more kinematically congruent and anatomically controlled sensorimotor experience during hand movements. While FES directly activates residual musculature, robotic assistance may facilitate more reproducible joint trajectories and movement amplitudes, particularly in patients with severe paresis. Both approaches

should therefore not be considered mutually exclusive. Future work could investigate hybrid paradigms integrating robotic assistance and FES in order to combine active muscular recruitment with synchronized proprioceptive and visual feedback.

The findings suggest that the proposed BCI-based therapy incorporating multimodal feedback may facilitate motor recovery and confer additional cognitive and functional benefits beyond those achieved with conventional rehabilitation in stroke patients. However, we acknowledge that the study's modest sample size ($n = 16$) limited statistical power and precluded subgroup analyses based on clinically relevant variables such as affected side or hand. Furthermore, the absence of longitudinal follow-up prevents conclusions regarding the durability of observed gains. Future research should investigate the durability of these improvements. Additional efforts may focus on the implementation of more advanced signal processing and adaptive decoding techniques, alongside strategies to enhance user engagement and perceived sense of agency through augmented reality integration and, where appropriate, the incorporation of gamification elements to promote sustained motivation and participation. In addition, recovery stage was heterogeneous across participants. Notably, the two late-subacute participants were allocated to the CG, whereas the EG included one chronic participant. Since spontaneous recovery dynamics differ substantially across recovery phases, this heterogeneity may have influenced group-level outcomes.

Another important limitation is that the experimental group received additional therapy exposure in the form of 23 BCI sessions, whereas the control group only received conventional rehabilitation. Since motor imagery itself may promote motor recovery after stroke, part of the observed improvements may reflect non-specific training effects independent of neural decoding accuracy. At the same time, it would not be ethically appropriate to withhold conventional rehabilitation from the EG in order to isolate the effect of the BCI intervention, as conventional rehabilitation constitutes the current standard of care for stroke rehabilitation. Future studies should therefore incorporate intensity-matched control conditions, such as sham-feedback or robotic-only interventions added to conventional therapy, in order to better isolate the specific therapeutic contribution of the BCI feedback while maintaining equivalent rehabilitation exposure across groups.

CONCLUSION

This study investigated the effects of a noninvasive MI-based BCI delivering combined visual avatar and robotic-assisted feedback in post-stroke rehabilitation. Over 23 one-hour sessions, participants receiving the BCI intervention (EG, $n = 10$) in addition to standard multidisciplinary therapy exhibited significant pre-post improvements in upper-limb motor performance, selected cognitive domains, and functional independence, whereas no significant changes were detected in the

participants receiving conventional therapy alone (CG, $n = 6$). Motor gains encompassed clinically relevant Fugl-Meyer subscales and palmar range of motion, while cognitive improvements were observed in verbal episodic memory and visuoconstructive performance. Functional outcomes also improved, as reflected by enhanced self-perceived health status and greater autonomy in daily activities. Although between-group comparisons did not reach statistical significance, potentially owing to limited sample size and unequal therapy exposure between groups, the consistent pattern of within-group improvements in the EG cohort suggests a potential additive therapeutic effect associated with intention-contingent multimodal feedback and sensorimotor closed-loop. Overall, these preliminary findings suggest that MI-based BCI interventions may promote not only motor recovery but also cognitive and functional gains, particularly in subacute stroke populations.

ACKNOWLEDGMENTS

This research was mainly supported by the project RTC2019-007350-1 funded by Ministry of Science and Innovation of Spain, and partially by project 0124_EUROAGE_MAS_4_E, funded by EU-ERDF through the Interreg VI-A Spain-Portugal Program (POCTEP) 2021-2027; and VA140P24, by the Regional Government of Castilla y León and EU-ERDF. This work was supported by ‘Centro de Investigación Biomédica en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)’ through ‘Instituto de Salud Carlos III’ co-funded with EU-ERDF funds. A. Martín-Fernández was in receipt of a PIF grant by the ‘University of Valladolid’. We would also like to acknowledge the Benito Menni Hospital Center in Valladolid (Spain) for its assistance in recruiting participants and for its support in the implementation of the evaluation protocol.

REFERENCES

- [1] Feigin VL et al. World Stroke Organization: Global Stroke Fact Sheet 2025. *International Journal of Stroke*. 2025;20(2):132–144.
- [2] Angerhöfer C, Colucci A, Vermehren M, Hömberg V, Soekadar SR. Post-stroke Rehabilitation of Severe Upper Limb Paresis in Germany – Toward Long-Term Treatment With Brain-Computer Interfaces. *Frontiers in Neurology*. 2021;12:772199.
- [3] Langhorne P, Coupar F, Pollock A. Motor recovery after stroke: A systematic review. *The Lancet Neurology*. 2009;8(8):741–754.
- [4] Li D et al. Effects of brain-computer interface based training on post-stroke upper-limb rehabilitation: A meta-analysis. *Journal of NeuroEngineering and Rehabilitation*. 2025;22(1):44.
- [5] Cervera MA et al. Brain-computer interfaces for post-stroke motor rehabilitation: A meta-analysis. *Annals of Clinical and Translational Neurology*. 2018;5(5):651–663.
- [6] Santamaría-Vázquez E et al. MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. *Computer Methods and Programs in Biomedicine*. 2023;230:107357.
- [7] Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. A method for evaluation of physical performance. *Scand J Rehabil Med*. 1975;7(1):13–31.
- [8] Mathiowetz V, Volland G, Kashman N, Weber K. Adult norms for the Box and Block Test of manual dexterity. *The American journal of occupational therapy*. 1985;39(6):386–391.
- [9] Srinayanti Y, Widiyanti W, Andriani D, Firdaus FA, Setiawan H. Range of motion exercise to improve muscle strength among stroke patients: A literature review. *International Journal of Nursing and Health Services (IJNHS)*. 2021;4(3):332–343.
- [10] Palomo R et al. Spanish normative studies in young adults (NEURONORMA young adults project): Norms for the Rey–Osterrieth Complex Figure (copy and memory) and Free and Cued Selective Reminding Test. *Neurología (English Edition)*. 2013;28(4):226–235.
- [11] Ware Jr JE. SF-36 health survey update. *Spine*. 2000;25(24):3130–3139.
- [12] Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: A quantitative review. *Archives of physical medicine and rehabilitation*. 1996;77(12):1226–1232.
- [13] Ren C et al. The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: A systematic review and meta-analysis. *Frontiers in Human Neuroscience*. 2024;18:1438095.
- [14] Gomez-Pilar J, Corralejo R, Nicolas-Alonso LF, Álvarez D, Hornero R. Neurofeedback training with a motor imagery-based BCI: Neurocognitive improvements and EEG changes in the elderly. *Medical & Biological Engineering & Computing*. 2016;54(11):1655–1666.
- [15] Marcos-Martínez D et al. Motor imagery-based neurofeedback in older adults: Neural signatures and feasibility in a randomized controlled trial targeting age-related cognitive decline. *Journal of NeuroEngineering and Rehabilitation*. 2026;23(1):107.
- [16] Zanona AdF et al. Brain-computer interface combined with mental practice and occupational therapy enhances upper limb motor recovery, activities of daily living, and participation in subacute stroke. *Frontiers in neurology*. 2023;13:1041978.
- [17] Irimia D et al. recoveriX: A new BCI-based technology for persons with stroke. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE: Orlando, FL, USA, Aug. 2016, 1504–1507.

Noninvasive self-paced BCI for lower limb exoskeleton control: a multi-session investigation of user learning

Alessio Lo Faro¹, Tommaso Cortecchia², Miroljub Mihalovic¹, Edoardo Trombin¹, Emanuele Menegatti^{1,3}, Luca Tonin^{1,3}, Stefano Tortora^{1,3}

¹Department of Information Engineering University of Padova, Padova, Italy

²Politecnico di Torino, Torino, Italy

³Padova Neuroscience Center, Padova, Italy

E-mail: stefano.tortora@unipd.it

ABSTRACT: Brain-Computer Interfaces (BCIs) represents a promising approach to incorporate the user's motion intention into the control of assistive technologies. However, its efficacy for the control of real lower-limb exoskeletons (LLE) remain largely overlooked, with most of the work in the literature focusing on single-session studies. To overcome these limitations, this work proposes and investigates the usage of a noninvasive electroencephalography (EEG)-based BCI for self-paced control of a LLE across multiple sessions. The proposed framework and training protocol are explicitly designed to promote progressive familiarization through repetitive training, continuous feedback, and minimal decoder calibration, encouraging the emergence of more stable and physiologically interpretable neural patterns. Experimental results show that all participants achieved functional LLE control within three sessions, paired with classification improvements and stabilization of cortical activation patterns. Overall, the findings support the feasibility of practical LLE-BCI systems while highlighting the importance of stable decoding strategies in online and multi-session experiments promoting user learning, which are key to achieving a more reliable and effective BCI-driven LLE control.

INTRODUCTION

In recent decades, the burden of motor disability resulting from neurological conditions has increased, mainly due to population aging and improved survival [1], resulting in an unmet need for new rehabilitation and assistive strategies. Lower-limb exoskeletons (LLEs) represent a promising assistive technology to compensate for motor impairments, enabling standing and walking capabilities that are otherwise impossible with conventional devices such as wheelchairs. Moreover, exoskeletons are increasingly considered as rehabilitation tools, as they allow intensive, repetitive, and task-oriented training with limited external assistance compared to traditional methods [2]. Nevertheless, current LLE are still characterized by unnatural control interfaces to translate user intentions into robot commands; in this context, Brain-Computer

Interfaces (BCIs) enable a direct communication pathway between the brain and the external actuator. Therefore, the coupling of BCIs and exoskeletons has emerged as a promising approach both for motion substitution [3] and motor rehabilitation [4]. In particular, BCIs capturing the user's motion intention (MI) represent a natural solution for the control of limb prosthetics and exoskeletons without the need of overt movements or muscular activations [5]. While MI-based BCIs are well established for upper-limb exoskeleton control [6], the decoding of gait or feet MI presents intrinsic challenges: the cortical representation of the feet is deeper and more medial, resulting in weaker and less spatially distinct EEG features compared to upper-limb MI. These limitations have contributed to marked inter-subject variability, and a limited application for the control of real exoskeleton devices.

Among the challenges preventing the effective usage of BCI, non-stationarity of EEG signals and the high inter-session variability lead to requiring repeated calibrations of the BCI decoder [7]. To overcome these limitations, a new perspective, that has become recently popular, describes the BCI system not as a static decoder, but as a dynamic system composed of a learning user and an adaptive decoding algorithm [8]. Within this framework, performance gains may stem not only from more complex classification methods, but also from the user's increasing ability to produce stable and well-separated neural patterns through interaction with the system [9]. This view motivates training strategies that emphasize user learning and limit frequent decoder recalibration [10].

Within this context, the present work proposes a noninvasive EEG-based BCI system for the self-paced control

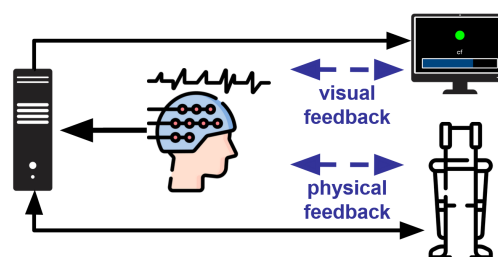


Figure 1: Experimental setup

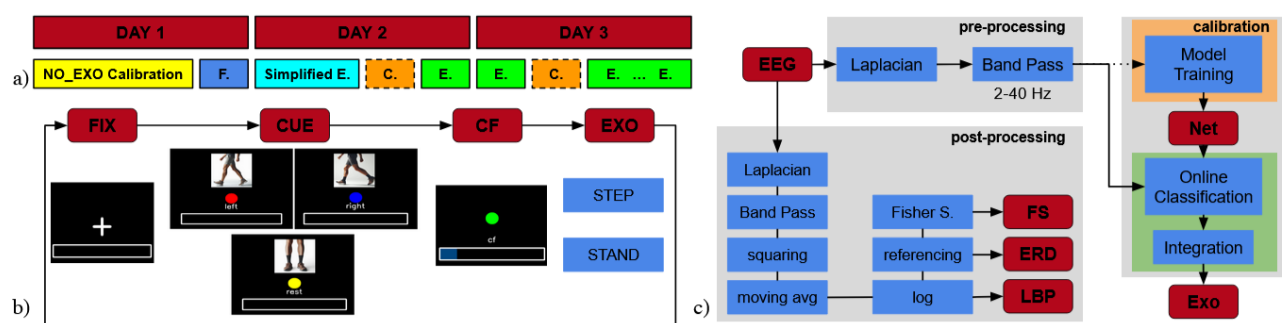


Figure 2: a) Sessions design: F: familiarization, C: calibration, E: evaluation. b) Experimental trial organization with the corresponding visual instructions. c) Data processing and classification pipelines.

of a LLE. Unlike existing studies, the proposed method specifically target the user learning, allowing the subjects to progressively familiarize with the BCI-LLE system, and improve their performance through closed-loop interactions with real-time neurofeedback. Analysis of multi-session performance and EEG features reveals that methodologies promoting user learning are critical for developing the stable, discriminant neural signatures necessary for effective LLE-BCI operation.

MATERIALS AND METHODS

Experimental setup: The experimental setup (shown in Fig. 1) consists of a 64-channel EEG acquisition system (Waveguard cap with eego64 amplifier, AntNeuro, Netherlands) connected to a master PC running the software for data acquisition, processing, and classification. A monitor guided the experimental protocol and provided live feedback to the participant on the outcome of the BCI decoder. Physical feedback was provided by the open-source LLE ALICE (Interactive Technologies Studio INDI, France). During the experiments, all participants used hand crutches to maintain a correct and stable posture. The integration of all the different hardware and software components required for the experiment was achieved through ROS-Neuro [11], an extension of the Robot Operating System (ROS) for BCI applications.

Participants: Four healthy participants (3 male and 1 female) were recruited, with an age in the range of 23 to 25 years old. No participants had previous experience with BCI systems or LLE-assisted gait exercises. To identify them, an anonymous alphanumeric code was assigned to each participant: K9, L1, L2, L3.

Protocol and session design:

The protocol consists of a multi-session experiment organized in three different days, as shown in Fig. 2a. Based on the session, the protocol is organized in five different phases for user familiarization, decoder calibration, and online evaluation. During familiarization, participants interacted with the LLE without EEG acquisition, and an external operator triggered the command for the exoskeleton. The calibration phases involved EEG acquisition during visually guided MI of left or right stepping (Movement class) or relaxation (Rest class), orga-

nized in runs of 16 trials. The organization of each trial is shown in Fig. 2b, and includes a 2-second fixation period, a 2-second task cue, and a continuous feedback phase (4 s for Movement classes, 8 s for Rest), during which participants performed the instructed mental task. Only in the first day, calibration runs are performed without the exoskeleton and with the subject straight standing in front of the monitor (No-Exoskeleton Calibration). These data are used to calibrate the first subject-specific decoder. In the online evaluation phases, the classifier outputs are processed in real time and used to control the visual feedback. The evaluation runs consist of 16 trials where movement is discriminated from rest, with the same organization as in Fig. 2b. Due to the difficulty in reliably distinguishing left from right feet imagery [12], movement classes were merged during classification, also motivated by the natural alternation of left-right steps in gait. At trial completion, the exoskeleton executed the cued movement to provide physical feedback. A simplified evaluation protocol was also introduced in early online testing and bias inspection, providing exclusively positive feedback by responding only to correct classifier outputs. Following the first calibration, we maintained the same decoding model for each subject also in the following days whenever possible. New calibration runs with the subjects wearing the exoskeleton and a decoder retraining (either a complete recalibration or a fine tuning of the existing model) were performed if the recognition rate of a class dropped below 55% in two consecutive runs. Each session lasted less than two hours, including setup preparation, user instruction, and experimental execution, to minimize fatigue and habituation. At least 8 evaluation runs per subject were acquired to ensure a sufficiently large evaluation dataset.

Data analysis: Two independent pipelines were ideated as shown in Fig. 2c: one to prepare the dataset for the online decoders, and a second pipeline for the analysis of the neural features and user learning. EEG signals are sampled at 512 Hz and processed using a 1 s sliding window shifted every 62.5 ms (frame rate of 16 Hz). Thirty-two central electrodes around Cz were selected from the 64-channel montage according to the international 10–10 system (frontal to parietal sites); AFz and CPz were used as ground and reference, respectively.

For the online experiments the signals were firstly filtered by a spatial Laplacian operator. Subsequently a 5th-order Butterworth bandpass filter with cut-off frequencies set at 2 and 40 Hz was employed. Offline, the EEG signals were transformed into a set of spatial-spectral features. Log-band power (LBP) was computed in the $\delta - \theta$ (1–8 Hz), μ (8–13 Hz), and β (13–32 Hz) bands using band-pass filtering, signal squaring, moving-averaging, and logarithmic scaling. Temporal modulations were quantified via event-related desynchronization/synchronization (ERD/ERS) utilizing the signal recorded during the fixation period as the reference. Feature discriminability between rest and movement tasks was assessed using the Fisher’s Score (FS), yielding discriminant features maps across electrodes and frequency bands.

Classification: For each subject, three deep learning models were selected and trained individually on the pre-processed calibration data: EEGNet [13]; PhiNet [14], and a Temporal Graph Convolutional Network [15]. For online evaluation, the model yielding the best validation performance on the calibration data was selected. To improve the robustness of the system against single-sample classification noise, classifier outputs $C_i(t)$ were integrated over time at 16 Hz as follow: $v_i(t) = v_i(t - 1) + (C_i(t) - 0.5)/ev$. The integrated values v_i are clipped to the $[0, 1]$ interval, therefore in our two-class configuration, v_{Rest} and $v_{Movement}$ remain symmetric around 0.5. For controlling the exoskeleton, the overall decision was treated as a one class problem: rest state is maintained until $v_{Movement}$ (represented by the bar on the visual feedback) reached a threshold th , triggering therefore a step command and ending the trial. Dynamics were governed by tuning the confidence threshold th and the ev evidence-velocity accumulation parameter, allowing control over system’s sensitivity.

Evaluation metrics: System performance was evaluated using both standard classification metrics and a task-specific temporal measure. Classification accuracy and F1 score were computed to assess single-sample and trial-level discrimination performance. In addition, a normalized time score [9] was used to quantify the trade-off between system reactivity to MI and robustness against false activations during rest.

$$normTimeScore = \frac{(timeout - \hat{T}_{Rest}) * n_{rest} + \hat{T}_{Mov} * n_{Mov}}{timeout * 2 * n_{tot}} \quad (1)$$

$\hat{T}_i$ is the mean time required to complete the task of the i^{th} class and n_i is the total number of task performed; therefore, lower values indicate a faster and more reliable system. The *timeout* is set at 10 seconds as it is the maximal duration of the continuous feedback for the rest trials.

RESULTS

Feature analysis: For all the participants, a total of 10 calibration runs per subject were acquired in different days and used to analyze the evolution of the discriminant features over time in the $\delta - \theta$, μ and β bands. The

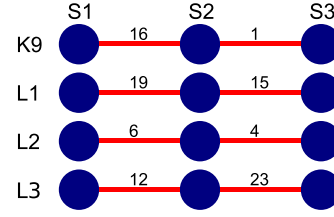


Figure 3: interval (in days) between two consecutive sessions.

interval in days between two consecutive sessions is reported in Fig. 3, while Fig. 4 reports the FS maps for each calibration run calculated on the β band. To assess the stability of the generated neural patterns over time, Fig. 4 shows the euclidean distances between the FS distributions of consecutive runs, measuring the similarity between the discriminant features used for MI classification [10]. This analysis was performed for each frequency bands. To enable inter-subject comparison, vectors were normalized by the mean norm of subject-specific FS vectors. Interestingly, all the subjects show a statistically significant decrease of the between-run distance throughout the experiment, particularly in the μ and β bands, highlighting an increase in the temporal consistency of the neural features in these bands.

Classification models: For each participant, the adopted policy aimed to maintain the same classifier for as long as possible. To analyze the decoding stability for the LLE control, Fig. 5 illustrates, for each session and participant, whether it was necessary to train a new classifier, perform a fine-tuning of the existing model with the new data, or whether the classifier was preserved as it was. Additionally, the preferred model for each subject is also reported. All the subjects required a full model recalibration at the beginning of the second session, while the previous model was preserved in the third session with a model fine-tuning for subjects L1, L2 and L3.

Online evaluation results: Fig. 6 reports the evolution of the performance metrics over time, showing an overall increase of performance with training for all the subjects. To analyze the adjustments of the control dynamic, Tab 1 reports the values of the hyperparameter ev and threshold th used during each evaluation run. Finally, to provide a qualitative perspective of the evolution of the system performance, Fig. 7 reports the single-trial results obtained during the first and last runs of the two evaluation sessions. Green lines correspond to successful trials, while red lines to failures.

DISCUSSION

Features analysis: The evolution of the FS spatial distributions suggests a progressive consolidation of discriminant neural patterns across sessions. Despite inter-subject variability, all participants exhibit increasing centralization of discriminant features in at least one frequency band, mainly along midline sensorimotor regions. This behavior supports the hypothesis that repeated inter-

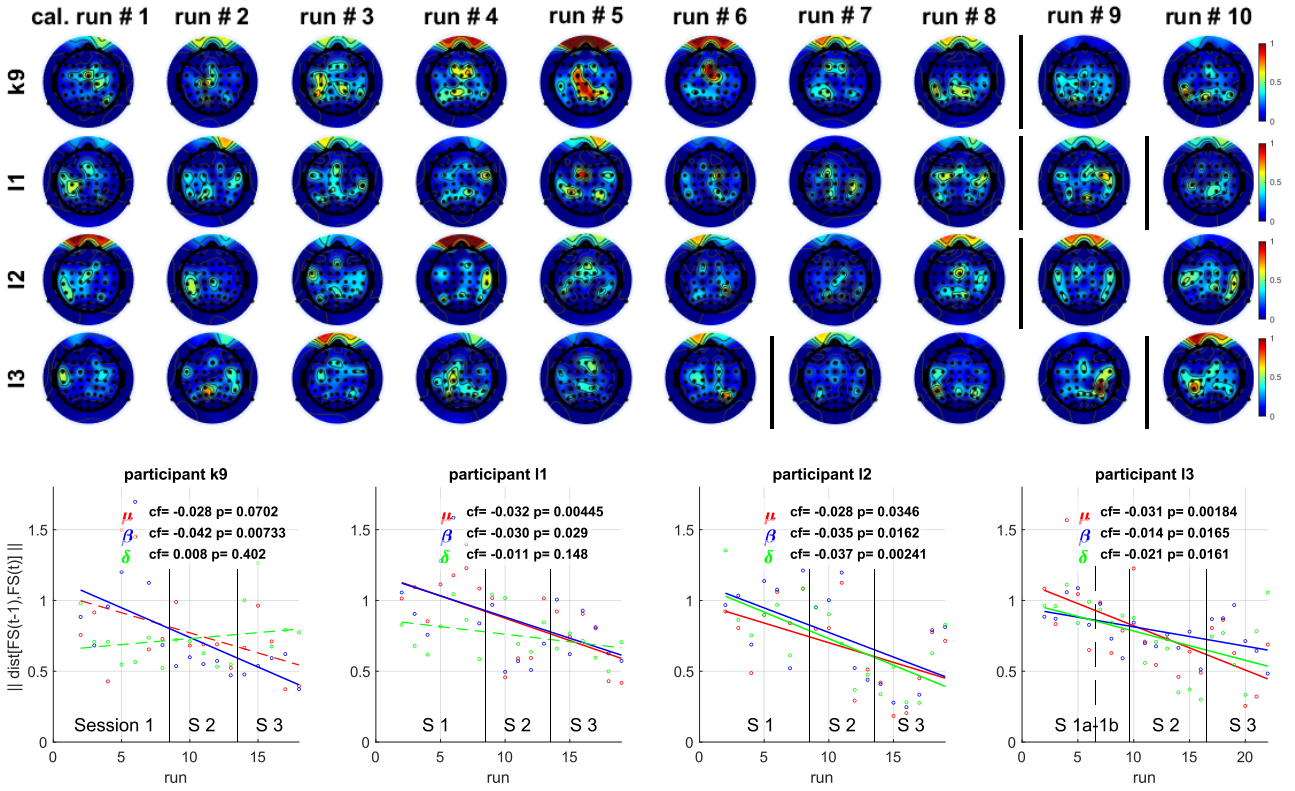


Figure 4: (Top) Features maps computed in the β frequency band for each calibration run; vertical lines correspond to the start of a new session. (Bottom) Normalized between-run FS distances across runs for each participant and frequency band. Correlation calculated using Spearman correlation test (full line when significant, i.e., $p < 0.05$ two-tailed t-test, dashed line otherwise).

participant	S2 start	S2 end	S3 start	S3 end
K9	TGCN	TGCN	TGCN	TGCN
L1	EEGNet	EEGNet	EEGNet	EEGNet
L2	PhiNet	PhiNet	PhiNet	PhiNet
L3	PhiNet	PhiNet	PhiNet	PhiNet

complete recalibration
 fine tuning
 no recalibration

Figure 5: Classifier updates before or during each evaluation sessions.

Table 1: Hyperparameter values during online evaluation runs

	K9		L1		L2		L3	
	ev	th	ev	th	ev	th	ev	th
run 1 S2	56	0.8	48	0.7	48	0.7	56	0.7
run 2 S2	48	0.8	48	0.8	48	0.7	48	0.6
run 3 S2	48	0.8	48	0.8	56	0.8	48	0.7
run 4 S2	48	0.8	48	0.8	—	—	48	0.65
run 5 S2	—	—	—	—	—	—	48	0.65
run 6 S2	—	—	—	—	—	—	48	0.65
run 1 S3	48	0.8	56	0.8	48	0.75	48	0.8
run 2 S3	48	0.8	56	0.8	48	0.8	48	0.7
run 3 S3	48	0.8	48	0.8	48	0.85	48	0.75
run 4 S3	48	0.8	48	0.8	56	0.8	56	0.8
run 5 S3	—	—	48	0.8	56	0.85	56	0.7
run 6 S3	—	—	—	—	56	0.8	—	—

action with the BCI promotes user-driven neural adaptation. Among the analyzed bands, the β rhythm shows the most consistent involvement across participants, combining high discriminative power with a gradual alignment of spatial hot spots toward centrally located electrodes. Although complete spatial stabilization is not achieved, the concurrent steady decrease in run-to-run variability of FS vectors indicates increasing stability of cortical activation patterns. The $\delta - \theta$ band exhibits weaker and

less consistent trends, which may be related to its involvement in movement-related neural processes, as suggested by studies focusing on locomotion decoding [16]. On the other hand, the major effect is visible in the μ and β bands. Since for the subjects the novelty lies in the mental exercise, and not in the performance of a complete gait cycle, it would therefore seem reasonable that the most evident improvements occur in the μ and β sensory motor rhythms (SMR), suggesting an emergence of more stable and discriminant movement-related features consistent with existing literature on motor learning and cortical plasticity [17].

Online evaluation results: When performance metrics are aggregated across subjects, a general tendency toward improved classification accuracy over time emerges, although the magnitude and statistical robustness of these effects vary across individuals. Most participants exhibit significant increases in trial- or sample-level accuracy, while one subject shows no clear performance trend, confirming the pronounced inter-subject variability typical of MI-based BCI systems [8, 10]. Across sessions, the normalized time score decreases, particularly for participants showing trial accuracy improvements. This concurrent evolution suggests that performance gains are not merely the result of integrator tuning, but reflect a genuine performance improvement. From an individual perspective, participant K9 demonstrates the most consistent learning behavior, with significant improvements in

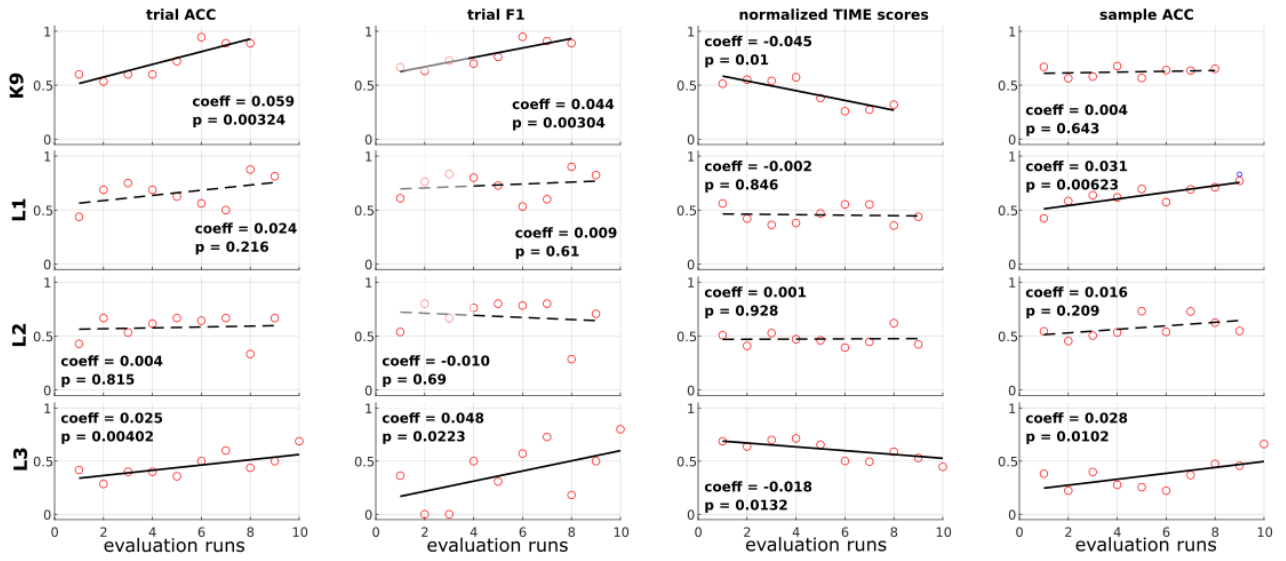


Figure 6: Performance metrics across the evaluation runs for each participant. Correlation calculated using Spearman correlation test (full line when significant, i.e., $p < 0.05$ two-tiled t-test, dashed line otherwise).

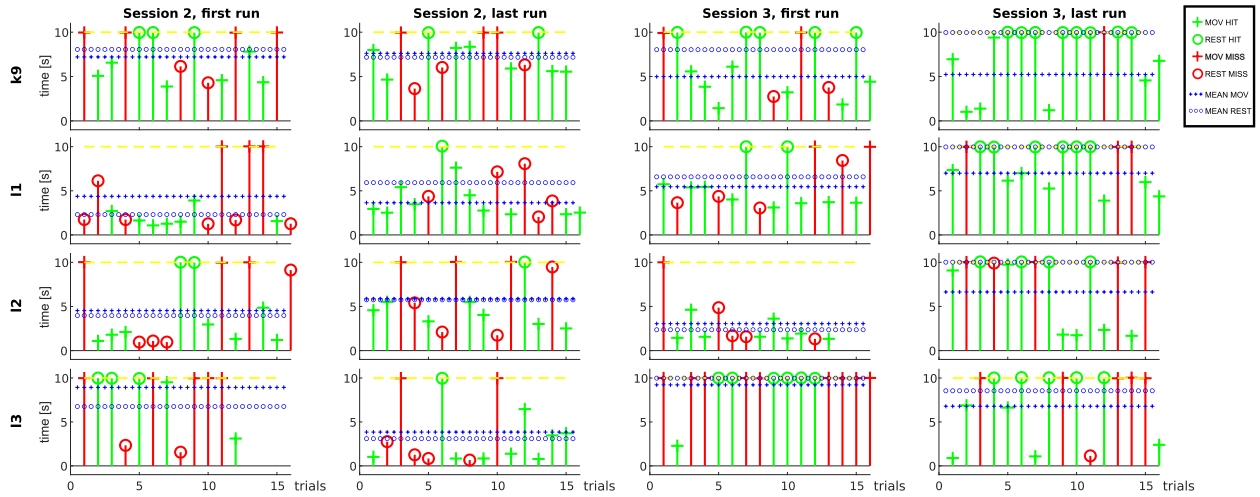


Figure 7: Single trial outcomes of online evaluation runs. For each evaluation session the first and the last run are considered. Green labeled stems represent a positive outcome while red stem correspond to failed attempts.

trial-level metrics and response efficiency despite minimal parameter adjustments and no decoder recalibration. L1 and L3 participants exhibit milder but steady statistically significant improvements, while L2 shows the lowest improvements, even if with classification accuracies always above chance level. Despite this, after classifier updates and hyperparameters tuning processes, also L2 showed some improvements in the control of the LLE at the end of the experiment, as visible in Fig. 7, similar to the other participants. These observations emphasize the need for subject-specific strategies to accommodate heterogeneous learning and control capabilities. Overall, the results support the interpretation that even a limited number of training sessions can induce measurable user adaptation, contributing to system-level performance improvements. In particular, the dissociation observed between single-sample and trial-based accuracy in some

participants suggests that learning may manifest as increased temporal stability of neural patterns, rather than as instantaneous classification gains, reinforcing the role of user learning in effective BCI control.

CONCLUSION

The experimental results obtained confirm that the proposed protocol based on repeated sessions, continuous feedback, and limited calibration is capable of supporting the development of a functional LLE-BCI system while simultaneously promoting neural learning processes in the user. The analysis of feature maps reveals a progressive centralization of discriminant information in midline channels toward physiologically meaningful frequency bands (μ and β) over the course of the sessions. This suggests that cortical adaptation occurs associated with

the user's familiarization with the motor imagery task. Although the ERD/ERS represent a simplified characterization of the EEG input compared to the internal representations of deep networks, they are widely used in the literature to interpret task-related dynamics, and still provide meaningful insight into subject-specific improvements. In parallel, improvements in reaction times and an increase in online test accuracy metrics further support the hypothesis that the protocol promotes genuine BCI learning. At the same time, the results highlight the need for a dynamic evolution of classification methodologies capable of adapting to individual peculiarities.

With appropriate training of the user and a shared control policy with the robot, this approach could support continuous lower-limb exoskeleton control driven by motor intention and not constrained by predefined cues; however, this requires highly reliable classification of the rest state, particularly in complex environments.

Limitations and future developments: Although the study achieved positive results, some limitations must be acknowledged. First, the participant cohort is relatively small and does not allow population-level analyses: larger studies would be required to explore whether individual factors influence the modulation of right/left step imagination. Secondly, additional sessions could have provided a clearer assessment of user adaptation over time and allowed the evaluation of the stability of the observed results across sessions, which is currently visible at the end of the third sessions. Another relevant limitation concerns the use of healthy volunteers, as this choice may introduce a translation gap [5]. Future validation on motor-impaired end users will be necessary to assess the clinical applicability of the proposed framework. Finally, although the framework allows the acquisition of lateralized stepping data, classifiers were trained to decode a generic stepping command. This design choice enabled robust control for novice users, but precluded the evaluation of left-right motion intention discrimination. Future work may investigate lateralized decoding after extended training.

ACKNOWLEDGMENT

This work was funded by Next Generation EU - "Age-It - Ageing well in an ageing society" project (PE0000015), NRRP - PE8.

REFERENCES

- [1] GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* 2024;23(4):344–381.
- [2] Hsu TH, Tsai CL, Chi JY, Hsu CY, Lin YN. Effect of wearable exoskeleton on post-stroke gait: A systematic review and meta-analysis. *Annals of Physical and Rehabilitation Medicine.* 2023;66(1):101674.
- [3] Hramov AE, Maksimenko VA, Pisarchik AN. Physical principles of brain–computer interfaces and their applications for rehabilitation, robotics and control of human brain states. *Physics Reports.* 2021;918:1–133.
- [4] Frolov AA et al. Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface (bci)-controlled hand exoskeleton: A randomized controlled multicenter trial. *Frontiers in Neuroscience.* 2017;11:400.
- [5] Tonin L, Millán JdR. Noninvasive brain–machine interfaces for robotic devices. *Annual Review of Control, Robotics, and Autonomous Systems.* 2021;4:191–214.
- [6] Khan MA, Das R, Iversen HK, Puthusserypady S. Review on motor imagery based bci systems for upper limb post-stroke neurorehabilitation: From designing to application. *Computers in Biology and Medicine.* 2020;123:103843.
- [7] Saha S, Baumert M. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: A review. *Frontiers in computational neuroscience.* 2020;13:87.
- [8] Perdakis S, Tonin L, Millán JdR. Brain-machine interfaces: A tale of two learners. *IEEE Systems, Man, and Cybernetics Magazine.* 2020;6(2):32–42.
- [9] Ceradini M, Tortora S, Micera S, Tonin L. The effect of user learning for online eeg decoding of upper-limb movement intention. *IEEE Transactions on Medical Robotics and Bionics.* 2025;7(2):633–641.
- [10] Tortora S et al. Neural correlates of user learning during long-term bci training for the cybathlon competition. *Journal of NeuroEngineering and Rehabilitation.* 2022;19:69.
- [11] Tonin L, Beraldo G, Tortora S, Menegatti E. Ros-neuro: An open-source platform for neurorobotics. *Frontiers in Neurorobotics.* 2022;16:886050.
- [12] Tariq M, Trivailo PM, Simic M. Eeg-based bci control schemes for lower-limb assistive-robots. *Frontiers in human neuroscience.* 2018;12:312.
- [13] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering.* 2018;15(5):056013.
- [14] Ishikawa S, Yamada M, Bao H, Takezawa Y. Phinets: Brain-inspired non-contrastive learning based on temporal prediction hypothesis. In: *International Conference on Representation Learning.* 2025, 68386–68412.
- [15] Hou Y et al. Gcns-net: A graph convolutional neural network approach for decoding time-resolved eeg motor imagery signals. *IEEE Transactions on Neural Networks and Learning Systems.* 2022;35(6):7312–7323.
- [16] Tortora S et al. Effect of lower limb exoskeleton on the modulation of neural activity and gait classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering.* 2023;31:2988–3003.
- [17] Espenhahn S et al. Sensorimotor cortex beta oscillations reflect motor skill learning ability after stroke. *Brain Communications.* 2020;2(2):fcaa161.

The Effect of Single Session Implicit Motor Learning on Finger Motor Imagery for EEG BCI

Laura M. Damm, Dai Jiang, Andreas Demostheneous

Department of Electronic and Electrical Engineering, University College London, London, UK

E-mail: laura.damm.23@ucl.ac.uk

ABSTRACT

Motor imagery (MI) based brain computer interfaces (BCIs) rely on cue driven paradigms, yet little is known about how the order in which cues are presented influences neural discriminability. Here the effect of sequential versus random cue order on finger-MI is investigated. Ten healthy, right-handed participants (7 F, mean age = 31.9 ± 10.7 years) completed two runs of 300 trials each (30 cycles of 5-finger trials). Across subjects, the γ -band integrals showed the most consistent cue-order effect. Depending on frequency band 7-8 subjects yielded significant effects. Participants who began with the sequential cues exhibited higher mean amplitudes and reduced variability across runs, suggesting that early exposure to a predictable cue sequence facilitates implicit motor-learning and stabilises the forward model. These findings demonstrate that cue-order manipulation is a potent factor influencing MI-BCI performance and underscore the importance of designing training protocols that promote implicit learning.

INTRODUCTION

The development of brain computer interfaces (BCIs) towards more complex systems, introduces a need to consider the effect of paradigm choice for data collection. The combined efforts of non-invasive BCI and finger movement research enables the consideration of a neurorehabilitative system that can decode imagined finger movement [1]. This is especially crucial for recovery after stroke, where the motor imagery (MI) of finger movement is still executable, but motor execution (ME) frequently is not [2]. The development towards such intricate tasks in the context of MI BCI research requires the evaluation of each aspect that affects the data discriminability in a classification context.

The neural nature of individual finger movement is complex and delicate. Pioneering studies utilising invasive recording methodology such as electrocorticography (ECoG) have allowed for initial analysis of the neural correlates underlying contextual ME and MI [3]. However, invasive recording is not feasible or even possible for all required implementations of finger movement analysis. In the context of neurorehabilitation for example a non-

invasive approach such as electroencephalography (EEG) is more feasible as it is comparatively low in cost and recovery efforts. Despite drawbacks such as low signal to noise ratio (SNR) and decreased spatial resolution, EEG has been successfully implemented for finger movement studies [4].

Established feature extraction methodologies such as filter bank common spatial patterns (FBCSP) have shown moderate success in a variety of EEG BCI studies for finger movement [5]. Drawing from these observations, it is possible that the delicate nature of finger movement requires to branch out and explore feature patterns outside the status quo. Especially where the essence of MI studies must be considered, as event initiation and execution are not as controllable as for ME.

Understanding the physiological nature underlying finger movement and how it is recorded in EEG, is important to consider for feature engineering approaches. Where the impact of implicit motor skill learning has been assessed in paradigms such as the serial reaction time tasks, the effect on imagined finger movement has not been contemplated [6]. As such this paper focuses on determining the effect of randomly versus sequentially cued finger MI instructions within a single session design. Specifically focusing on the potential effect of participants being introduced to the overall task with either random or sequential cue order. This permits to consider the effect of implicit motor learning in the context of finger MI. Implicit motor learning refers to the gradual and unconscious sensorimotor mapping that occurs during familiarisation with instruction patterns [7]. This ties into the predictive coding models that assume that predictability leads to heightened β -band synchrony and α -band inhibition [8].

The hypothesis addressed in this paper is that participants introduced to the task with sequential cue order will show less discriminability between cue orders. The research question considered whether cue order has a significant effect on implicit motor learning for finger movement MI tasks.

MATERIALS AND METHODS

Dataset:

A total number of 10 right-handed subjects was recruited (7 women, mean age 31.9, SD = 10.7) to participate in a single session EEG MI task. Prior to EEG cap application

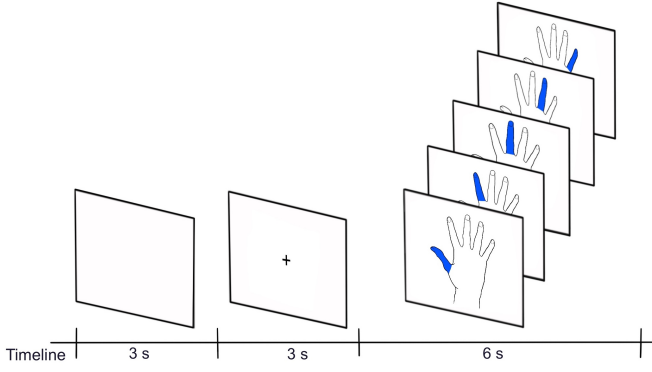


Figure 1: Sequential cue order instructions for finger MI.

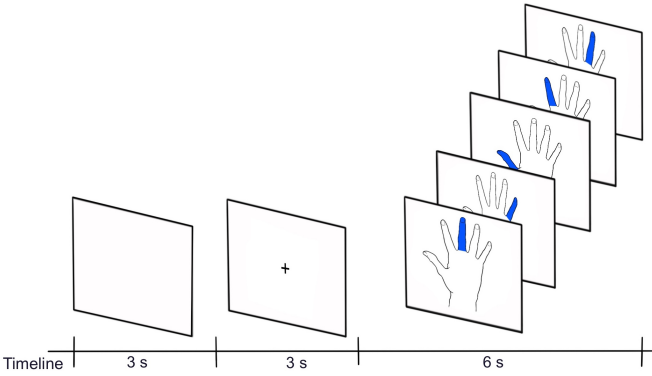


Figure 2: Random cue order instructions for finger MI.

participants were informed about the proceedings of the test as well as potential risks to provide informed consent. The EEG data was amplified via the g.tec HiAmp which was connected to a laptop via USB-A [9]. The recording was executed with 64 g.Scarabeo active Ag/AgCl electrodes of which 2 were placed on the mastoids as reference. Electrode placement was done with the international 10-10 system, the locations of which were associated with the electrode number during processing. Here the recording loop was built in Simulink incorporating g.tec's acquisition system for accurate recording. The paradigm presentation was initialised within the recording loop to allow for maximal alignment at a sampling rate of 512 Hz.

Each subject completed 30 cycles of 5 finger trials for each cue order, leading to 300 trials per subject. Each cue order was separated into two 15 minute cycles between which participants had a short break. Each individual finger trial was initialised with a 3 second white screen, followed by a 3 seconds of focus cross prior to the event initialisation visual which lasted for 6 seconds. A visualisation of a single trial run is given in Fig. 1 for sequential cue order and Fig. 2 for random cue order. The cue order specifically refers to the order in which each finger was presented to the participants within a single trial. Meaning sequential cues were ordered in thumb, index, middle, ring, and pinky for each trial, while random cues were in different order for each trial. When switching to the next cue order participants were given several minutes of rest.

Subject	Run 1	Run 2
1	Sequential	Random
2	Random	Sequential
3	Sequential	Random
4	Random	Sequential
5	Sequential	Random
6	Sequential	Random
7	Random	Sequential
8	Random	Sequential
9	Sequential	Random
10	Random	Sequential

Table 1: Subjects and the cue of each run.

The order in which each subject was presented with the paradigm was interchangeably controlled to assess any potential effects of exposure to random or sequential in the first run where Table 1 shows the order for each subject.

Analysis:

Prior to feature extraction and analysis the EEG data was preprocessed. A bandpass filter of 0.5 to 45 Hz was applied to filter out noise as is standard practice for EEG preprocessing. In the same vein, common average referencing (CAR) was utilised for offline re-referencing, before the data was segmented into epochs with a 7 second length. This included 2 seconds of pre-event data and a 5 second post-event data. This was repeated for each run and order and executed using the *mne* python library [10]. The epoched data was analysed with a sliding window integral moving across each epoch: Here the EEG data was initially bandpass filtered to analyse the individual frequency bands of alpha, beta, gamma, and theta. Each window was the same length as the sampling frequency of 512 Hz, where an overlap of 256 samples permitted to consider transient changes within each window:

$$I_n = \int_{\frac{n \cdot \text{step}}{f_s}}^{\frac{n \cdot \text{step} + L_{\text{win}}}{f_s}} |x(t)| \quad (1)$$

where the trapezoid weighting was considered:

$$dt \approx \Delta t \left[\frac{1}{2} |x_{n \cdot \text{step}}| + \sum_{k=1}^{L_{\text{win}}-2} |x_{n \cdot \text{step}+k}| + \frac{1}{2} |x_{n \cdot \text{step}+L_{\text{win}}-1}| \right] \quad (2)$$

where $\Delta t = 1/f_s$ and $\text{step} = L_{\text{win}} - 256$.

For equation 1 each channel and label was assessed individually. In equation 2 the first and last step of the discrete timeframe were weighted to consider transient changes across each window. This process created a dataframe containing epoch number, finger label, frequency band, channel, window index, and the associated integral value informing about the summed area under the curve in this window. This detailed dataframe enabled an in-depth analysis of the differences between cue order and between runs. For cue order analysis it was first determined if the data at hand was normally distributed, which allowed for ANOVA measurement. If the data was not normally distributed Kruskal-Wallis was applied.

The significance levels were corrected with the bonferoni method to ensure validity of the results in the face of repeated measures. For analysis of the discrepancies between runs either a paired t-test for normally distributed data or a Wilcoxon Signed Rank (WSR) for non-normally distributed data was used. All of the code utilised for statistical analysis was taken from the *statsmodels* python library [11].

RESULTS

The analysis of differentiability between fingers based on sliding window integral activity found that no significant difference occurred between fingers. This indicates that all five fingers initiate correlated temporal-spectral patterns.

Orderwise Comparison:

This section contains Tables 2 through 5 considering the effect of random versus sequential order independent of the order it was introduced to the participants. As such the effect of the order alone and if it had a significant impact was considered. For the θ -, α - and β -band 7 out of 10 subjects yielded significant difference between orders. For the γ -band 8 out of 10 subjects were significant. Notably across all four bands only one subject S5 consistently showed no significant difference between cue orders.

Subject	H Value	Corrected P
1	1.94	0.66
2	23.19	0.00
3	129.33	0.00
4	134.82	0.00
5	1.69	0.78
6	10.5	0.01
7	66.31	0.00
8	124.27	0.00
9	5.19	0.09
10	20.92	0.00

Table 2: Kruskal-Wallis analysis of the α -band sliding window integral differences between sequential and random cue order.

Subject	H Value	Corrected P
1	64.68	0.00
2	8.02	0.00
3	98.2	0.00
4	134.92	0.00
5	0.29	1.00
6	69.93	0.00
7	48.06	0.00
8	133.76	0.00
9	1.98	0.64
10	2.32	0.5

Table 3: Kruskal-Wallis analysis of the β -band sliding window integral differences between sequential and random cue order.

Runwise Comparison:

Subject	H Value	Corrected P
1	56.6	0.00
2	38.15	0.00
3	50.78	0.00
4	135.03	0.00
5	3.17	0.3
6*	64.88	0.00
7	92.66	0.00
8	139.75	0.00
9	10.75	0.00
10	0.46	1.00

Table 4: Kruskal-Wallis analysis of the γ -band sliding window integral differences between sequential and random cue order. Subject S6 was normally distributed in this band and ANOVA was applied.

Subject	H Value	Corrected P
1	12.77	0.00
2	18.56	0.00
3	138.86	0.00
4	106.06	0.00
5	0.29	1.0
6	2.12	0.58
7	79.18	0.00
8	93.29	0.00
9	11.49	0.00
10	4.88	1.00

Table 5: Kruskal-Wallis analysis of the θ -band sliding window integral differences between sequential and random cue order.

Where the previous section showcased the results of indiscriminately analysing differences between cue orders, this section considers cue order introduction. Here Figures 3 and 4, consisting of eight plots, visualise the effect of random versus sequential cue order as first task introduction. Specifically how each frequency band was impacted and how the mean sliding window integral for each subject differs between runs. Noticeably, the subjects who started with random order showcased a significant difference between order runs compared to those who started with sequential. Most display an increase in mean integral value when engaging with the sequential cue order as their second run. This distinction was not as strong in those who started with sequential and had random as their second run, where only S3 showcases significant change.

DISCUSSION

This paper sought out consider how cue order affects implicit MI learning for finger movement, where the main hypotheses cannot be rejected. Participants introduced to sequential cue order first overall elicited stronger signals that were more stable across the two cue orders, where a significant difference between them was given in all but one subject.

When examining Tables 2 through 5 it is discernable that most subjects experience significant difference across cue order in at least one of the frequency bands. What stands

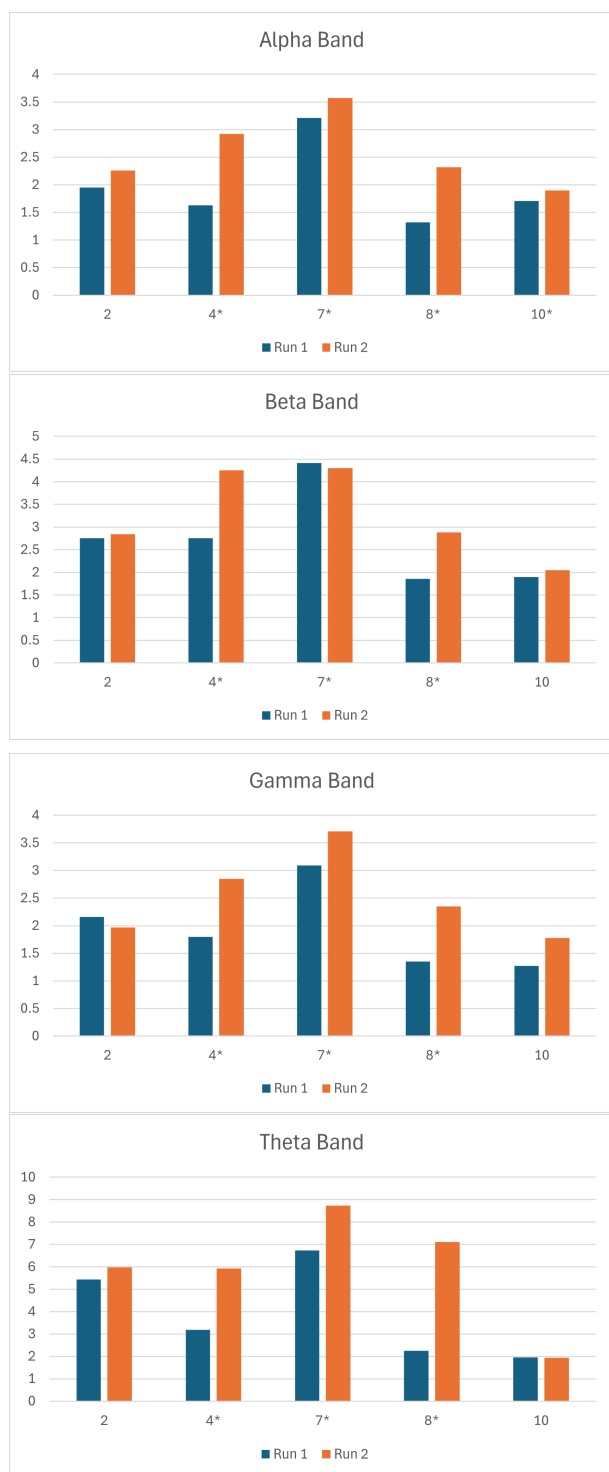


Figure 3: Sliding-window integral mean for each frequency band (α , β , γ , θ) separated by run order where subjects first executed **random order**. The y-axis holds the signal amplitude in that band. An asterisk (*) marks a statistically significant difference between runs.

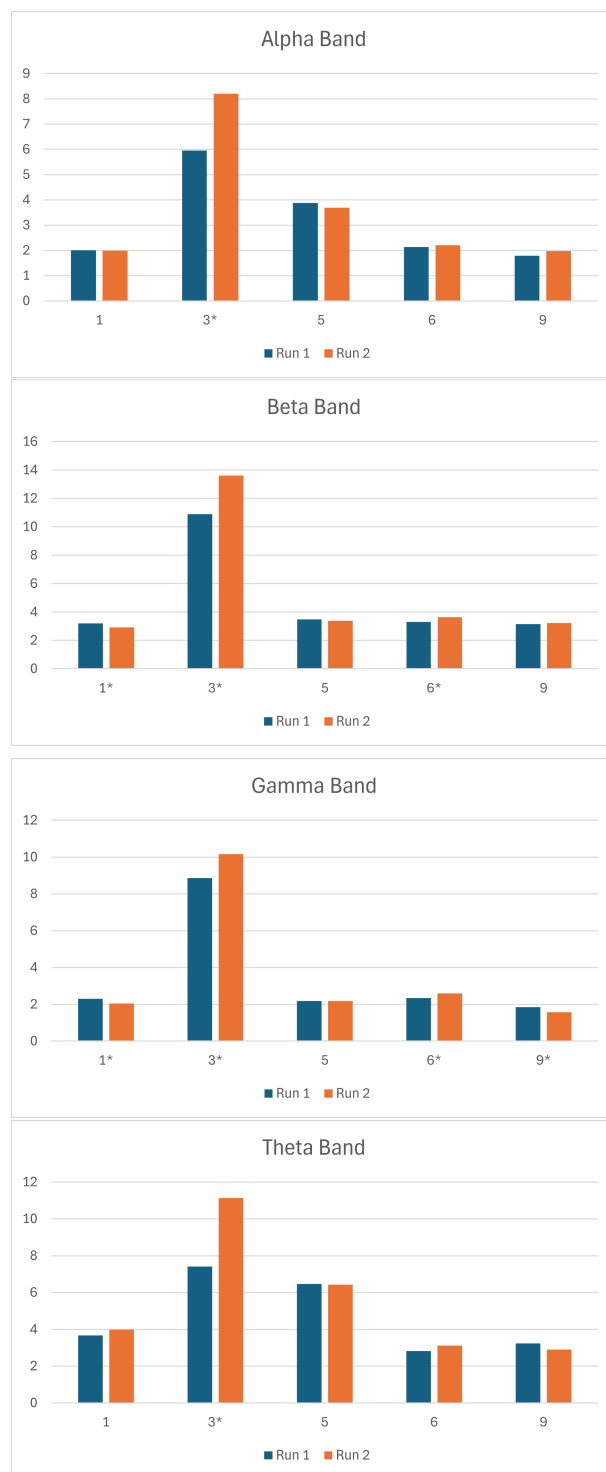


Figure 4: Sliding-window integral mean for each frequency band (α , β , γ , θ) separated by run order where subjects first executed **sequential order**. The y-axis holds the signal amplitude in that band. An asterisk (*) marks a statistically significant difference between runs.

out is that the α - and β -band discriminability is not as strong as anticipated for every subject. Instead most subjects demonstrate cue order effect in the γ -band. The relevance of the γ frequency band for finger movement has been demonstrated, where the effect of cue order mirrors this [12].

An important observation is that the subjects who did not demonstrate a cue order effect in α - or β -band show a significant effect in the γ - and θ -band, see subjects S1 and S9. In a similar vein, subjects whose γ - and θ -band did not allow for significant differentiability demonstrate it in α - and β -band. Especially subject S1 stands out, where only the α -band reports a significant effect.

The effect that introduction order can have on the discernability between cue orders is visualised in Figures 3 and 4. Here it is shown that random cue order as task introduction potentially inhibits the forward model to develop reliable sensorimotor mapping. As such it is reasonable to consider that the task introduction cue order has a significant impact on the overall performance. This is mirrored in the strength mean power compared between sequential and random across all bands as demonstrated in Figure 4. In direct comparison to Figure 3 where differences between Run 1 and Run 2 are more commonly significant and stronger overall. For participants who were introduced via sequential order only subject S3 performed consistently with discriminability across cue orders, signifying a stronger forward model initiation.

The nature of the data at hand allows for detailed temporal-spectral analysis of the amplitude power within each frequency band. This permits to infer underlying neural behaviour, such as synchronisation and desynchronisation patterns [13]. The mean amplitude reported for the β -band for example demonstrates a stronger amplitude in the β -band for those who did the sequential order in their first runs. In contrast, those exposed to random order first achieved significantly lower mean amplitude. This mirrors the expected event related synchronisation (ERS) and desynchronisation (ERD) patterns associated with task familiarity and implicit motor learning [14]. This behaviour is mirrored across all frequency bands, indicating that while the discriminability might not always be significant, the ERDS behaviour mirrors known physiological patterns.

In the context of EEG BCI for finger movement decoding these results allow to determine the positive effect of sequential task presentation. This is extrapolated from the mean amplitude achieved for those who started with sequential cue order. Furthermore, the low discriminability between cue orders if started with sequential showcases the reliability of signal over task adaptability.

Future research must consider how to utilise this observed effect to extrapolate delicate patterns such as finger MI more reliably. While no observed effect could be reported between fingers for sliding window integrals in this instance, the information on cue order effect on implicit motor learning enables the development of a stronger effect paradigm.

CONCLUSION

This paper examined the effect of cue order on delicate MI tasks, specifically for the temporal-spectral domain. The results allow for future feature engineering to consider temporal effects of neural behaviour. Specifically, that sequential cue order as introduction to a task promotes implicit motor learning in comparison to random cue introduction.

ETHICS AND FUNDING

The data used in this paper was collected with approval of the UCL Humanities, Arts and Sciences Research Ethics Committee. Ethics approval number: 26907/001. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/S022139/1 - The Centre for Doctoral Training in Connected Electronic and Photonic Systems.

REFERENCES

- [1] Lee HS et al. Individual finger movement decoding using a novel ultra-high-density electroencephalography-based brain-computer interface system. *Front. Neurosci.* 2022;16.
- [2] Hardwick RM, Caspers S, Eickhoff SB, Swinnen SP. Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neuroscience & Biobehavioral Reviews*. 2018;94:31–44.
- [3] Flamary R, Rakotomamonjy A. Decoding Finger Movements from ECoG Signals Using Switching Linear Models. *Front. Neurosci.* 2012;6.
- [4] Kaya M, Binli MK, Ozbay E, Yanar H, Mishchenko Y. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Sci Data*. 2018;5(1).
- [5] Damm LM, Jiang D, Demosthenous A. Autoregressive Artefact Detection in Finger Movement Decoding for Electroencephalography Brain Computer Interfaces. *IEEE Sensors Journal*. 2025.
- [6] Robertson EM. The Serial Reaction Time Task: Implicit Motor Skill Learning?: Figure 1. *J. Neurosci.* 2007;27(38):10073–10075.
- [7] Friston K, Kiebel S. Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B*. 2009;364(1521):1211–1221.
- [8] Pfurtscheller G, Brunner C, Schlögl A, Lopes Da Silva F. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*. 2006;31(1):153–159.
- [9] G.tec Medical Engineering GmbH Austria HiAmp 64 Channel Biosignal Amplifier. [Online]. Available: <https://www.gtec.at/product/g-hiamp-256-channel-biosignal-amplifier/>.
- [10] Zubarev I, Vranou G, Parkkonen L. MNEflow: Neural networks for EEG/MEG decoding and interpretation. *SoftwareX*. 2022;17:100951.

- [11] Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference. 2010.
- [12] Chen Q, Flad E, Gatewood RN, Samih MS, Krieger T, Gai Y. Gamma oscillation optimally predicts finger movements. *Brain Research*. 2025;1848:149335.
- [13] Pfurtscheller G, Lopes Da Silva F. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [14] Zhang S et al. Analysis and Classification for Single-Trial EEG Induced by Sequential Finger Movements. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE: Berlin, Germany, Jul. 2019, 4541–4544.

Diversity in EEG: Why is it important?

Merlin Angel Kelly¹ and Rishan Patel¹, Peter Bryan², Lexi Kier³, Tom Carlson¹

¹Aspire Create, University College London, London, United Kingdom

²Department of Bioengineering, Imperial College London, London, United Kingdom

³Department of Kinesiology, University of North Carolina at Greensboro, Greensboro, North Carolina, USA

E-mail: ucabmk5@ucl.ac.uk, uceerjp@ucl.ac.uk

Electroencephalography (EEG)-based brain-computer interfaces (BCIs) are increasingly deployed across clinical and consumer contexts, yet the research foundations underpinning these systems reflect a profound lack of demographic diversity. Hardware limitations, particularly the inability of standard EEG electrode systems to achieve reliable scalp contact across curly, kinky, and afro-textured hair types, represent a concrete and under acknowledged barrier to inclusive data collection. We present findings from a mixed-methods survey of 31 EEG researchers and practitioners, finding that 74% reported hardware-related challenges with diverse hair types, with difficulties clustering at cap placement and preparation. Qualitative themes revealed patterns of signal degradation, extended setup burden, explicit participant exclusion, downstream data loss, and a systemic researcher knowledge gap. Existing hardware solutions only partially address this problem. We argue that electrode redesign capable of achieving reliable scalp contact across all hair types is a prerequisite for BCI systems that work equitably for all.

INTRODUCTION

Brain-computer interfaces (BCIs) represent a transformative technology that enables users to control external devices by decoding neural activity, bypassing the need for muscular involvement [1, 2]. Electroencephalography (EEG)-based BCIs have been widely adopted across applications ranging from motor rehabilitation for stroke and spinal cord injury patients to communication aids for individuals with locked-in syndrome [3], and increasingly into domains serving healthy users such as education, entertainment, and occupational settings [4]. As BCI technology matures and its market expands globally, a critical question emerges: for whom are these systems actually being developed and validated? A growing body of evidence from the broader neuroscience literature reveals that the research foundations upon which BCI systems are built suffer from a profound lack of diversity, a problem that threatens to undermine the efficacy, safety, and equitable deployment of BCI technologies for large segments of the global population.

The diversity crisis in neuroscience is well documented

[5, 6]. The overwhelming majority of neuroscience research has been conducted with participants drawn from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) societies [6, 7], with studies finding that 96% of cognitive psychology participants came from Western industrialised countries and 90% of neuroimaging studies were conducted in Western nations [8]. Even more troublingly, most studies fail to report basic demographic information about their participants: fewer than 4% of articles in leading cognitive neuroscience journals reported the race and ethnicity of their samples, with not a single EEG study in that review reporting racial demographics [9]. Where demographic data are reported such as the UK Biobank, the world's largest neuroimaging database, comprises approximately 95% white participants [10, 11], and a review of 20 global neuroscience clinical trials found that 85.6% of the participants were white [12]. An examination of the major publicly available BCI datasets that drive the current algorithm development reinforces this concern.

Existing motor imagery datasets, while representing important scientific contributions, consistently reflect a narrow pattern of participant recruitment. For instance, the OpenBMI dataset [1] includes 54 subjects drawn exclusively from Korea University, and a multi-day motor imagery dataset [2] comprises 62 healthy, right-handed university students aged 17–30 recruited from the World Robot Conference Contest in China. Similarly, a five-session motor imagery (MI) dataset [13] was recorded from 25 subjects at Shanghai University. The cross-dataset variability study by Xu et al. [14] further illustrates this trend by aggregating eight MI datasets, all obtained under controlled, single-site laboratory conditions. The multi-centre P300 study by Flores et al. [15], the BCI competitions [16], and the EEG foundation challenge by Aristimunya et al. [17] made explicit progress by combining datasets from different countries and hardware configurations. Across these efforts, participants are overwhelmingly drawn from single institutions within single countries, and while some datasets represent geographic diversity beyond Western contexts, none report racial, ethnic, or hair-type characteristics. This absence of demographic reporting makes it impossible to evaluate the representativeness of these datasets, reinforcing

a broader pattern in the neuroscience literature in which homogeneity is not documented but also not challenged. The consequences of this demographic homogeneity for BCI performance are substantial and multifaceted. At the most fundamental level, neural structure and function vary meaningfully across populations: race/ethnicity and socioeconomic status contribute to individual differences in episodic memory, attention, and emotion [5]. Cultural factors shape brain development, influencing everything from economic security and nutrition to daily routines, with documented differences in neural function across cultures for theory of mind, emotion regulation, and face perception [18]. Herbert et al. [4] highlight that language, a factor characterised by enormous global diversity, with over 6,500 languages spoken worldwide, is almost entirely unaddressed in BCI research, with most implementations restricted to a single language, primarily English. Given that bilingual individuals exhibit different functional brain organisation compared to monolinguals, including differences in cognitive functions directly relevant to BCI performance such as attention, memory, and imagery, the current monolingual and monocultural approach to BCI development represents a significant limitation [4].

Documented differences between racial and ethnic groups have been demonstrated in neuroscience. EEG studies have found that Black participants show decreased power in delta and beta bands during sleep, independent of socioeconomic factors [19]. While direct evidence in BCI contexts is limited, documented population-level differences in EEG characteristics suggest that assuming universal neural signatures across ethnicities is unwarranted. Lim et al. [20] have documented that Black/African American individuals are twice as likely, and Hispanic/Latino individuals 1.5 times as likely, to develop Alzheimer's disease compared to non-Hispanic white individuals. Critically, this is not merely a difference in prevalence, Black/African American and non-Hispanic white individuals exhibit opposite patterns of Default Mode Network resting-state connectivity in Alzheimer's disease, potentially demonstrating race-specific disease trajectories [21]. BCI-based diagnostic tools trained predominantly on white participants may therefore not only underperform but actively mischaracterise disease states in Black patients, precisely the population bearing the greatest disease burden.

Beyond specific clinical populations, demographic differences in neural characteristics have broader implications for BCI algorithms. Inter-subject variability is already recognised as one of the most significant challenges in the BCI field, independent of demographic diversity. Maswanganyi et al. [22] demonstrated that merging EEG signals acquired from different subjects resulted in significant declines in intention detection rate, with accuracy drops of up to 21% across sessions. Lee et al. [1] found that a lack of BCI response affected 53.7% of users in motor imagery paradigms, with large performance variations between both subjects and sessions.

Shen et al. [23] showed that EEG variability poses a great challenge to affective BCI frameworks, with inter-subject emotion commonality reaching only 30–40% even under optimised conditions. If this variability is already problematic within demographically homogeneous samples, the challenge will be considerably amplified as BCI systems are deployed to genuinely diverse populations whose neural data may differ systematically from the training distributions.

The implications compound further as BCI research increasingly relies on transfer learning and deep learning approaches to overcome inter-subject variability. Xu et al. [14] demonstrated that deep learning models trained on one dataset are unable to generalise well to other datasets, with models becoming highly specialised to their training dataset structure. Xu et al. [24] further showed that cross-dataset variability, arising from differences in equipment, electrode configurations, and recording environments, significantly reduces the transfer performance of deep learning models. Han et al. [25] developed graph neural network frameworks specifically to address the challenge of heterogeneous electrode configurations across datasets, demonstrating that even technical differences between recording setups create substantial barriers to model generalisation. Critically, these studies identify both physiological variability (between subjects) and environmental variability (between recording setups) as sources of distribution shift [14]. When the training data underlying these algorithms systematically excludes certain populations, the models will not merely fail to generalise, they will embed and amplify existing biases. As Kam et al. [26] warn, AI models trained on unrepresentative data will inherit and amplify existing biases, a concern that extends directly to the machine learning classifiers at the heart of modern BCI systems.

The under-representation of diverse participants in EEG research is not incidental, it is structurally embedded in the hardware upon which data collection depends. Non-invasive BCI systems rely fundamentally on EEG signal acquisition, yet EEG electrode and cap hardware is not designed to work with the full range of human hair types [27]. Traditional electrode systems allow only approximately one centimetre of space between electrode and scalp, meaning that thick, voluminous, curly, and coily hair types (more common among individuals with African or Caribbean heritage) cannot be accommodated as shown in Figure 1 [7, 27, 28]. This phenotypic bias in instrumentation results in black participants being excluded from EEG research at substantially higher rates, with a review of 81 EEG papers finding that only five included Black participants and none confirmed whether their data survived quality checks [7]. When participants with thick or curly hair are successfully fitted, they face longer and more burdensome data collection sessions and higher rates of data exclusion during post-processing [27]. These systematic hardware barriers mean that the neural signatures, decoding algorithms, and normative baselines that underpin BCI systems have been derived

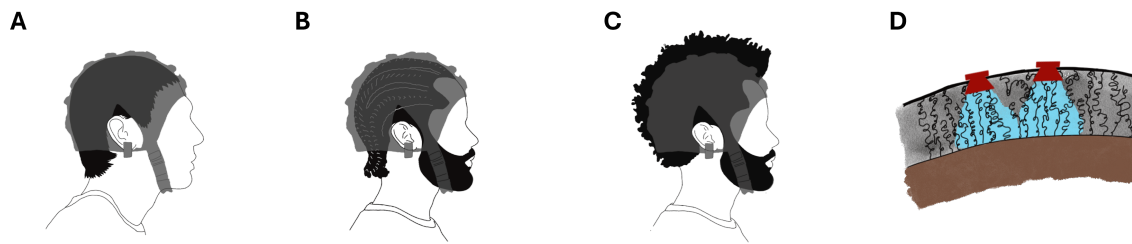


Figure 1: Image A is an example of an EEG cap fitting on a user with straight and thin hair, allowing minimal gel to obtain a signal. Image B demonstrates placing a cap on curly hair in a protective style, creating further distance from the scalp leading to bridging and excess gel as shown by image D. Image C is an example of curly and long hair, which would lead to the cap not fitting over the users head, leading to discomfort and further distance from the electrode and the scalp.

from an unrepresentative subset of the global population. In this paper, we asked EEG researchers for their perspective on this phenotypic bias, if it impacts their research, and what they believe the solution is to better understand this systematic barrier. Further, we discuss previous solutions which address this hair bias and the pitfalls faced by these prototypes. Finally, we conclude with recommendations for the field of BCI to improve diversity in datasets, in turn improving outcomes for ethnicities that are often neglected in academic studies.

MATERIALS AND METHODS

A mixed-methods survey was designed to characterise the prevalence and nature of EEG hardware challenges encountered when recording from participants with curly, kinky, or afro-textured hair. The survey was distributed via LinkedIn, Discord, and WhatsApp within communities of EEG researchers and practitioners, and was open to individuals with any level of experience conducting or participating in EEG research. As the surveys were fully anonymous on non-sensitive topics, this study was considered minimal risk by UCL ethics and did not require ethical approval. A total of 31 respondents completed the survey, though not all respondents answered every question; response counts are reported per item where relevant. The survey comprised both closed-response items (yielding quantitative data on challenge prevalence, stage of occurrence, and appetite for solutions) and open-ended questions (capturing specific challenges, perceived causes, downstream research impacts, and suggested solutions). Qualitative responses were synthesised thematically to identify dominant patterns across the sample.

We acknowledge several limitations: the sample is self-selected and modest in size, respondents were not asked to report their own demographic characteristics or geographic location, and recruitment through research-adjacent online communities may over-represent individuals already sensitised to diversity concerns in EEG research. The survey findings should therefore be interpreted as practitioner-level qualitative evidence that complements the published literature, rather than as a representative epidemiological estimate of challenge prevalence across the field.

RESULTS

Respondent Profile: Respondents represented a broad range of EEG research contexts, including BCI paradigms (motor imagery, SSVEP, P300, speech decoding), clinical applications (stroke rehabilitation, neurodegenerative disease), and cognitive and developmental neuroscience. This breadth indicates that the hardware challenges described below are not confined to a single subfield but are distributed across the landscape of EEG research.

Prevalence of Hardware Challenges: Of the 31 respondents, 23 (74%) reported having experienced challenges when using EEG systems with participants who have curly, kinky, or afro-textured hair. This rate is consistent with the systematic exclusion patterns documented in the published literature and suggests that hardware-related barriers are a routine, rather than exceptional, feature of EEG data collection practice.

Stage of Challenge: Among those who reported challenges, difficulties were most frequently encountered during cap placement ($n=24$), preparation ($n=19$), recording ($n=15$), and post-session physical cleaning ($n=6$). The fact that challenges cluster at cap placement and preparation before any neural data has been collected underscores that the barrier is primarily one of hardware design rather than signal processing.

EEG Systems in Use: Respondents reported using a range of commercial EEG systems, with g.tec the most commonly used ($n=9$), followed by BioSemi ($n=7$), Brain Products ($n=3$), NeuroScan ($n=3$), EGI/Philips ($n=2$), and a further six respondents specifying other systems including Clarity and custom setups. The distribution of challenge reports across multiple major hardware platforms reinforces that the problem is not specific to any single manufacturer but reflects an industry-wide design norm.

Nature and Consequence of Hardware Challenges: Qualitative responses converged on five interrelated themes. The most consistently reported challenge was signal quality degradation. Respondents described persistently high impedances, elevated noise, and unreliable scalp contact arising from hair volume and density preventing electrodes from reaching the scalp. Several noted that standard remediation strategies - additional conduc-

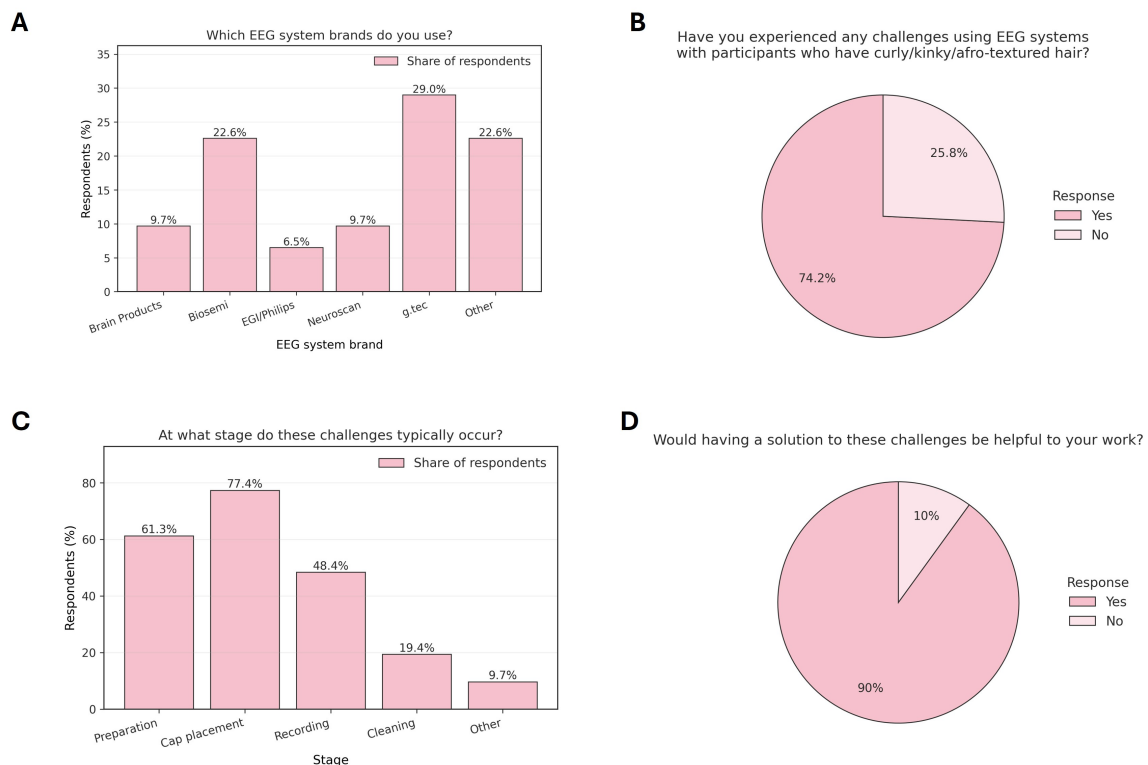


Figure 2: Bar and pie charts highlighting the quantitative results on (A) Users EEG system brand, (B) If users have experienced difficulties with participants hair types/styles, (C) At what stage were there difficulty with hair types/styles, and (D) Would a solution to these challenges be helpful

tive gel and repeated re-gelling - were less effective with curly hair than with straight hair, and one respondent noted that classification performance was markedly poorer for participants with curly or kinky hair compared to the rest of their sample.

The second theme was extended and burdensome setup. Respondents described substantially longer preparation times, with one noting an additional 30–40 minutes of capping time that still did not achieve full contact, and another describing a session in which scalp abrasion during electrode placement resulted in minor bleeding. Gel application was frequently characterised as difficult, with conductive gel failing to penetrate dense hair reliably and drying faster in thicker/curlier hair textures. Post-session gel removal was also identified as a significant burden for participants, with one respondent noting that gel residue persisted in their hair for days after the session.

Third, respondents described patterns of explicit participant exclusion. Several reported that protective hairstyles (braids, cornrows, locs) constituted de facto exclusion criteria in their labs, and that even participants with loose kinky hair required workarounds that were not always successful. One respondent noted they had never had a Black participant return for a follow-up session, pointing to the compounding effect of poor participant experience on longitudinal research and dataset diversity.

Fourth, the downstream consequence was data loss and analytical compromise: noisy data was discarded, preprocessing pipelines required stricter thresholds, and in some

cases entire participant datasets were excluded post-hoc. These losses are commonly attributed to hardware in published methods sections, rendering the demographic skew in resulting datasets invisible.

Finally, respondents identified a researcher knowledge gap as a structural contributor to the problem. Multiple respondents noted that they had received no training in how to handle kinky and curly hair types, and one characterised systematic exclusion as a norm echoed by colleagues across institutions worldwide. This suggests that even where workarounds exist, they are not being disseminated effectively within the field.

Appetite for Hardware Solutions: Of the 30 respondents who answered the final closed question, 27 (90%) indicated that having a solution to these hardware challenges would be helpful to their work. Suggested approaches included hair-penetrating or pin-based electrodes, longer electrode stems to reach through dense hair, more flexible and stretchable cap designs, and electrode systems designed to function independently of hair type. The convergence of respondents on electrode redesign, rather than protocol modifications or participant preparation, as the preferred solution reinforces the framing of this problem as an engineering challenge requiring novel hardware solutions.

DISCUSSION

From our findings, as shown from Figure 2, we can

acknowledge the 74% of researchers mentioning challenges with EEG setup for thick and curly hair mimics the exclusion patterns documented in the literature [7, 27]. The fact that the majority of challenges are faced during the cap placement and preparation highlights that non-inclusive EEG hardware is the leading factor for participant removal as opposed to participant or protocol problems. The qualitative responses of the survey demonstrated themes of high noise and impedance during recordings, leading to extended setup times. Once the data processing starts, researchers often identified data loss or artefacts, leading to removal of participants from the dataset. It was also mentioned that labs often listed hairstyles predominately used by Black demographics in their exclusion criteria.

The compounding exclusion pipeline in EEG research creates a cascade effect. Limitations in EEG hardware lead to participant exclusion or data removal. This in turn creates homogenous datasets for machine learning training and due to this bias in training data would lead to underperformance in the very populations excluded from their development. As there is clear evidence that physiological variability is a source of distribution shift and accuracy reduction in classifiers [14, 22] a systematic exclusion of specific ethnicities will lead to a lack of robustness and generalisability in BCI applications. This highlights that exclusion at the hardware stage does not remain isolated but propagates downstream and impacts all methods which use BCI datasets.

Even though this issue is still faced in both neuroscience and BCI research, meaningful progress has been made in literature. The Sevo electrode [29] is one example where the electrodes utilise braided hair as an anchor to maintain scalp contact. While Sevo electrodes are the most cited solution for interfacing with thick and curly hair, their reliance on braided styles introduces extended setup times and perpetrate the practitioner knowledge gap with diverse hair types. As a result, the implementation of these electrodes in academia does not fully address researcher's needs for an ideal solution: a system that functions for all hair types and styles, matches wet electrode impedance, and requires no hair preparation or modification.

The diversity deficit in BCI research is not merely an abstract concern about scientific rigour, it is an immediate and escalating barrier to the equitable deployment of a technology that has the potential to transform the lives of millions. BCI systems are moving beyond laboratory prototypes into clinical rehabilitation, assistive communication, smart home control, education, and entertainment [3, 4]. As these systems transition from controlled research settings to real-world deployment, they will encounter the full diversity of human populations: diversity in neural characteristics shaped by genetics, culture, and lived experience [5, 18], diversity in the languages users speak and the cultural frameworks through which they process information [4], and diversity in the physical characteristics that determine whether current EEG hardware can reliably acquire their brain signals at all [7,

27, 30].

If BCI technology replicates patterns of exclusion similar to the field of neuroscience, it risks not only producing inequitable outcomes but further deepening the distrust that already limits participation of under-represented groups in neuroscience research. Conversely, addressing diversity in BCI represents a uniquely actionable opportunity. Unlike many systemic barriers, the technical limitations of EEG hardware represent a clear engineering problem: developing electrodes that achieve reliable scalp contact across all hair types, without requiring participants to alter their hairstyles. Solving this would remove one of the most tangible sources of exclusion [7, 27, 30]. Coupled with systematic demographic reporting, diverse dataset development, and culturally informed BCI design, such innovations could establish an inclusive foundation for BCI systems that work reliably for all users.

This paper highlights the urgent need for BCI researchers to confront the diversity deficit in their field, whilst also analysing survey results that demonstrate the systematic barriers present in EEG hardware. By drawing on the extensive evidence of exclusion documented in broader neuroscience and translating these findings to the specific context of BCI applications, we argue that diversity is not a peripheral concern but a central challenge that will determine whether BCI technology fulfils its transformative potential for all of humanity.

CONCLUSION

The findings presented here confirm that EEG hardware limitations are a routine and systemic barrier to inclusive BCI research, not an edge case. The 74% challenge rate reported by practitioners, concentrated at the point of hardware-participant interaction, demonstrates that exclusion begins before a single neural signal is recorded. This exclusion propagates downstream: into homogenous datasets, into algorithms that cannot generalise to the populations they are intended to serve, and ultimately into clinical and consumer BCI systems that will underperform for those with the greatest unmet neurological need. Existing electrode solutions represent meaningful but incomplete progress, and the field has not yet produced hardware that functions reliably across all hair types and styles while maintaining wet electrode signal quality. As BCI systems transition from laboratory settings to global deployment, the window to address this structural bias is narrowing. Electrode innovation is not a peripheral concern, it is a prerequisite for equitable BCI.

REFERENCES

- [1] Lee MH et al. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience*. 2019;8(5):giz002.
- [2] Yang B et al. A multi-day and high-quality EEG dataset for motor imagery brain-computer interface. *Scientific Data*. 2025;12(1):488.

- [3] Medina ASA, Bonilla MIA, Giraldo IDR, Palacios JFM, Gutiérrez DAC, Liscano Y. Electroencephalography-Based Brain-Computer Interfaces in Rehabilitation: A Bibliometric Analysis (2013–2023). *Sensors*. 2024;24(22).
- [4] Herbert C. Brain-computer interfaces and human factors: The role of language and cultural differences—Still a missing gap? *Frontiers in Human Neuroscience*. 2024;18.
- [5] Dotson VM, Duarte A. The importance of diversity in cognitive neuroscience. *Annals of the New York Academy of Sciences*. 2020;1464(1):181–191.
- [6] Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? *The Behavioral and Brain Sciences*. 2010;33(2-3):61–83; discussion 83–135.
- [7] Webb EK, Etter JA, Kwasa JA. Addressing racial and phenotypic bias in human neuroscience methods. *Nature Neuroscience*. 2022;25(4):410–414.
- [8] Shaaban CE, Rosso AL. Racial, Ethnic, and Geographic Diversity in Population Neuroscience. *Current topics in behavioral neurosciences*. 2024;68:67–85.
- [9] Goldfarb MG, Brown DR. Diversifying participation: The rarity of reporting racial demographics in neuroimaging research. *NeuroImage*. 2022;254:119122.
- [10] Miller KL et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*. 2016;19(11):1523–1536.
- [11] Sudlow C et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*. 2015;12(3):e1001779.
- [12] Rutten-Jacobs L et al. Racial and ethnic diversity in global neuroscience clinical trials. *Contemporary Clinical Trials Communications*. 2024;37:101255.
- [13] Ma J, Yang B, Qiu W, Li Y, Gao S, Xia X. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*. 2022;9(1):531.
- [14] Xu L, Xu M, Ke Y, An X, Liu S, Ming D. Cross-Dataset Variability Problem in EEG Decoding With Deep Learning. *Frontiers in Human Neuroscience*. 2020;14.
- [15] Flores C, Contreras M, Macedo I, Andreu-Perez J. Transfer Learning with Active Sampling for Rapid Training and Calibration in BCI-P300 Across Health States and Multi-centre Data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024;32:3794–3803.
- [16] Blankertz B et al. The BCI competition III: Validating alternative approaches to actual BCI problems. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2006;14(2):153–159.
- [17] Aristimunha B et al. EEG Foundation Challenge: From Cross-Task to Cross-Subject EEG Decoding. *arXiv:2506.19141 [eess]*. Sep. 2025. Accessed: Feb. 13, 2026. [Online]. Available: <http://arxiv.org/abs/2506.19141>.
- [18] Qu Y, Jorgensen NA, Telzer EH. A Call for Greater Attention to Culture in the Study of Brain and Development. *Perspectives on Psychological Science*. 2021;16(2):275–293.
- [19] Hall MH et al. Race and Financial Strain are Independent Correlates of Sleep in Midlife Women: The SWAN Sleep Study. *Sleep*. 2009;32(1):73–82.
- [20] Lim AC et al. Quantification of race/ethnicity representation in Alzheimer’s disease neuroimaging research in the USA: A systematic review. *Communications Medicine*. 2023;3(1):101.
- [21] Misiura MB et al. Race modifies default mode connectivity in Alzheimer’s disease. *Translational Neurodegeneration*. 2020;9:8.
- [22] Maswanganyi RC, Tu C, Owolawi PA, Du S. Statistical Evaluation of Factors Influencing Inter-Session and Inter-Subject Variability in EEG-Based Brain Computer Interface. *IEEE Access*. 2022;10:96821–96839.
- [23] Shen YW, Lin YP. Cross-Day Data Diversity Improves Inter-Individual Emotion Commonality of Spatio-Spectral EEG Signatures Using Independent Component Analysis. *IEEE Transactions on Affective Computing*. 2024;15(1):210–222.
- [24] Xu L, Xu M, Ma Z, Wang K, Jung TP, Ming D. Enhancing transfer performance across datasets for brain-computer interfaces using a combination of alignment strategies and adaptive batch normalization. *Journal of Neural Engineering*. 2021;18(4):0460e5.
- [25] Han J, Wei X, Faisal AA. EEG decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks. *Journal of Neural Engineering*. 2023;20(6):066027.
- [26] Kam JW et al. Creating diverse and inclusive scientific practices for research datasets and dissemination. *Imaging Neuroscience*. 2024;2:imag–2–00216.
- [27] Adams EJ et al. Fostering inclusion in EEG measures of pediatric brain activity. *npj Science of Learning*. 2024;9(1):27.
- [28] Kier L, Brooks DD, Reifsteck EJ, Drollette ES. “Think About Us”: Addressing Racial Biases in Current EEG Cap Design and Methodologies for Inclusion of Black Student-Athletes. *International Journal of Kinesiology in Higher Education*. 2025;0(0):1–15.
- [29] Etienne A et al. Novel Electrodes for Reliable EEG Recordings on Coarse and Curly Hair. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Jul. 2020, 6151–6154.
- [30] Ricard JA, Parker TC, Dhamala E, Kwasa J, Allsop A, Holmes AJ. Confronting racially exclusionary practices in the acquisition and analyses of neuroimaging data. *Nature Neuroscience*. 2023;26(1):4–11.

Decoding Sleep and Wakefulness from Chronically Implanted Intracortical Signals in a Communication BCI User

Elena C. Offenberg¹, Zachary V. Freudenburg¹, Mariana P. Branco¹, Jonas B. Zimmermann², Janir R. da Cruz², Mariska J. Vansteensel¹, and Nick F. Ramsey^{1,3}

¹ Brain Center, University Medical Center Utrecht, Utrecht, Netherlands

² Wyss Center, Geneva, Switzerland

³ Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

E-mail: e.c.offenberg@umcutrecht.nl

ABSTRACT: As brain-computer interfaces (BCIs) move toward long-term home use, continuous operation across day and night introduces new challenges. Decoders trained during wakefulness may produce unintended activations during sleep due to large differences in neural activity between vigilance states. Users must be able to communicate at night, making reliable detection of sleep and wakefulness essential for safe and autonomous BCI operation.

We investigated whether awake-sleep stages can be inferred from chronically implanted intracortical neural signals in a communication BCI user with complete locked-in syndrome. Simultaneous non-invasive electroencephalography (EEG) and electrooculography (EOG) recordings provided reference sleep labels over a 44-hour period. A convolutional neural network was trained to distinguish between wakefulness and N2 sleep using intracortical data and achieved high accuracy ($96.1 \pm 1.6\%$). The model was applied to six independent BCI spelling sessions, where more sleep-classified epochs coincided with reduced spelling performance. These results demonstrate the feasibility of decoding vigilance state from intracortical signals and support the development of state-aware BCIs for continuous home use.

INTRODUCTION

For implanted BCIs to be clinically viable for unsupervised home use, they must provide reliable and continuous operation over extended periods, ideally with 24/7 availability. This requirement is particularly critical for communication BCIs intended for individuals with severe motor impairments, such as those with amyotrophic lateral sclerosis or brainstem stroke, for whom the BCI may represent the primary or sole means of communication. For these individuals, the ability to signal caregivers during nighttime hours is essential for preserving autonomy and safety. At the same time, continuous operation introduces the risk of unintended activations when the user is asleep, or not intending to interact with the device, which may lead to false alarms, caregiver burden, and reduced trust in the system. Robust identification of periods in which the user intends

to engage with the BCI is thus a key requirement for safe and autonomous long-term deployment. One promising approach is the integration of an automated classifier of the user's vigilance state, specifically distinguishing between wakefulness and sleep. This can be used to infer whether BCI control is likely intended (during wakefulness) or clearly not intended and should be suppressed (during sleep), thereby reducing false positives and improving usability while maintaining continuous availability. Ideally, this classification should rely exclusively on neural signals already available from the implanted BCI system, avoiding the need for additional sensors of recording modalities that would complicate system design and limit clinical scalability. Automatic detection of wakefulness, sleep, and sleep stages has been extensively studied using non-invasive electrophysiological recordings, including electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG). Decades of work have established characteristic spectral and temporal features associated with vigilance, and numerous algorithms have achieved high accuracy in polysomnography-based sleep staging [1]. However, the BCIs most suitable for long-term at-home use have been based on chronically implanted electrodes [2, 3], which sample neural activity from a more spatially restricted cortical or subcortical region. The extent to which sleep-wake information can be reliably extracted from such localized intracranial recordings remains insufficiently explored.

Several studies have demonstrated the feasibility of discriminating wake and sleep states using temporary intracranial recordings, including electrocorticography [4], stereoelectroencephalography [5], and intracortical microwires [6], often reporting high classification performance. These investigations suggest that sleep-related neural signatures are detectable from implanted electrodes. However, most of this evidence derives from short-term clinical monitoring in epilepsy or research settings, from relatively distributed brain regions, rather than from spatially focused recordings from chronically implanted BCI users in naturalistic environments. Observations in individuals with long-term implanted BCIs remain rare and reports on nocturnal BCI use, and

the effects of vigilance state highly scarce. In one report, a dedicated “night mode” was introduced to reduce unintended activations, relying on prolonged patterns of user behavior rather than direct neural markers of sleep [7]. Other work has identified neural correlates of drowsiness and linked them to fluctuations in BCI performance [3]. Despite these initial findings, systematic investigation of sleep-wake classification using chronic implanted BCI signals is still sparse, and the generalizability of existing results remains unclear.

MATERIALS AND METHODS

Participant and data acquisition: The data was collected from a participant with late-stage amyotrophic lateral sclerosis (ALS) who had been chronically implanted with two intracortical microelectrode arrays (8×8 electrodes per array, 1.5 mm electrode length, 0.4 mm inter-electrode pitch; Blackrock Microsystems LLC) in the left motor cortex in March 2019 (Figure 1). Neural signals were acquired using a digitizing headstage and a neural signal processor (CerePlex E and NSP, Blackrock Microsystems LLC) and sampled at 30kS/s. All procedures were approved by the Ethical Committee of the Medical Faculty of the Technische Universität München Rechts der Isar, and the participant and his legally responsible family members gave informed consent for the surgical procedure [8].

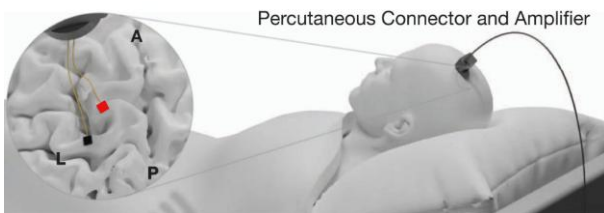


Figure 1: Experimental setup, adapted from [8]. Two microelectrode arrays were placed in the precentral gyrus and superior frontal gyrus. L marks the left central sulcus and A-P the midline from anterior to posterior. The Utah array used in this study, located in the Supplementary Motor Area (SMA), is marked in red.

A dedicated simultaneous recording session lasting over 44 hours (almost two days) was conducted, with the exact duration determined by setup constraints. During this session, non-invasive electrophysiological signals, including EOG and EEG data, were acquired along with the intracranial recordings. EOG data was recorded from two periocular electrodes positioned above the right eye and below the left eye, and EEG data was recorded from 12 scalp electrodes with a sampling rate of 500 Hz and using a 16 channel EEG amplifier (V-Amp DC, Brain Products, Germany) with gel-based Ag/AgCl electrodes. Electrode impedance was checked at the beginning of the recording to be below 5 kΩ. All channels were referenced to an electrode on the right mastoid and grounded to the electrode placed at the FPz location on the scalp.

The participant lacked reliable voluntary eye-movement control and was therefore unable to use eye-tracking-based communication systems, relying instead on an

intracortical BCI spelling interface described previously [8]. Data used in the present study was recorded 459-461 days after implantation.

Sleep stage classification using non-invasive signals: EEG and EOG data recorded during the 44-hour period was preprocessed by band-pass filtering between 0.5 and 40 Hz and downsampling to 256 Hz to match the sampling rate used for preprocessed intracortical data. Outliers were removed using the interquartile range (IQR) criterion by clipping samples outside the range $[Q_1 - 5 \cdot IQR, Q_3 + 5 \cdot IQR]$.

Sleep staging was performed using the SleepyLand toolbox [9], which implements several state-of-the-art automated sleep stage classification algorithms using EEG and EOG data as input. Specifically, we applied DeepResNet [10], SleepTransformer [11] as well as USleep [12] to generate sleep stage predictions for non-overlapping 30-second epochs across the full recording period. Final sleep stage labels were obtained using an unweighted soft-voting ensemble of the three classifiers. The inclusion of both EEG and EOG signals within this multimodal ensemble implicitly accounted for ocular artifacts. These ensemble predictions served as reference labels for subsequent intracortical signal classification.

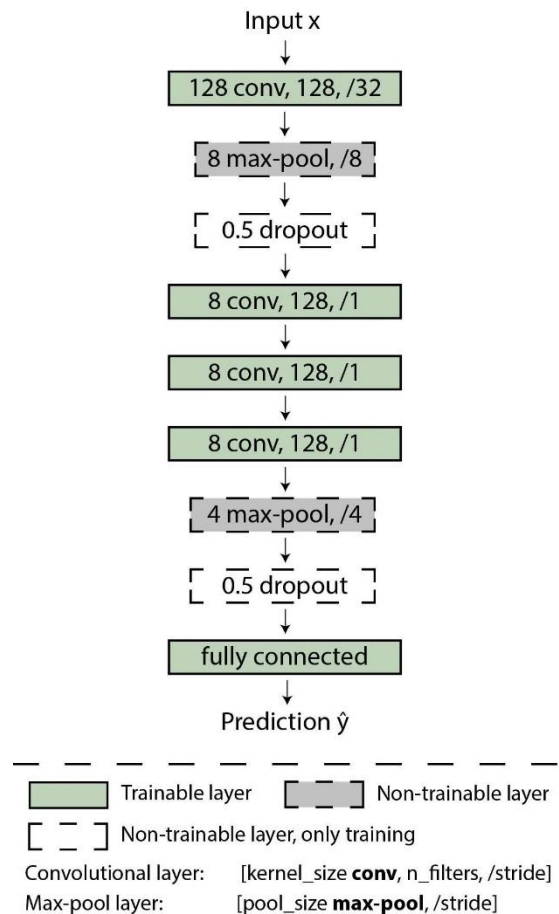


Figure 2: Architecture of the convolutional neural network used for awake-sleep classification. The model was adapted from TinySleepNet [13]. Rectangular blocks denote individual network layers, and arrows indicate the flow of data from 30-second neural signal epochs (x) to the binary sleep state prediction ($\hat{y}$).

Intracortical awake-sleep classification: Intracortical signals recorded from the Utah arrays were downsampled from 30 kHz to 256 Hz using polyphase filtering to prevent aliasing and notch-filtered at 50 Hz and its harmonics to remove line noise. Signals were averaged across electrodes from the array located in the SMA to obtain a single representative time series for classification. Averaging across electrodes was performed to capture global state-dependent changes in neural activity while reducing the influence of channel-specific artifacts and electrode instabilities, and to improve the signal-to-noise ratio of features relevant for vigilance state discrimination.

A convolutional neural network (CNN) was trained to discriminate between wakefulness and sleep (N2 sleep, see next paragraph) using 30-second intracortical signal epochs. The network architecture was based on the convolutional feature extraction component of TinySleepNet [13], excluding the recurrent neural network (RNN) module used for temporal sequence modeling. This architecture was chosen because it enables classification from individual 30-second epochs without relying on temporal context across multiple windows. The CNN comprised four convolutional layers interleaved with two max-pooling layers and two dropout layers to reduce dimensionality and mitigate overfitting. The complete architecture is shown in Figure 2 and was implemented using PyTorch [14].

For model training, all epochs labeled as N2 sleep by the SleepyLand ensemble classifier ($n=743$) were included, along with an equal number of randomly selected epochs labeled as awake to ensure class balance. The N2 sleep stage was selected because it was the most prevalent sleep stage in the dataset.

Each epoch was normalized by subtracting its mean and dividing by its standard deviation. Model performance was evaluated using 10-fold stratified cross-validation. The CNN was trained for 200 epochs using a cross-entropy loss function and the Adam optimizer with a learning rate of 0.001.

Frequency band contribution analysis: To assess the contribution of distinct frequency bands to awake-sleep classification, intracortical signals were decomposed into canonical frequency bands using bandpass filters

following notch filtering at 50 Hz and its harmonics. For the spike band component (250-5000 Hz), bandpass filtering and power extraction were performed prior to downsampling by squaring the signal to obtain a slowly varying amplitude envelope and to prevent aliasing effects. The following bands were analyzed individually: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), low gamma (30-50 Hz), high gamma (60-150 Hz) and spike band power (250-5000 Hz).

For each band-specific dataset, the convolutional neural network was retrained using the same preprocessing, training, and cross-validation procedures described above. Classification performance obtained using each isolated frequency band was then compared with performance obtained using the full broadband signal to quantify the relative contribution of each frequency range to awake-sleep discrimination.

Application to independent BCI spelling sessions: To assess generalizability, the trained classifier was applied to neural data from six independent BCI spelling sessions conducted in the afternoon and evening following the conclusion of the 44-hour continuous recording around noon. During each session, the participant was free to generate self-selected text and produced several sentences in both English and German using an independent spelling interface described in [8]. Spelling performance was evaluated by a fluent bilingual speaker, who assessed the percentage of output words that were comprehensible in either language.

Spelling session data was preprocessed identically to the training data, including downsampling, notch filtering, and normalization. The trained classifier was then applied to the first 20 minutes of each spelling session using sliding 30-second windows with a step size of 1 second. This fixed-duration segment was selected to enable direct comparison across sessions of unequal length. For each epoch, the model generated a binary prediction corresponding to either wakefulness or sleep (N2).

RESULTS

Sleep stage classification using non-invasive signals: Sleep stage predictions for each 30-second epoch are overlaid on the averaged EEG spectrogram across all 12

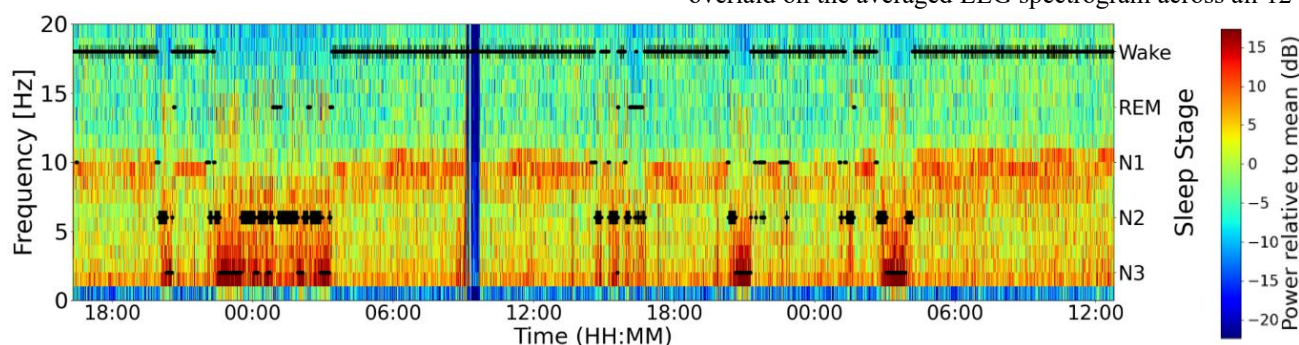


Figure 3: Averaged EEG spectrogram (0-20 Hz) across all 12 channels over the full 44-hour recording period. Predicted sleep stages by the SleepyLand classifier for each 30-second epoch are overlaid with black markers. Vertical lines indicate epochs used for training the intracortical classifier, consisting of all epochs labeled as N2 and an equal number of randomly selected epochs labeled as awake.

recorded channels in Figure 3. The resulting temporal and spectral distribution of vigilance states is consistent with known physiological patterns of human sleep. Periods classified as sleep occurred predominantly during nighttime hours, while wakefulness dominated daytime intervals, with the expectation of an afternoon nap on the second day of recording.

Furthermore, the sequence of predicted sleep stages followed the expected cyclic structure of human sleep architecture [15], providing qualitative validation of the automated sleep staging obtained from EEG and EOG signals.

Intracortical awake-sleep classification: Training and testing the convolutional neural network on the 44-hour intracortical recording yielded a mean classification score of $96.1 \pm 1.6\%$ (Figure 4) of distinguishing wakefulness from N2 sleep, substantially exceeding chance level (practical chance level 48.7–51.3% [16]).

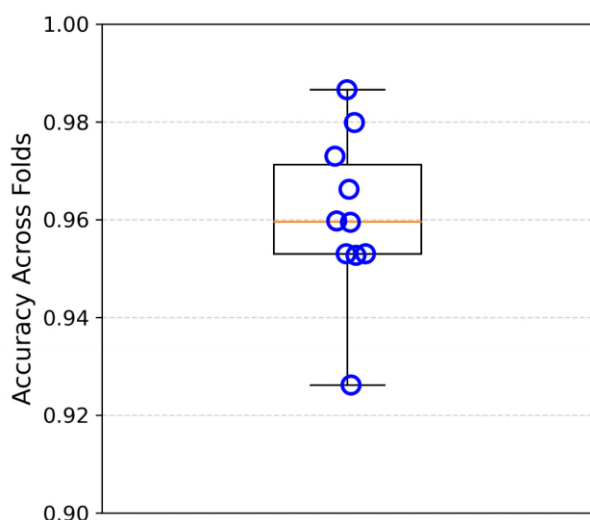


Figure 4: Classification accuracy across 10 cross-validation folds for intracortical awake-N2 sleep discrimination. Individual fold accuracies are overlaid on a boxplot showing the median and interquartile range (IQR); whiskers indicate the full range of observed values. Chance-level performance for this prediction is 48.7–51.3% [16].

Frequency band contribution analysis: Classification accuracies obtained when training the CNN on isolated frequency bands is shown in Figure 5. All individual frequency bands yielded significantly lower accuracy compared with the full broadband signal (Bonferroni corrected for number of bands, $p < 0.05$, Wilcoxon signed-rank test), indicating that no single frequency range alone was sufficient to achieve optimal discrimination between wakefulness and sleep. Additionally, only spike band power did not differ significantly from the practical chance-level range of 48.7–51.3% ($p = 0.77$ vs lower bound, $p = 0.49$ vs upper bound, Wilcoxon signed-rank test).

Application to independent BCI spelling sessions: The trained CNN was next applied to six independent BCI spelling sessions following the 44-hour recording period. For each session, overlapping 30-second epochs were

extracted from the first 20 minutes of data, resulting in 1,172 epochs per session.

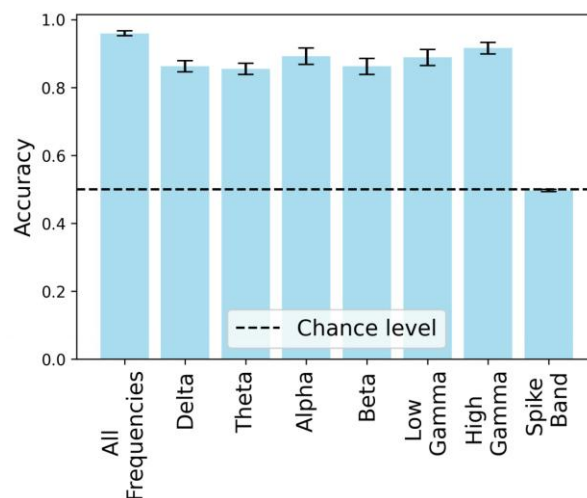


Figure 5: Classification accuracy for awake-sleep discrimination across cross-validation folds for individual frequency bands.

During the first two sessions (13:40 and 14:45), no epochs were classified as sleep by the model across any cross-validation fold. In the third session (16:02), only a single epoch was classified as sleep. The number of epochs predicted as sleep increased progressively in later sessions, reaching an average of 42 epochs in the final session at 22:20 (Figure 6).

Spelling performance declined in parallel with this increase in sleep-classified epochs. While 100% of words produced during the session at 20:37 were comprehensible, only 75% of produced words were comprehensible in the session at 21:51. In the final session, none of the produced words were comprehensible. This pattern suggests a relationship between reduced vigilance and impaired communication performance during prolonged BCI use.

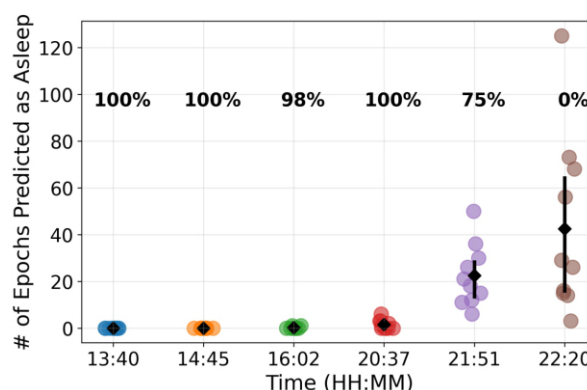


Figure 6: Number of epochs (out of 1,172 per session) classified as sleep by the trained CNN for each of six independent BCI spelling sessions recorded on the day following the 44-hour dataset. For each session, spelling performance is indicated as the percentage of comprehensible words produced during free spelling.

DISCUSSION

This study demonstrates the feasibility of detecting sleep and wakefulness from spatially focused, chronic neural signal recordings in a real-world BCI user, providing an important proof of concept for vigilance-aware autonomous BCI systems. Reliable identification of periods in which the user is asleep or drowsy is critical for reducing unintended activations, improving safety, and supporting continuous home use of communication BCIs. While the present findings are encouraging, substantial future work is required to extend and validate these findings.

The utilized reference sleep stage labels were derived from automated sleep staging of EEG and EOG signals without manual expert verification. While this approach enables scalable annotation, it may introduce some degree of label noise. Incorporating human scoring on a subset of the data could provide an estimate of labeling accuracy and help quantify the impact of potential misclassifications on model performance.

In this initial investigation, classification was restricted to discriminating between wakefulness and N2 sleep. We observed an increase in the number of epochs classified as sleep later in the day, which coincided with a decline in spelling performance. This binary framework could be expanded to include all sleep stages, enabling a more detailed characterization of vigilance state and potentially supporting adaptive BCI behavior across different levels of arousal. An alternative and complementary approach would be to estimate a continuous “drowsiness” or vigilance score rather than discrete sleep stages. Even during wakefulness, vigilance fluctuates considerably, and intermediate states of reduced alertness may already impair BCI performance. Such states may be detectable, as prior work has shown that increased low-frequency neural activity, including theta oscillations, is associated with fatigue and diminished task performance [17].

In the present study, non-invasive EEG and EOG recordings were used to generate reference sleep stage labels. While this approach enabled robust training and validation of the intracortical classifier, an important next step will be to identify sleep and vigilance signatures directly from intracortical data alone. Achieving reliable classification without simultaneous scalp recordings would simplify system design and facilitate long-term deployment in home environments. A related open question concerns the temporal stability of such classifiers: it remains unclear whether a model trained during a single calibration session would remain valid over weeks or months, or whether periodic recalibration using multimodal recordings would be required. Understanding this tradeoff will be essential for determining practical protocols for long-term clinical use.

A further limitation of this study is that it was conducted in a single participant. Future work must evaluate the generalizability of the proposed approach across individuals, electrode types, implantation sites, and BCI

configurations. In addition, the participant had late-stage ALS, a condition known to alter sleep architecture, including reduced sleep efficiency and decreased proportions of N3 and REM sleep [18]. These disease-related factors may influence both the neural signatures of sleep and the performance of classification models. Validation in additional participants, including those with different etiologies and sleep profiles, will therefore be necessary.

Analysis of frequency band contributions indicated that the highest classification performance was obtained using alpha and gamma-band activity. Alpha-band activity encompasses sleep spindles, which are characteristic of N2 sleep, while gamma oscillations are prominent during active wakefulness and diminish as sleep onset approaches [19]. These findings are therefore consistent with established electrophysiological markers of vigilance state. In contrast, spike band power alone was insufficient to reliably discriminate between sleep and wakefulness in this dataset. This result may partly reflect limitations of the current network architecture, which includes max-pooling operations that may not optimally capture very high-frequency temporal patterns. In addition, spike band activity is known to vary substantially across individual electrodes [20], and averaging across channels may have attenuated informative features present at the single-channel level. Alternative architectures and channel-wise classification approaches may therefore be better suited to exploit spike band information.

Overall, this work provides an important first step toward vigilance-aware intracortical BCIs capable of continuous operation in naturalistic settings. By demonstrating that sleep-wake information can be extracted from chronically implanted electrodes and that these predictions relate meaningfully to communication performance, the results motivate further research into adaptive BCI systems that respond dynamically to user state. Future studies incorporating larger cohorts, longer timescales, and real-time implementation will be crucial for translating these findings into clinically robust and autonomous BCI technologies.

CONCLUSION

In this work, we investigated whether sleep and wakefulness can be inferred from chronically implanted intracortical BCI signals during naturalistic use. Using simultaneous non-invasive sleep staging as reference labels, we trained a convolutional neural network to distinguish between wakefulness and N2 sleep with high accuracy over a 44-hour recording period. We further demonstrated that the classifier generalized to independent BCI spelling sessions and that its predictions were associated with changes in communication performance. Together, these results establish the feasibility of intracortical vigilance state decoding and provide a foundation for future studies exploring state-aware BCI operation.

ACKNOWLEDGEMENTS

This publication is part of the project Dutch Brain Interface Initiative (DBI2) with project number 024.005.022 of the research programme Gravitation, which is financed by the Dutch Ministry of Education, Culture, and Science (OCW) via the Dutch Research Council (NWO). In addition, this project is funded by the European Union's HORIZON-EIC-2021-PATHFINDER CHALLENGES program under grant agreement No 101070939 (INTRECOM) and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00198.

REFERENCES

- [1] Gaiduk M, Serrano Alarcón Á, Seepold R, Martínez Madrid N. Current status and prospects of automatic sleep stages scoring: Review. *Biomed. Eng. Lett.* 2023;13(3):247–272
- [2] Vansteensel MJ, Pels EGM, Bleichner MG, Branco MP, Denison T, Freudenburg ZV et al. Fully Implanted Brain–Computer Interface in a Locked-In Patient with ALS. *N. Engl. J. Med.* 2016;375(21):2060–2066
- [3] Card NS, Singer-Clark T, Peracha H, Iacobacci C, Hou X, Wairagkar M et al. Long-term independent use of an intracortical brain-computer interface for speech and cursor control. *bioRxiv* 2025.06.26.661591
- [4] Pahwa M, Kusner M, Hacker CD, Bundy DT, Weinberger KQ, Leuthardt EC. Optimizing the Detection of Wakeful and Sleep-Like States for Future Electrocorticographic Brain Computer Interface Applications. *PLOS ONE* 2015;10(11):e0142947
- [5] Von Ellenrieder N, Peter-Derex L, Gotman J, Frauscher B. SleepSEEG: automatic sleep scoring using intracranial EEG recordings only. *J. Neural Eng.* 2022;19(2):026057
- [6] Lamsam L, Gu B, Liang M, Sun G, Khan KJ, Sheth KN et al. The human claustrum tracks slow waves during sleep. *Nat. Commun.* 2024;15(1):8964
- [7] Leinders S, Arnoutse EJ, Brainco MP, Freudenburg ZV, Geukes SH, Schippers A et al. DO NOT LOSE SLEEP OVER IT: IMPLANTED BRAIN-COMPUTER INTERFACE FUNCTIONALITY DURING NIGHTTIME IN LATE-STAGE AMYOTROPHIC LATERAL SCLEROSIS. *medRxiv* 2024.10.11.24315027
- [8] Chaudhary U, Vlachos I, Zimmermann JB, Espinosa A, Tonin A, Jaramillo-Gonzalez A et al. Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training. *Nat. Commun.* 2022;13(1):1236
- [9] Rossi AD, Metaldi M, Bechny M, Filchenko I, van der Meer J, Schmidt MH et al. SLEEPYLAND: trust begins with fair evaluation of automatic sleep staging models. *Npj Digit. Med.* 2025;9(1):55
- [10] Olesen AN, Jennum PJ, Mignot E, and Sorensen HBD. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep* 2021;44(1):zsaa161
- [11] Phan H, Mikkelsen K, Chen OY, Koch P, Mertins A, De Vos M. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. *IEEE Trans. Biomed. Eng.* 2022;69(8):2456–2467
- [12] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *Npj Digit. Med.* 2021;4(1):72
- [13] Supratak A, Guo Y. TinySleepNet: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG, in *Proc. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 2020, 641–644
- [14] Paszke A, Gross S, and Massa F. PyTorch: an imperative style, high-performance deep learning library, in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, Canada, 2019, 8026–8037
- [15] Feinberg I, Floyd TC. Systematic Trends Across the Night in Human Sleep Cycles. *Psychophysiology* 1979;16(3):283–291
- [16] Mueller-Putz G, Scherer R, Brunner C, Leeb R, Pfurtscheller G. Better than random? A closer look on BCI results. *International Journal of Bioelectromagnetism* 2008;10:52–55.
- [17] Hung CS, Sarasso S, Ferrarelli F, Riedner B, Ghilardi MF, Cirelli C et al. Local Experience-Dependent Changes in the Wake EEG after Prolonged Wakefulness. *Sleep* 2013;36(1):59–72
- [18] Boentert M. Sleep disturbances in patients with amyotrophic lateral sclerosis: current perspectives. *Nat. Sci. Sleep* 2019;11:97–111
- [19] Adamantidis AR, Gutierrez Herrera C, Gent TC. Oscillating circuitries in the sleeping brain. *Nat. Rev. Neurosci.* 2019;20(12):746–762
- [20] Nason SR, Vaskov AK, Willsey MS, Welle EJ, An H, Vu PP, et al. A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces. *Nat. Biomed. Eng.* 2020;4(10):973–983

Proteus Effect in VR-Based BCI: How Avatar Age Influences Motor Imagery Behavior and EEG

Daniela Esteves¹, Marta Vasconcelos¹, Athanasios Vourvopoulos¹

¹Institute for Systems and Robotics (ISR-Lisboa), Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

E-mail: daniela.c.esteves@tecnico.ulisboa.pt

ABSTRACT: The Proteus Effect suggests that embodying avatars in virtual reality (VR) with specific characteristics can activate related stereotypes and alter behavior. While prior work has shown that avatar age influences motor imagery (MI) performance, its impact on MI-related EEG activity remains unclear. This pilot study investigated whether avatar age modulates both behavioral and EEG markers relevant for MI-based Brain-Computer Interfaces (BCIs). Sixteen healthy young adults performed a walking MI task in VR while embodied in either a young or elderly gender-matched avatar. Age-related beliefs, embodiment, mental chronometry, and Alpha event-related desynchronization/synchronization (ERD/ERS) were assessed. Participants endorsed age-related physical decline stereotypes and reported comparable embodiment across conditions. MI duration tended to be consistently longer when embodied in the elderly avatar. However, ERD/ERS analyses revealed modest and highly variable responses, with no consistent neural modulation across conditions.

INTRODUCTION

Brain-Computer Interfaces (BCIs) have received increasing attention in recent years in the field of post-stroke neurorehabilitation [1]. When combined with Motor Imagery (MI), without actual motor execution [2], BCIs can help stimulate damaged motor-related brain areas and promote neuroplasticity through repeated and vivid imagination of movements [3], helping post-stroke patients to recover lost motor function [4]. Providing visual feedback of the imagined movement further enhances this process, as action observation also activates motor-related brain regions. In this way, Virtual Reality (VR) has become an important tool to deliver such feedback in MI-BCI systems, since it creates accessible, immersive, and ecologically valid environments [5]. A key advantage of VR is its ability to induce a Sense of Embodiment (SoE), the experience of perceiving a virtual avatar as one's own body [6], which research suggests strengthens the potential for increased rehabilitation outcomes [7]. Beyond embodiment itself, avatar characteristics also shape users' cognition and behavior. In 2007, Yee, Bailenson [8] demonstrated that individuals adjust their behavior to match the traits and stereotypes associated

with their embodied avatars, a phenomenon termed as the Proteus Effect. In their experiments, participants who embodied taller or more attractive avatars subsequently behaved more confidently in social interactions. Since then, numerous studies have replicated and extended these findings, showing that avatar characteristics can influence social attitudes [9], cognitive performance [10], motor behavior and physical performance [11], emotional states [12], and creativity [13]. Despite increasing evidence, the mechanisms underlying the Proteus Effect remain insufficiently understood. Most studies have focused on behavioral and self-reported outcomes, while few have examined physiological correlates. Some evidence suggests that the occurrence of the Proteus Effect may influence physiological responses such as skin conductance[14], muscle activity[15], or heart rate[16]. However, direct investigation of neural activity associated with the Proteus Effect remains lacking.

Importantly, emerging evidence indicates that the Proteus Effect may also influence MI performance. For example, Beaudoin et al. [17] embodied participants in either a female young or elderly avatar while performing a walking MI task. The results showed that participants completed the task slower when embodying the elderly avatar, with this effect being stronger in those holding more negative beliefs about older adults' motor abilities. These findings suggest that avatar characteristics can modulate mentally represented physical activity, highlighting the potential relevance of the Proteus Effect in MI-based paradigms.

Given that MI-BCI performance relies on specific neural signatures, such as Alpha event-related desynchronization/synchronization (ERD/ERS) [2], any factor that can modulate MI-related brain activity may directly influence MI-BCI performance. Therefore, the present study aims to induce the Proteus Effect within an MI context and evaluate its impact on EEG activity relevant to MI-BCI training, specifically examining whether avatar age can modulate lower-limb MI-related brain patterns. Given the previous findings on the impact of Proteus Effect on MI performance [17] and evidence suggesting its influence on physiological signals [14–16], this study aims to replicate the findings of previous research and extend it further by investigating whether such effects are reflected in neural activity patterns.

MATERIALS AND METHODS

Participants: A total of 16 participants were included in this study (8 males (50%); 8 females (50%); age: 21.94 ± 2.77 years). 15 participants were right-handed and one was left-handed, without any known neurological conditions. All participants provided written informed consent in accordance with the 1964 Declaration of Helsinki. This study was approved by the Ethics Committee of Instituto Superior Técnico (CE-IST) with Ref. no. 22/2024.

Experimental design: The experimental design was based on the procedure described in Beaudoin et al. [17], with extensions to include EEG acquisition and additional MI tasks (Figure 2). Upon arrival, participants provided their informed consent and completed a set of questionnaires, specifically, demographic information, and the General Beliefs about Physical Functioning Instrument (GBPFI) [18]. The GBPFI assessed beliefs regarding the extent to which a typical person aged 20–90 years (in decades) possesses four physical abilities (strength, balance, gait speed, and flexibility). Responses were provided on a continuous vertical visual analogue scale ranging from 0 (lowest possible ability) to 1200 (highest possible ability), with participants being blind to the numerical values.

The VR headset was placed over the electrodes and remained on throughout the session. A 4-minute resting-state EEG recording was acquired in VR consisting of a black background with a central white fixation cross: 2 minutes eyes-open (fixating the cross), followed by 2 minutes eyes-closed.

Participants subsequently completed two experimental conditions in a randomized order (within-subject design): Elderly condition (participants embodied an elderly avatar) and Young condition (participants embodied a young avatar), with a break in between. In both, participants embodied a gender-matched avatar from a first-person perspective, while the alternative-age avatar (i.e., the non-embodied avatar) was seated beside the self-avatar and remained visible throughout the task. This ensured continuous visual exposure to both age representations and the activation of age-related stereotypes (Figure 1). Each condition comprised four runs (with a brief pause midway), and each run included three tasks:

1. Embodiment task: Participants remained seated and performed repetitive elbow flexion (approximately one movement per second, $\approx 45^\circ$ angular amplitude), while observing the self-avatar mirror their movements in real time. The task consisted of 10s baseline, 30s right-arm movement, 10s baseline, and 30s left-arm movement, with auditory cues marking movement onset and offset.

2. MI upper limbs task: Participants performed an MI task consisting of repeatedly imagining the same elbow flexion movement performed during the Embodiment task. Each trial began with a 10 s baseline (fixation cross between the avatar's arms), followed by a directional cue indicating the left or right arm, and 10 s of

the MI task. No avatar movement was present. Each run included 10 trials (5 per arm), presented in randomized order.

3. MI lower limbs task: Participants were instructed to perform a walking MI task. They first turned their head to the right toward a virtual corridor, where a chair appeared at one of five randomized distances. After the chair appeared, they imagined standing up, walking to the chair, and sitting down (MI task). Trials were initiated by pressing a button and closing their eyes; upon completing the imagined movement, participants pressed the button again and reopened their eyes. Button presses were used to record mental chronometry. Each run included five randomized trials (one per distance).

Thus, across the four runs per condition, participants completed 4 embodiment trials per arm, 20 MI upper limb trials per arm, and 20 MI lower limb trials (4 per chair distance).

At the end of each condition, participants completed the Embodiment Questionnaire (EQ) [19], consisting of 16 items rated on a 7-point Likert scale and estimated the self-avatar's age. At the end of the session, they were also immersed in gender-mismatched versions of both the elderly and young avatars and estimated their age to control for potential confounding effects of avatar characteristics (e.g., gender) on perceived age.

Experimental setup: The EEG setup used 32 active electrodes arranged according to the 10–20 system, with the reference electrode placed at FCz and ground electrode on the forehead. EEG signals were acquired at a sampling rate of 250 Hz using a LiveAmp 32 EEG amplifier (Brain Products GmbH, Gilching, Germany). Resting-state EEG signals were recorded using BrainVision Recorder (Brain Products GmbH, Gilching, Germany). To synchronize signals with experimental events the Lab Streaming Layer (LSL) was used.

The VR environment was adapted from Beaudoin et al. [17] (Figure 1), and developed in Unity game engine (Unity Technologies, USA). It consisted of a three-dimensional room with parquet flooring and wallpaper-covered walls, with two avatars seated side by side at a virtual table facing a mirror (a young adult and an elderly adult avatar). Depending on the experimental condition, participants embodied either the young or the elderly avatar (self-avatar), while the other remained as a static other-avatar to their left, visible directly and in the mirror. A corridor became visible when participants rotated their head approximately 90° to the right, where a virtual chair could appear at five equally spaced distances: Chair 1 (closest) through Chair 5 (farthest).

Participants experienced and controlled the avatars using an Oculus Rift CV1 headset (Oculus VR, Meta, Inc., USA), together with Oculus Touch controllers. Head rotation was mapped to the self-avatar's head orientation, while upper-limb movements were tracked via the controllers, with sagittal inclination corresponding to elbow flexion.

EEG signal analysis: EEG signals recorded during



Figure 1: Overview of the experimental setup (upper-left corner) and the virtual environment. The top-right image shows the virtual corridor located on the right side. The bottom images present the young and elderly avatars for the male (left) and female (right) versions of the scenario.

the closed-eyes resting state and the MI lower limbs task were processed post-hoc using the EEGLAB toolbox (v2023.1) in MATLAB R2022a.

The signals were first downsampled to 125 Hz and band-pass filtered (1 to 40 Hz). Channels were evaluated using EEGLAB's automatic rejection criteria, including flatline detection (>5s), low correlation with neighboring electrodes (Pearson correlation < 0.8; neighbors defined in Euclidean sensor space), line-noise levels (>4), and power spectral density inspection. Bad channels were manually removed and subsequently interpolated. Next, Independent Component Analysis (ICA) was performed, and artifactual components were rejected using ICLabel combined with visual inspection. Artifact Subspace Reconstruction (ASR) was applied to correct high-amplitude artifacts without removing data segments. Finally, signals were re-referenced to the common average. For the MI lower limbs task, data segments corresponding to the period from task onset to task completion (as determined by individual mental chronometry timings, e.g., approx. 4-12s) were extracted for each MI trial. Since this task was not time-locked, epoch duration varied across trials and fixed-length epoching was not applicable.

Event-Related Spectral Perturbations (ERSPs) were computed for the total duration of each MI lower limbs trial individually, using the closed-eyes resting-state recording (5–120 s) as baseline. ERSPs capture changes in spectral power relative to baseline across the time–frequency domain, from which ERD/ERS was derived as the percentage change in power relative to baseline[2]. ERD/ERS values were averaged over the Alpha frequency band (8–12 Hz) and across time (entire trial duration), obtaining a single-trial measure. This approach was preferred over full time-frequency pattern analysis, as the variable epoch lengths preclude meaningful temporal alignment across trials. Mean ERD/ERS values were then computed (i) across all chair distances combined and (ii) separately for each chair distance, resulting in six values per subject:

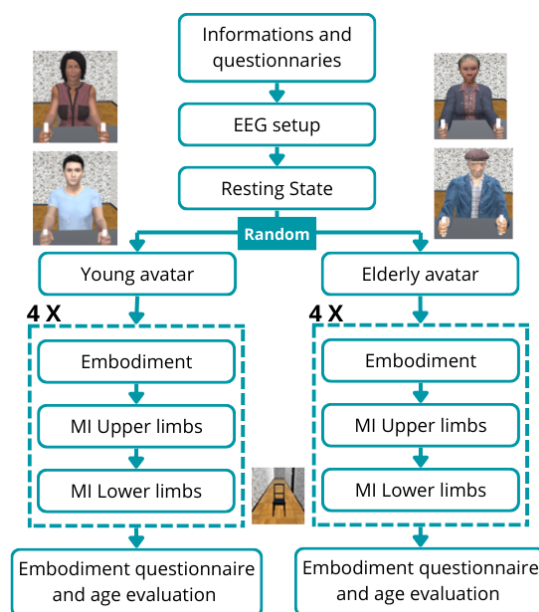


Figure 2: Schematic representation of the data collection procedure.

one representing the mean across all chairs and five corresponding to Chairs 1 through 5 individually, for both conditions (Young vs. Elderly).

Statistical analysis: Mental chronometry and ERD/ERS during lower-limb MI trials were compared between Young and Elderly conditions. ERD/ERS values were restricted to electrodes C3, C4, and Cz, and analyzed considering (i) all chair distances combined and (ii) each chair distance individually. These electrodes were selected because they overlie sensorimotor cortical areas, which are expected to be engaged during MI tasks[2]. Due to the reduced sample size, the non-parametric Wilcoxon signed-rank test was applied for all condition comparisons, with a significance level of 0.05. P-values were corrected using FDR multiple-comparison correction, applied separately for mental chronometry times and ERD/ERS values.

RESULTS

Perceived Aging and Embodiment: Participant beliefs about age-related changes, as reported by the GBPFI, show that physical strength, walking speed, balance, and flexibility decrease with age, indicating that the age-related stereotype necessary to elicit the Proteus Effect was indeed present (Figure 3). Regarding the SoE, as assessed by the Embodiment Questionnaire, participants reported comparable levels between conditions (Young condition: 4.31 ± 1.17 and Elderly condition: 4.42 ± 1.08 , in a 7-point Likert scale; $p = 0.50$). Additionally, participants perceived the elderly avatars as significantly older than the young avatars (male avatars: elderly 74.06 ± 7.81 years vs. young 19.00 ± 2.68 years, $p < 0.001$; female avatars: elderly 81.31 ± 6.31 years vs. young 34.06 ± 4.54 years, $p < 0.001$).

Mental chronometry: Mental chronometry results indi-

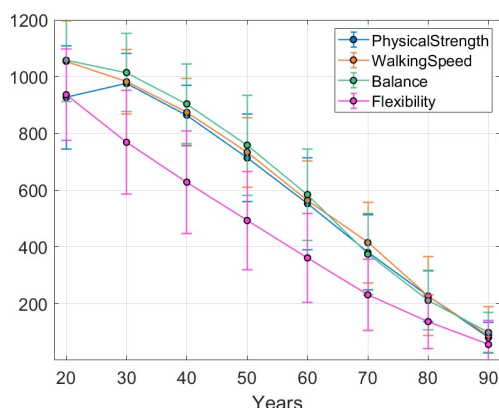


Figure 3: Mean ($\pm$ SD) ratings of the GBPFI questionnaire, on the 0–1200 scale, for perceived physical strength (blue), walking speed (orange), balance (green), and flexibility (pink) across age decades (20–90 years).

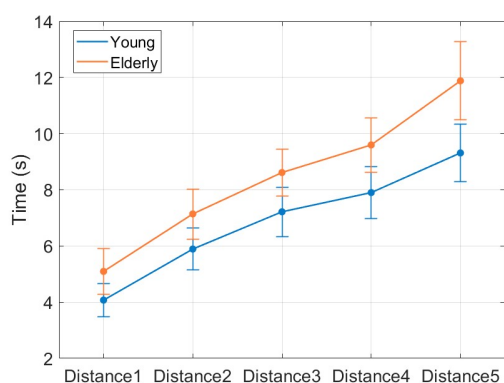


Figure 4: Mean ($\pm$ SD) mental chronometry time (in seconds) for the MI lower limb task across chair distances (1 to 5) in the Young (blue) and Elderly (orange) conditions.

cate that the time required to perform the MI lower limb task increased with the virtual distance of the chair, and participants consistently tended to take longer to perform the MI task when embodied in the elderly avatar compared to the young avatar (Figure 4). Although no significant differences between conditions were observed, a trend towards a main increase in MI duration in the Elderly condition was observed for chair distance 1 (Young: 4.07 ± 2.36 s; Elderly: 5.10 ± 3.24 s; $p = 0.08$), chair distance 2 (Young: 5.89 ± 2.98 s; Elderly: 7.14 ± 3.56 s; $p = 0.06$), chair distance 3 (Young: 7.22 ± 3.51 s; Elderly: 8.62 ± 3.32 s; $p = 0.06$), and chair distance 4 (Young: 7.90 ± 3.71 s; Elderly: 9.60 ± 3.87 s; $p = 0.06$). Only for chair distance 5 (Young: 9.31 ± 4.06 s; Elderly: 11.88 ± 5.56 s; $p = 0.15$) this trend was not as strongly observed, despite the Elderly condition still showing considerably longer MI times compared to the Young condition.

ERD/ERS Patterns During Lower-Limb MI: During lower-limb MI, ERD was induced over central sensorimotor regions as expected (Figure 5). However, this pattern was not consistently reflected at the electrodes of in-

terest (C3, Cz, C4), and no consistent ERD/ERS pattern emerged when considering all chair distances. No significant difference between conditions was observed at any electrode (Table 1), and both conditions showed ERS rather than the expected MI-induced ERD, with this effect being more pronounced in the Elderly condition.

A chair-by-chair analysis yielded similar results. The C4 electrode showed ERD for chair distance 1, 3, 4, and 5 in the Young condition, and for chair distance 1 in the Elderly condition. However, these effects were small and neither condition demonstrated robust or consistent ERD. Still, significantly lower values were observed in the Young condition compared to the Elderly condition for chair distance 4 across all electrodes of interest. No significant differences were found for the remaining chairs (Table 1), although values were generally lower in the Young condition.

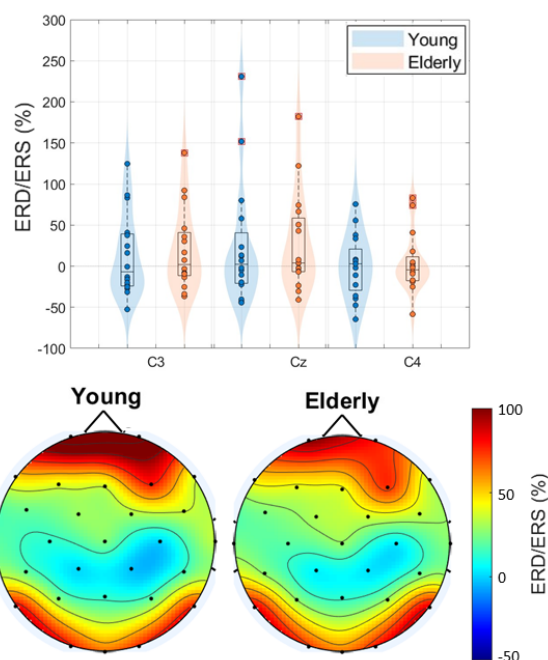


Figure 5: Grand-average ERD/ERS for the Young (blue) and Elderly (orange) conditions during the MI lower limbs task. The top panel shows boxplots of ERD/ERS at C3, Cz and C4, and the bottom panel presents the topographical map of mean changes, where negative values (blue) indicate ERD and positive values (red) indicate ERS.

DISCUSSION

This study aimed to extend the behavioral impact of the Proteus Effect on MI performance reported by Beaudoin et al. [17] and to examine whether it also modulates neural activity during lower-limb MI, specifically in Alpha ERD.

Participants exhibited the expected age-related stereotypes (Figure 3), reporting that physical abilities decline with age. This was essential, as the Proteus Effect relies on the activation of internalized stereotypes asso-

Table 1: Mean ERD/ERS (%) across all chair distances in the MI lower limbs task, and p-values from the Wilcoxon signed-rank test comparing Young vs. Elderly self-avatar conditions at C3, Cz and C4. * denotes statistically significant differences.

	All Chairs			Chairs 1			Chairs 2			Chairs 3		
	Young	Elderly	p-value	Young	Elderly	p-value	Young	Elderly	p-value	Young	Elderly	p-value
C3	12.72	19.40	0.06	4.60	13.07	0.71	20.81	9.33	0.69	14.66	21.10	0.41
Cz	24.36	27.52	0.47	27.50	23.09	0.71	32.06	24.14	0.80	22.42	25.27	0.41
C4	0.36	3.81	0.47	-2.76	-7.11	0.72	7.76	5.02	0.80	-0.96	4.46	0.41
	Chairs 4			Chairs 5								
	Young	Elderly	p-value	Young	Elderly	p-value						
C3	8.09	29.50	0.03*	15.43	23.99	0.12						
Cz	18.71	33.54	0.03*	21.13	31.55	0.12						
C4	-0.69	10.17	0.03*	-1.54	6.48	0.12						

ciated with avatar characteristics [8]. Embodiment ratings were moderate (approximately 4 on a 7-point scale) and comparable across conditions, indicating that participants experienced a similar SoE in both elderly and young avatars. Moreover, these levels were consistent with those reported by Beaudoin et al. [17]. Although embodiment is inherently subjective, these results suggest that participants were sufficiently embodied in the avatars, another critical factor for eliciting the Proteus Effect[8]. Together, these findings indicate that the necessary psychological conditions (age-related stereotypes, accurate age perception, and comparable embodiment) were present.

Behaviorally, the effect was reflected in the mental chronometry task (Figure 4), supporting the successful induction of the Proteus Effect. Imagined walking time increased with chair distance, demonstrating temporal congruency between imagined duration and implied physical effort. Importantly, participants consistently tended to take longer to complete the MI task in Elderly condition compared to the Young condition across all distances, even though no statistical significance was observed. This lack of significance may be attributed to the small sample size, inter-individual variability, and the relatively short task duration (4-12 seconds) which may not be enough to see differences. Moreover, these results closely replicate the pattern reported by Beaudoin et al. [17], supporting a behavioral manifestation of the Proteus Effect and suggesting that avatar age influenced the temporal representation of imagined movement.

In contrast to the behavioral findings, the EEG results were less conclusive. When considering all chair distances together, ERD values were generally modest and highly variable across participants. In several cases, Alpha-ERS rather than ERD was observed, and no consistent pattern emerged for any condition. Even in the chair-by-chair analysis, where subtle fluctuations were present, no robust or systematic ERD modulation was detected. Several factors may explain the absence of clear ERD modulation. First, ERD/ERS was computed using a closed-eyes baseline acquired at the beginning of the session and potential baseline drifts may have reduced sensitivity. Second, the MI task was performed without visual feedback, which typically strengthens ERD responses [20], possibly attenuating sensorimotor engagement. Additionally, lower-limb MI generally elicits weaker and more centrally localized ERD due to its deeper cor-

tical representation [21], making detection with scalp EEG more challenging, particularly in untrained participants, as in the present study. Finally, substantial inter-individual variability (Figure 5) may have masked subtle condition-related effects.

Nevertheless, Alpha power in the Young condition generally showed lower ERS compared to the Elderly condition for chair distance 1, 2, 4 and 5, as well as for all chairs together (Table 1), even reaching statistical significance for chair distance 4. This finding may suggest a potential interaction between avatar age representation and ERD/ERS patterns. Moreover, the fact that only chair distance 3 (Table 1) present a different trend between conditions may also indicate a possible influence of distance. However, given the absence of consistent ERD patterns, interpretation remains speculative.

Overall, under the present experimental conditions, the Proteus Effect did not produce a clear modulation of ERD during walking MI. While behavioral effects were robust, neural differences were subtle and inconsistent, suggesting that avatar-induced modulation may primarily influence cognitive aspects of MI rather than strongly altering brain patterns. Importantly, these findings should be interpreted cautiously, as they are preliminary and subject to several limitations. The sample size was relatively small, participants were mostly inexperienced in MI tasks, and the sample included only young adults. Additionally, the absence of visual feedback during MI, the button press used to initiate mental chronometry, and methodological factors such as baseline selection for ERD computation may have influenced neural activity and reduced sensitivity to task-related effects.

Future work will focus on increasing the sample size to improve statistical power and sensitivity to subtle neural effects, as well as expanding EEG analyses during lower-limb MI in both time and frequency domains. Exploration of potential correlations between ERD, mental chronometry, and questionnaire measures (e.g., GBPFI) will also be conducted, as well as explore possible effects of imagined walking distances. Lastly, examining EEG activity during embodiment and MI upper limbs tasks may provide further insight into possible neural modulation associated with the Proteus Effect.

CONCLUSION

This study extended and confirmed prior findings on the

behavioral impact of the Proteus Effect on MI and examined whether it also modulates neural activity during a walking MI task. Clarifying how avatar-induced effects influence MI-related neural signatures, such as ERD, is relevant for optimizing visual feedback in MI-BCI systems. Behavioral results were consistent with the Proteus Effect, as participants required longer imagined walking times when embodied in an elderly avatar compared to a young avatar. However, this modulation did not translate into robust or consistent ERD/ERS differences. Although subtle trends toward lower Alpha synchronization in the Young condition were observed, neural effects remained modest and highly variable. These findings suggest that avatar appearance can influence behavioral representations of movement without necessarily inducing measurable changes in MI-related EEG patterns. Future studies with larger samples, the inclusion of upper-limb MI, and extended EEG analyses are needed to determine under which conditions embodiment-related factors modulate neural signatures relevant for MI-BCI applications.

ACKNOWLEDGMENTS

This research was supported by the Fundação para a Ciência e Tecnologia (FCT) through the LARSyS grant (doi: 10.54499/LA/P/0083/2020, 10.54499/UIBP/50009/2020, and 10.54499/UIDB/50009/2020), and the NOISyS project (doi: 10.54499/2022.02283.PTDC).

REFERENCES

- [1] Cervera MA et al. Brain-computer interfaces for post-stroke motor rehabilitation: A meta-analysis. *Annals of Clinical and Translational Neurology*. 2018;5(5):651–663.
- [2] Pfurtscheller G, Lopes da Silva F. Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clin Neurophysiol*. 1999.
- [3] Vourvopoulos A, Jorge C, Abreu R, Figueiredo P, Fernandes JC, Badia S Bermúdez i. Efficacy and brain imaging correlates of an immersive motor imagery bci-driven vr system for upper limb motor rehabilitation: A clinical case report. *Frontiers in human neuroscience*. 2019;13:244.
- [4] Mane R, Ang KK, Guan C. Brain-computer interface for stroke rehabilitation. *Handbook of neuroengineering*. 2023:1285–1315.
- [5] Lotte F et al. Combining bci with virtual reality: Towards new applications and improved bci. In: *Towards practical brain-computer interfaces: Bridging the gap from research to real-world applications*. Springer, 2012, 197–220.
- [6] Kiltner K, Groten R, Slater M. The Sense of Embodiment in Virtual Reality. *Presence Teleoperators & Virtual Environments*. 2012;21.
- [7] Fluet GG, Roy D, Llorens R, Badia S Bermúdez i, Deutsch JE. Basis and clinical evidence of virtual reality-based rehabilitation of sensorimotor impairments after stroke. In: *Neurorehabilitation technology*. Springer, 2022, 429–466.
- [8] Yee N, Bailenson J. The proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research*. 2007;33(3):271–290.
- [9] R R et al. Examining the proteus effect on misogynistic behavior induced by a sports mascot avatar in virtual reality. *Scientific Reports*. 2024;14.
- [10] Banakou D, Kishore S, Slater M. Virtually being einstein results in an improvement in cognitive task performance and a decrease in age bias. *Frontiers in Psychology*. 2018;Volume 9 - 2018.
- [11] Kocur M, Kloss M, Schwind V, Wolff C, Henze N. Flexing muscles in virtual reality: Effects of avatars' muscular appearance on physical performance. In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Association for Computing Machinery: Virtual Event, Canada, 2020, 193–205.
- [12] Pandita S. Affective embodiment: The effect of avatar appearance and posture representation on emotions in vr. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2020, 539–540.
- [13] Guegan J, Buisine S, Mantelet F, Maranzana N, Segonds F. Avatar-mediated creativity: When embodying inventors makes engineers more creative. *Computers in Human Behavior*. 2016;61:165–175.
- [14] Frisanco A, Schepisi M, Tieri G, Aglioti SM. Embodying the avatar of an omnipotent agent modulates the perception of one's own abilities and enhances feelings of invulnerability. *Scientific Reports*. 2022;12.
- [15] Kurata T, Kanoh M. Two-dimensional display-based the proteus effect on muscle activity level. *Nov*. 2024, 1–6.
- [16] Kocur M, Kloss M, Schaufler C, Schwind V, Henze N. Investigating the impact of customized avatars and the proteus effect during physical exercise in virtual reality. *Apr*. 2025.
- [17] Beaudoin M, Barra J, Dupraz L, Mollier-Sabet P, Guerraz M. The impact of embodying an “elderly” body avatar on motor imagery. *Experimental Brain Research*. 2020;238.
- [18] Lineweaver TT, Kugler J, Rabellino A, Stephan Y. Beliefs about age-related changes in physical functioning across the adult life span and their relationship with physical activity levels of older adults. *Aging, Neuropsychology, and Cognition*. 2018;25(4):613–631.
- [19] Peck TC, Gonzalez-Franco M. Avatar Embodiment. A Standardized Questionnaire. *Frontiers in Virtual Reality*. 2021;1.
- [20] Choi JW, Kim BH, Huh S, Jo S. Observing Actions Through Immersive Virtual Reality Enhances Motor Imagery Training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2020;28(7):1614–1622.
- [21] Luft AR et al. Comparing brain activation associated with isolated upper and lower limb movement across corresponding joints. *Hum Brain Mapp*. 2002;17.

Deep learning based generalizable articulatory feature extraction for speech-BCI applications

Ruoling Wu¹, Julia Berezutskaya¹, Zachary V. Freudenburg¹, Nick F. Ramsey^{1,2}

¹ University Medical Center Utrecht Brain Center, Utrecht, The Netherlands

² Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, The Netherlands

E-mail: nick.ramsey@donders.ru.nl

ABSTRACT: *Objective:* Speech brain-computer interfaces (BCIs) often rely on paired neural-speech data, which is challenging to acquire from individuals with vocal tract paralysis. We aim to bridge this gap by proposing an across-subject framework that synthesizes intelligible speech using generalizable articulatory features extracted from an able-bodied cohort. *Approach:* We develop a subject-independent articulatory-to-speech model to extract generalizable articulatory features from real-time MRI (rtMRI) videos of a healthy cohort and use these features to synthesize text. *Main Results:* Our framework successfully synthesized text from the extracted generalizable articulatory features by achieving a Phoneme Error Rate (PER) of 18.9 % on unseen subjects. The results confirm that the extracted generalizable features remain robust and decodable across different subjects. *Significance:* By reducing the dependence on subject-specific articulatory data, our work offers a viable pathway for developing generalizable speech BCIs that could be used by those who can no longer articulate.

INTRODUCTION

The loss of speech can substantially compromise an individual's quality of life as it creates profound barriers to social interaction [1]. Medical conditions such as amyotrophic lateral sclerosis (ALS) or brainstem stroke can cause motor impairments that lead to the inability to control vocal tract muscles to produce speech. While recent advances in brain-computer interfaces (BCIs) show a promising way to restore speech, most current applications rely on paired neural-articulatory or neural-acoustic data from the same individual, which is challenging to acquire from individuals who cannot articulate or vocalize.

To circumvent this challenge, a viable solution is mapping articulatory features shared across a healthy cohort to the neural data of the target users [2], since motor representations remain preserved years after paralysis [3, 4]. Therefore, it is possible to map the motor representations of healthy people to neural data of people with paralysis. This allows for the decoding of intended speech for users with severe paralysis.

Our previous work has demonstrated the possibility of reconstructed generalizable articulatory features extracted from neural data of a different group of participants [2]. However, we did not demonstrate the linguistic effectiveness of the generalizable articulatory features. In this study, to investigate the generalizability and linguistic effectiveness of the extracted features, we propose an across-speaker speech BCI framework (Fig 1). First, generalizable articulatory features are extracted from articulatory data of healthy people. This step is the focus of this study. In the next step, the extracted articulatory features are fed into a neural feature extractor to reconstruct these features from the neural data of a target attempting to articulate the same speech output. The reconstructed articulatory features are then converted into speech via an articulatory-to-speech model.

While previous studies have explored various articulatory-to-speech synthesis approaches using modalities such as Electromagnetic Articulography (EMA) [5, 6], real-time MRI (rtMRI) [7-12], and Electromyography (EMG) [13], existing models are not able to achieve a balance between feature generalizability and robust speech synthesis performance.

To strike the balance, we introduce an across-subject articulatory-to-speech deep learning model to extract generalizable articulatory features from real-time sagittal MRI. Using data from multiple healthy individuals producing the same set of sentences, the model transforms them into text.

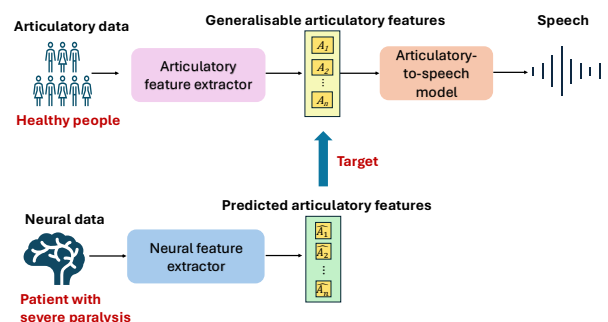


Figure 1: The across-speaker speech BCI framework. Articulatory data from healthy people is firstly used to

train an articulatory feature extractor. Then, the extracted generalizable articulatory features then serve as training targets for a neural feature extractor, which learns to map the neural data of patients with severe paralysis to the generalizable articulatory features. Finally, an articulatory-to-speech model synthesizes speech from the articulatory features predicted from neural data. We here report on the components on the upper panel.

MATERIALS AND METHODS

Across-speaker articulatory feature extraction model:

Our proposed model is designed with dual objectives: (1) to extract generalizable articulatory features from different subjects' rtMRI videos (2) to facilitate high-quality text synthesis from these generalizable articulatory features. Fig 2 shows our proposed model architecture. To extract generalizable articulatory features, we used AV-HUBERT, which is known for its robust performance in visual speech recognition [14, 15]. The standard AV-HUBERT consists of four parts: independent video and audio front-end encoders, a shared transformer backbone, and a cluster prediction head. As our study only focuses on the video modality, we disabled the audio front-end encoder and replaced the cluster-prediction head with a fully-connected layer as the phoneme prediction head, as in a previous study [10]. To prevent overfitting on our relatively small dataset, we replaced the original transformer backbone with a smaller transformer consisting of six layers and eight heads. To remove the subject-specific information from the articulatory features, we integrated an adversarial subject classifier inspired by the Domain-Adversarial Neural Networks (DANN) architecture [16]. The subject classifier is connected to the AV-HUBERT encoder via a Gradient Reverse Layer (GRL) [17]. During backpropagation, the GRL flips the sign of the gradients, forcing the AV-HUBERT encoder to learn articulatory features that minimize Connectionist Temporal Classification (CTC) loss [18] $\mathcal{L}_{ctc}$, while maximizing the subject classification loss $\mathcal{L}_s$, defined as:

$$\mathcal{L}_{ctc} = -\ln \sum_{\pi \in \mathcal{B}^{-1}(1)} \prod_{t=1}^T P(\pi_t | x, t)$$

$$\mathcal{L}_s = -\sum_{i=1}^K y_i \log(\hat{y}_i)$$

In the CTC loss $\mathcal{L}_{ctc}$, $\prod_{t=1}^T P(\pi_t | x, t)$ is the probability of a single specific predicted phoneme path across all time steps. In the subject classification loss $\mathcal{L}_s$, K is the number of subjects, y_i is the ground truth subject label and $\hat{y}_i$ is the predicted probability for subject label i . To obtain robust subject-invariant articulatory features, we employed a multi-stage training strategy. In the warm-up stage, we trained the AV-HUBERT encoder and the phoneme predictor together to converge only using $\mathcal{L}_{ctc}$ to ensure the AV-HUBERT encoder can extract linguistically effective subject-dependent articulatory features. Next, without influencing the

encoder and the phoneme prediction head, we froze their weights and trained only the subject classifier with $\mathcal{L}_s$ until convergence. After that, we unfroze all layers and trained the model with a combined loss function $\mathcal{L}_{combine}$, defined as

$$\mathcal{L}_{combine} = \mathcal{L}_{ctc} + \alpha \mathcal{L}_s$$

we regulated GRL using the “soft-start” strategy by gradually increasing the GRL’s scaling factor α from 0 to 1 via a sigmoid-based schedule. We used AdamW [19] as the optimizer with a learning rate that peaks at 0.001 after 30% updates, following the optimization strategy used [10].

Dataset and preprocessing: We used rtMRI data in the publicly available USC-TIMIT dataset [20]. The dataset consists of rtMRI videos recorded from ten American English speakers (five males M1-M5, five females F1-F5), each reading the same 460 sentences. The video resolution is 68 x 68 pixels, and the frame rate is 23.18 frame-per-second (fps). To prepare for the sentence-level analysis, we used the transcription API Whipex [21] to generate precise time-aligned transcriptions. Using these timestamps, each video recording was segmented into sentence-level clips. Following the sentence segmentation, a data preprocessing pipeline (shown in Fig 3) was run on the sentence-level clips: first, to isolate the vocal tract from the background, we defined the Region of Interest (ROI) for each speaker. Followed the pipeline in the previous work [22], we calculated a pixel variance heat map per speaker and identified the outermost borders (top, left, right, bottom) where pixel variance exceeded the mean pixel variance. Each frame was then cropped from its original size of 68 x 68 pixels to 47 x 47 pixels by the borders. After cropping, we rescaled pixel values to the range of a [0, 1] range and applied z-score normalization per clip to reduce flickering. After that we resized the video frame to 96 x 96 pixels and upsampled the temporal resolution from 23.18 fps to 25 fps to align with the model’s input requirements [15].

Evaluation metrics: To assess our dual objectives, we use a two-fold evaluation strategy focusing on linguistic effectiveness and generalizability of the extracted features. First, to evaluate the linguistic effectiveness, we quantify the phoneme decoding performance. We implemented a CTC decoder with a pretrained 4-gram language model [15] to decode text from the generalizable articulatory features using Phoneme Error Rate (PER) as the evaluation metric. PER measures the distance between the decoded and the ground-truth phoneme sequence. It is defined as,

$$PER = (S + D + I) / N,$$

where S, D, I represent the total number of substitutions, deletions, and insertions required to align the two sequences, respectively, and N represents the total number of phonemes in the ground-truth phoneme sequence. Second, to quantify feature generalizability, we trained a k-nearest neighbors (KNN) classifier on the extracted articulatory features to predict subject labels. The model is cross-validated across all features using a 10-fold cross-validation scheme, providing a

measurement of how effectively subject-specific information was removed across the entire cohort.

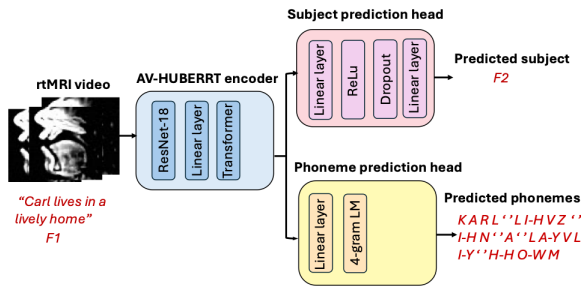


Figure 2: Model architecture. Our proposed model is composed of three main components: an AV-HUBERT Encoder that extracts articulatory features, followed by two parallel task heads: the phoneme prediction head and the subject prediction head. The phoneme prediction head consists of a linear layer and a 4-gram pretrained language model (LM). It converts the extracted features into phonemes to ensure the linguistic effectiveness, while the subject prediction head attempts to identify the speaker's identity.

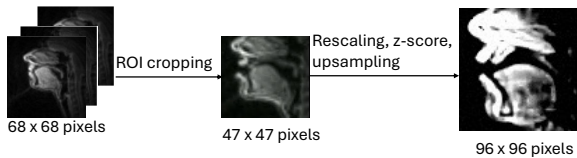


Figure 3: rtMRI preprocessing pipeline. Raw rtMRI videos were cropped from 68 x 68 pixels to 47 x 47 pixels ROIs. Next, pixels were rescaled to a range of [0-1], normalized by z-score. Finally, the frames are spatially upsampled to 96 x 96 pixels and temporally upsampled to 25 fps to meet the input requirement of AVHUBERT.

RESULTS

Feature generalizability across subjects: We evaluated the subject-invariance of features extracted from both models by using a KNN classifier to assess their ability to classify subject labels. As shown in Figure 4, the mean subject classification accuracy for features from the subject-variant AV-HUBERT model is 0.98. In contrast, the accuracy for the subject-invariant model is significantly lower at 0.51 ($p < 0.05$, paired t-test). This means that the subject-invariant model effectively lost the ability to differentiate individual subjects, and thus gained generalizability.

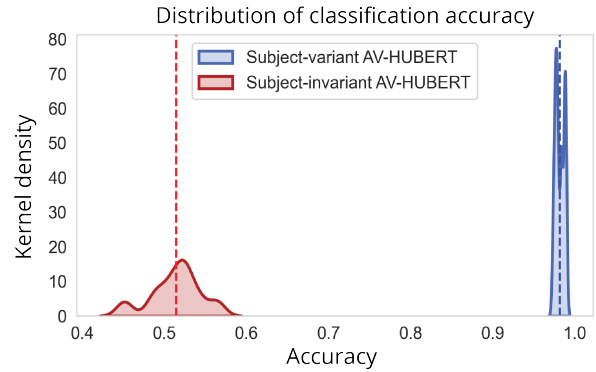


Figure 4: Distribution of subject-label classification accuracy. Curves represent the Kernel Density Estimation (KDE) of classification accuracies across subjects for both models. The subject-variant AV-HUBERT model (blue) achieves a mean accuracy of 0.98, while the subject-invariant AV-HUBERT model (red) shows a significantly reduced mean accuracy of 0.51. Vertical dashed lines indicate the respective mean values for each distribution.

Feature effectiveness evaluation: As previously described, we evaluated linguistic effectiveness of articulatory features based on the phoneme decoding performance. Table 1 shows the PER on each unseen subject. The subject-invariant AV-HUBERT consistently outperformed the subject-variant, achieving a mean PER of 18.9 ± 1.5 , compared to 32.1 ± 1.9 for the subject-variant model.

Table 1: PER (%) on each unseen subject and their average

Subject	subject-invariant model	subject-variant model
F1	18.2	32.6
F2	17.7	31.1
F3	19.1	29.6
F4	20.6	32.5
F5	16.1	28.6
M1	21.2	35.5
M2	18.5	32.9
M3	19.6	33.5
M4	20.7	34.0
M5	17.4	30.9
AVG	18.9 ± 1.5	32.1 ± 1.9

DISCUSSION

In this work, we proposed an across-speaker MRI-to-speech framework that extracts generalizable articulatory features across subjects. Our framework outperformed the baseline model [9] with a phoneme error rate of 18.9% on the synthesized sentences, compared to 32.9 % PER of the subject-variant model. These results indicate that the extracted articulatory features are both generalizable and linguistic effective.

Regarding the improved performance of the subject-invariant model, we hypothesize that the linguistic

effectiveness benefits from the feature generalizability. Since the target outputs, i.e., the sentence text, are invariant across subjects, we speculate that forcing the model to ‘unlearn’ the subject-specific information allows it to focus on the articulatory information most relevant to the phoneme prediction. To explore this speculation, we need to further examine the model’s learned behavior through a saliency study [23].

To quantify the feature generalizability, we trained a KNN classifier to identify the subject labels from the extracted articulatory features. Our result shows that the classification accuracy for features extracted from the subject-invariant model is significantly lower than that of the subject-variant model. This suggests that the subject-invariant model effectively eliminated a large part of subject-specific information during the encoding process. However, the classification accuracy for the invariant model remains above the chance level (0.10), indicating that some residual subject information persists within the extracted features. Future work will focus on further eliminating the residual subject-specific information while preserving linguistic information for text decoding.

Taken together, our proposed model extracts articulatory features from rtMRI videos that are both generalizable and linguistically effective. Next, we will map these features to neural data from a different subject, allowing us to develop across-speaker speech BCIs.

CONCLUSION

We demonstrated that our proposed framework can extract generalizable articulatory features from which speech can be synthesized with a low phoneme error rate. Using these articulatory features offers great potential for developing generalizable speech BCIs and reduces the need for a paired articulatory-neural or articulatory acoustic data, which is challenging to obtain from people with severe paralysis.

ACKNOWLEDGEMENT

This work is supported by Dutch Brain Interface Initiative (DBI2), project number 024.005.022 of the research programmed Gravitation, which is financed by the Dutch Ministry of Education, Culture, and Science (OCW) via the Dutch Research Council (NWO).

REFERENCES

- [1] M. Walshe and N. Miller, "Living with acquired dysarthria: the speaker's perspective," *Disability and rehabilitation*, vol. 33, no. 3, pp. 195–203, 2011.
- [2] R. Wu, J. Berezutskaya, Z. V. Freudenburg, and N. F. Ramsey, "Across-speaker articulatory reconstruction from sensorimotor cortex for generalizable brain-computer interfaces," *bioRxiv*, p. 2025.10. 08.680888, 2025.

- [3] M. L. Bruurmijn, I. P. Pereboom, M. J. Vansteensel, M. A. Raemaekers, and N. F. Ramsey, "Preservation of hand movement representation in the sensorimotor areas of amputees," *Brain*, vol. 140, no. 12, pp. 3166–3178, 2017.
- [4] Z. V. Freudenburg *et al.*, "Sensorimotor ECoG signal features for BCI control: a comparison between people with locked-in syndrome and able-bodied controls," *Frontiers in neuroscience*, vol. 13, p. 1058, 2019.
- [5] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep speech synthesis from articulatory representations," *arXiv preprint arXiv:2209.06337*, 2022.
- [6] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces," *PLoS computational biology*, vol. 12, no. 11, p. e1005119, 2016.
- [7] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *Interspeech*, 2023, pp. 127–131.
- [8] T. G. Csapó, "Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract," *arXiv preprint arXiv:2008.00889*, 2020.
- [9] L. Pandey and A. Sabbir Arif, "Silent speech and emotion recognition from vocal tract shape dynamics in real-time MRI," in *ACM SIGGRAPH 2021 Posters*, 2021, pp. 1–2.
- [10] N. Shah, A. Kashyap, S. Karande, and V. Gandhi, "MRI2Speech: Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025: IEEE, pp. 1–5.
- [11] P. Wu *et al.*, "Deep speech synthesis from MRI-based articulatory representations," *arXiv preprint arXiv:2307.02471*, 2023.
- [12] Y. Yu, A. H. Shandiz, and L. Tóth, "Reconstructing speech from real-time articulatory MRI using neural vocoders," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021: IEEE, pp. 945–949.
- [13] K. Scheck *et al.*, "Cross-speaker training and adaptation for electromyography-to-speech conversion," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024: IEEE, pp. 1–4.
- [14] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
- [15] B. Shi, W.-N. Hsu, K. Lakhota, and A. Mohamed, "Learning audio-visual speech

- representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.
- [16] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.
 - [17] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, 2015: PMLR, pp. 1180–1189.
 - [18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
 - [19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
 - [20] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.
 - [21] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.
 - [22] E. Stolwijk, "Towards speech-based brain-computer interfaces: finding most distinguishable word articulations with autoencoders," 2023.
 - [23] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

Optimizing Bit-wise Linear Decoding for Imperceptible Code-Modulated Visual Evoked Potentials (I-c-VEP)

M. A. Fodor^{1,2}, M. Tangermann³, J. Thielen³, I. Volosyak^{1,2}

¹Faculty of Technology and Bionics, Rhine-Waal University of Applied Sciences, Kleve, Germany

²Department of Informatics and Data Science, Graduate School for Applied Research in North
Rhine-Westphalia, Bochum, Germany

³Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The
Netherlands

E-mail: Ivan.Volosyak@hochschule-rhein-waal.de

ABSTRACT: Visual evoked potential (VEP)-based brain-computer interfaces (BCIs) are a well-established paradigm in the field, yet they remain limited by a fundamental usability constraint: they rely on perceptible low-frequency flicker, which induces visual fatigue and discomfort. To address this limitation, we previously proposed modulating an imperceptible high-frequency (60 Hz) carrier based on a pseudo-random binary code sequence. The resulting imperceptible code-modulated VEP (I-c-VEP) paradigm substantially reduces perceived flicker compared to conventional c-VEP approaches, while preserving key advantages over steady-state VEP (SSVEP) systems, namely efficient target scalability and robustness to narrowband interference.

In this study, we explore whether simple, efficient linear decoding pipelines suffice for bit-wise I-c-VEP classification across 16 participants. By validating our pipelines on strictly independent, temporally held-out data, we demonstrated that approaches based on linear ridge regression and linear discriminant analysis can achieve over 97% mean trial accuracy. These findings provide a solid foundation for efficient and high-performance real-time implementations of the I-c-VEP paradigm.

INTRODUCTION

Brain-computer interfaces (BCIs) can establish a direct communication pathway between the brain and external devices, enabling vital assistance for individuals with severe motor impairments by translating brain activity into control commands [1].

BCIs using the visual evoked potential (VEP) are favored for intuitive gaze control, offering robust evoked responses that can be reliably captured with non-invasive electroencephalography (EEG). The two prevalent VEP-based encoding strategies are steady-state VEP (SSVEP), which encodes targets by triggering VEPs through flashes at constant frequencies [2], and code-modulated VEP (c-VEP), which encodes targets by triggering VEPs in a pseudo-random sequence [3]. While SSVEP remains popular, c-VEP is often favored as it encodes informa-

tion in the temporal structure of pseudo-random stimulation sequences rather than narrow spectral peaks. The use of unique sequences or circular shifts of a base sequence facilitates efficient target scalability, while also offering inherent robustness to narrowband interference [3].

Despite their performance, the broader adoption of VEP-based BCIs is constrained by a fundamental usability limitation: their reliance on high-contrast flashing as stimulation. Commonly, a black and white low-frequency (5–25 Hz) flicker is used to elicit robust neural responses. However, prolonged use induces significant visual fatigue, cognitive load, and discomfort [4], and the highly visible flicker can act as a distraction that can influence task performance [5].

To mitigate these adverse effects, one of the established avenues has been the adoption of high-frequency SSVEP paradigms (> 30 Hz) to reduce flicker perceptibility [4, 6]. Visual comfort generally increases with frequency until the critical flicker fusion (CFF) threshold is reached, at which point a temporally modulated light source is perceived as steady [6]. Stimulation above 50–60 Hz, termed rapid invisible frequency tagging (RIFT) [5], typically is no longer consciously perceived. Importantly, neurophysiological evidence demonstrates that the visual cortex continues to entrain to these frequencies even without conscious awareness [5, 6]. However, because the amplitude of the evoked response decreases with increasing stimulation frequency, the resulting signal-to-noise ratio (SNR) deteriorates at higher frequencies [5, 6]. Consequently, although RIFT offers a compelling route toward an imperceptible BCI stimulus, its practical adoption remains constrained by this fundamental SNR trade-off.

In a preceding study [7], we introduced the *imperceptible code-modulated VEP* (I-c-VEP) paradigm, in which a high-frequency (60 Hz) carrier used in RIFT is amplitude-modulated (“on” and “off”) according to a predefined pseudo-random code sequence. Specifically, during the “off” state, the stimulus is displayed at a constant mean luminance, while during the “on” state, it flickers at 60 Hz between minimum and maximum luminance, reducing perceived flashing sensation as the target blends

into a more steady, quasi-uniform gray.

This approach bridges concepts from c-VEP and SSVEP research to address a fundamental usability limitation of VEP-based BCIs: visual discomfort caused by perceptible flicker. Although bit transitions themselves still introduce perceptible artifacts, this initial implementation may already represent one of the c-VEP BCI paradigms with the lowest amount of perceptible flicker to date, while preserving the key advantages of c-VEP approaches, efficient target scalability and robustness to narrowband interference.

When using code-modulated stimuli, one could classify the responses to unique events in the stimuli (i.e., bit-level decoding) instead of directly discriminating among multiple target classes (i.e., trial-level decoding). While decoding at the event level through models such as ridge regression [8] or reconvolution [9] can offer significant flexibility in stimulus design and high data efficiency, its adoption has been limited compared to prevalent and highly effective trial-level decoding strategies [3].

In this study, we explore whether simple, efficient linear bit-wise decoding pipelines suffice for the I-c-VEP paradigm. Across 16 participants, we evaluate ridge regression [8], shrinkage-LDA [10], block-Toeplitz LDA [11], and Riemannian tangent space LDA [12, 13]. We hypothesize that within this selection, we can identify a pipeline that achieves high accuracy while remaining data-efficient, providing a robust candidate for future implementations.

MATERIALS AND METHODS

Participants: The study was conducted at the Brain-Computer Interface laboratory of the Rhine-Waal University of Applied Sciences, in accordance with the Declaration of Helsinki, following approval from the Medical Faculty Ethics Committee of the Duisburg-Essen University (24-11957-BO). The full experiment involved 20 participants and included phases related to a separate study. Sixteen participants (9 female, 7 male, mean age of 24.81 ± 3.17 years) completed the full protocol within the scheduled 1.5-hour and were therefore included in the present analysis.

Hardware and Software: We used an Acer Nitro XV252Q F LCD monitor (1920×1080 resolution) at 360 Hz refresh rate, positioned 80 cm from the participant. To ensure timing and luminance stability, vertical synchronization (V-Sync) was enabled, while adaptive sync technologies and dynamic contrast settings were disabled. The panel operated at a maximum brightness with a gamma of 2.2, and with the “Overdrive” setting set to “Normal” to avoid overshoot artifacts. Stimulus presentation was done by a Dell workstation (AMD Ryzen 9 7950X3D CPU, NVIDIA RTX 4070 Ti Super GPU).

Continuous EEG data were acquired at 1200 Hz using a g.USBamp biosignal amplifier (Guger Technologies, Schiedlberg, Austria). Passive Ag/AgCl sensors (Easy-Cap GmbH, Wörthsee, Germany) were arranged in a 16-

channel montage focused on the parieto-occipital cortex, adhering to the international 10–5 system: P7, P3, Pz, P4, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2, O9, Iz, and O10. The reference and ground electrodes were placed at Cz and AFz, respectively, and impedances were maintained below 5 k Ω throughout the session.

Stimulus: The I-c-VEP paradigm displayed four gray square targets (7.08×7.08 cm; 250 px) at 360 Hz arranged horizontally on a black background with a 5.66 cm (200 px) spacing. Each target had a 1.23 cm (40 px) numeric label in the center. Trials presented two full cycles of a 31-bit m-sequence (2×2.583 s). Stimulation was implemented by gating a 60 Hz carrier signal on and off according to the m-sequence $s = 1111101100111000011010100100010$ at a presentation rate of 12 Hz (30 frames per bit; ≈ 83.33 ms). A logical bit ‘1’ encoded the active 60 Hz flicker, realized by a square wave alternating 3 black and 3 white frames per cycle at the native 360 Hz refresh rate. Conversely, bit ‘0’ encoded the static (off) state, displaying a mean luminance of 0.5 in linear color space (sRGB (188, 188, 188)) to ensure isoluminance (gray appearance) with the time-averaged intensity of the flicker (on) state.

Experimental Protocol: The full experiment spanned approximately ≈ 90 min, including 20–30 min for the EEG setup and post-experiment procedures. For this study, two runs were conducted at the beginning and the end of the session to maximize their temporal distance. The first run was used for selecting and training the decoding models, and the second served as a hold-out test set. The interim period (50–60 min) was used for tasks of a separate study. This included a first sequential target selection task (≈ 6 min), another training (≈ 5 min), a second sequential target selection task (≈ 6 min), and approximately 10 min to collect subjective rating of various c-VEP stimuli. Each run contained 40 trials, during which participants sequentially fixated one of the four targets. A single trial comprised 5.16 s of continuous stimulation. Following each trial, a 2 s gaze-shift interval was provided to ensure participants had sufficient time to comfortably fixate the next target.

Preprocessing: For all models, the raw continuous EEG data ($f_{s_raw} = 1200$ Hz) were high-pass filtered at 1 Hz (second-order Butterworth) to remove baseline drift, and notch-filtered at 50 and 100 Hz to mitigate power line interference. To prevent aliasing prior to downsampling, a fourth-order Butterworth low-pass filter was applied at 120 Hz (80% of the target 150 Hz Nyquist limit). The data were then decimated by a factor of 4 to a working sampling rate of $f_s = 300$ Hz. Finally, 300 ms epochs time-locked to the bit-transition onsets were extracted to capture the corresponding VEP morphology.

While focusing on decoding approaches, we evaluate multiple frequency bands to explore which ranges yield better performance and to account for potential synergies between specific pipelines and spectral filters. Three categories of frequency bands were considered: (i) broadband c-VEP filters (1–40 Hz, 1–65 Hz, and 1–90 Hz), (ii)

low, medium and high frequency bands (1–15 Hz, 15–40 Hz, and 40–90 Hz), and (iii) I-c-VEP-specific bands (55–65 Hz, and 58–62 Hz and 10–65 Hz, 20–65 Hz). The latter bands were intended to include the 60 Hz carrier as well as the broadband c-VEP response. The exclusion of the alpha band was motivated by the hypothesis that the perceptually subtle I-c-VEP stimulus may provide a lower visual drive than traditional flicker; consequently, endogenous alpha activity might be less suppressed. All filters were applied causally (forward-pass only) via cascaded second-order sections (SOS).

For spatial filtering, xDAWN [14] (2, 4, or 6 components) was applied across all LDA models to maximize the signal-to-signal-plus-noise ratio of the evoked responses. The ridge pipeline used canonical correlation analysis (CCA; 2 or 4 components), a common spatial filter for c-VEP decoding [9], to maximize correlation with class-specific templates, as in EEG2Code [8]. To isolate the 60 Hz carrier, common spatial patterns (CSP) [15] was evaluated alongside xDAWN to extract filters that maximize variance differences between classes. From the resulting squared amplitudes, either the full time-series or the mean power within a 210–270 ms post-transition window (centered on the observed peak of class discriminability) was extracted. Both raw and log-transformed inputs were evaluated.

Decoding Models: Five bit-wise linear decoding models were evaluated. For each bit, a single EEG epoch time-locked to the bit transition was extracted and linearly projected to a scalar output, which was subsequently used for bit-level decoding.

The first model, *linear ridge regression (Ridge)*, is a proven decoding approach for c-VEP [8]. It applies L_2 regularization to mitigate overfitting, with the penalty parameter (α) optimized via cross-validation on the training set. While inspired by the EEG2Code framework, our implementation extracts the epochs with a one-bit stride to yield exactly one prediction per bit, ensuring direct comparability with the other evaluated linear models.

The second, *shrinkage LDA (sLDA)* with Ledoit-Wolf shrinkage regularization of the covariance matrix, systematically shrinks the empirical covariance matrix toward a structured target, mitigating the ill-conditioning inherent to high-dimensional EEG feature spaces and limited training data [10].

The third, *block-Toeplitz LDA (btLDA)* assumes that the covariance between two time points depends only on their temporal distance rather than absolute position, and that covariance decreases with temporal distance. This assumption allows to regularize the covariance matrix into a block-Toeplitz structure with drastically reduced free parameters. Applying this temporal regularization alongside standard shrinkage significantly improves performance, particularly in limited-data scenarios [11].

Fourth, *Riemannian tangent space LDA (tsLDA)* was implemented [12]. Standard spatial covariance estimation discards the time-locked temporal morphology critical for decoding VEPs. To map this temporal morphol-

ogy onto the Riemannian manifold, “super-epochs” were constructed by concatenating the raw single-trial EEG epochs with target-specific template responses. These templates were generated by calculating the grand average of the time-locked training epochs for each respective class (i.e., the mean evoked responses for bit ‘0’ and bit ‘1’). Following this, spatial covariance estimation was performed directly on the super-epochs, an approach proven successful for the c-VEP domain [13] to effectively embed the temporal phase information directly into the spatial covariance matrix. Resulting symmetric positive definite (SPD) matrices were projected into the tangent space using the Fréchet mean of the training data as a fixed reference, enabling robust classification via sLDA.

Finally, *feature-level (early) and score-level (late) fusion* strategies were assessed to integrate complementary temporal, spatial, and spectral information. Early-fusion concatenated features extracted across a filter bank into a unified vector, which was standardized using parameters estimated from the training set and classified by a single sLDA model. Late-fusion aggregated the outputs of independently trained base models. Raw decision scores (hyperplane distances) from each model were z-scored using the mean and standard deviation derived from the training set, averaged, and transformed via a logistic sigmoid function to yield a final posterior probability. This approach effectively fuses predictions in a straightforward way without requiring a separate meta-classifier.

Evaluation: Final pipeline configurations were selected in a staged, domain-informed way to avoid exhaustive permutation testing. To ensure zero data leakage, pipeline configuration and final evaluation were strictly decoupled. Pipelines were ranked and selected based on 5-fold cross-validation performance on the training run, using the normalized area under the learning curve (nAUC; computed with a 4-trial step size) as the primary metric to capture both accuracy and data efficiency.

The selected models from each approach were then evaluated once on the temporally held-out test data (second run, recorded ≈ 1 h later). Learning curves were computed using a fine-grained 1-trial step size to assess performance under both rapid (≈ 55 s; 8 trials) and full (≈ 4.75 min; 40 trials) training scenarios. Decoding curves were generated per-bit (83.33 ms steps) to track classification accuracy over the trial duration. We report the final peak accuracy ($\pm$ SD) and a “plateau point”—defined as the earliest evaluation time not significantly lower than the peak (one-sided Wilcoxon, $p > 0.05$). Differences between models were assessed via two-sided Wilcoxon signed-rank tests with step-down Holm-Bonferroni correction. Finally, simulated information transfer rate (ITR; see <https://bci-lab.hochschule-rhein-waal.de/en/itr.html>) was calculated to quantify system throughput.

RESULTS

The section first describes the staged selection of the

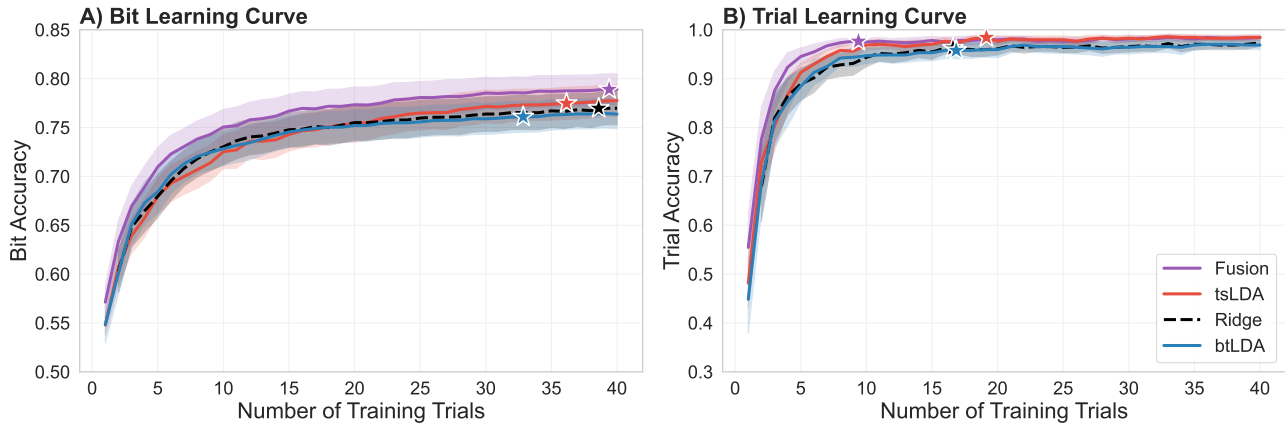


Figure 1: **Grand average offline learning curves across $N = 16$ participants.** (A) Bit-wise (binary) classification accuracy as a function of training trials. (B) Trial-level (4-class) decoding accuracy evaluated using the full 5.16 s trial duration. Solid lines represent the mean performance across subjects, while the shaded regions indicate ± 1 standard error of the mean (SEM). The star markers denote the plateau point for each curve, defined statistically via a one-sided Wilcoxon signed-rank test as the earliest time point not significantly worse ($p > 0.05$) than the peak accuracy.

pipelines and then reports and compares the performance of the selected, best implementation for each pipeline. To maintain clarity, the learning and decoding curves present only the best configurations for each approach: the *Ridge*, *btLDA*, *tsLDA*, and the *Late-fusion* LDA model, all evaluated on the held-out test set. Figure 1 illustrates the mean bit- and trial-level accuracies as a function of training data volume, while Figure 2 depicts trial decoding accuracies over evaluation time. Of the 62 bits comprising each trial, the final 4 were excluded to prevent classification of data from the post-stimulus interval. Consequently, a single 40-trial run yielded $40 \times 58 = 2,320$ epochs, corresponding to one epoch per evaluated bit.

Optimal Models:

Time-Domain Pipeline (*btLDA*): First, *sLDA* was evaluated with the broadband c-VEP and I-c-VEP-specific frequency bands using xDAWN with 4 spatial components. The broadband 20–65 Hz filter yielded the highest performance (CV-nAUC: 76.0%), significantly outperforming the second-best 10–65 Hz band (CV-nAUC: 74.8%; $W = 7.0$, $p = .001$). Consequently, using the 20–65 Hz band and xDAWN (4 components), we compared *sLDA* against *btLDA*. The *btLDA* model demonstrated significantly better data efficiency, achieving a higher CV-nAUC (76.5% vs. 76.0%; $W = 22.0$, $p = .016$). Going forward with *btLDA* and xDAWN, we evaluated the number of spatial components (2, 4, and 6). Extracting 2 components yielded the highest CV-nAUC (76.7%), though the difference compared to 4 components was not statistically significant ($p = .159$). Consequently, the final optimal *btLDA* model used a 20–65 Hz bandpass filter, xDAWN extracting 2 components.

Riemannian Pipeline (*tsLDA*): For the *tsLDA* pipeline, we utilized xDAWN with two components and evaluated three bandpass configurations: the previously established 20–65 Hz band, a broad 1–90 Hz band, and an early-fusion approach. The latter concatenated features from low (1–15 Hz), mid (15–40 Hz), and high (40–90 Hz)

bands prior to classification. This *Early-fusion* configuration significantly outperformed the 20–65 Hz band, yielding a CV-nAUC of 76.9% compared to 74.3% ($W = 1.0$, $p < .001$). Consequently, the early-fusion xDAWN implementation with two components was selected as the optimal *tsLDA* pipeline.

60 Hz Carrier Pipeline: To isolate the carrier (55–65 Hz), we compared spatial filters (xDAWN vs. CSP), component counts (2 vs. 4), and features (raw vs. log-power; binned vs. full time-series). xDAWN significantly outperformed CSP ($W = 0.0$, $p < .001$), and raw power proved superior to log-transformation. Temporal binning was used as it minimized dimensionality with negligible ($W = 66.0$, $p = .940$) CV-nAUC loss compared using the full time-series. Using binned raw power, 2-component xDAWN yielded a significantly higher CV-nAUC than 4-component (64.4% vs. 64.2%; $W = 18.0$, $p = .008$). This configuration achieved $81.4\% \pm 20.4\%$ trial accuracy with 40 training trials.

Linear Ridge Regression (*Ridge*): For the ridge regression model, evaluating the number of CCA components revealed that extracting 2 components significantly outperformed 4 components (76.9% vs. 76.5%; $W = 21.0$, $p = .013$).

Early and Late Fusion: The aim was to combine the best complementary frequency bands and model architectures. In the *Early-fusion* approach, concatenating the 20–65 Hz band with narrowband 60 Hz features improved upon the isolated pipelines, achieving a CV-nAUC of 77.4%. The *Late-fusion* approach, which aggregated the probabilistic outputs of the best-performing *btLDA*, *tsLDA*, and 60 Hz models, yielded the highest overall performance (CV-nAUC: 78.9%). This configuration significantly outperformed the next best model (*tsLDA*) ($W = 0.0$, $p < .001$), establishing it as the best performing model.

Table 1 provides the mean bit-wise classification accuracies, nAUC scores, and trial-level accuracies ($T = 40$) for

Table 1: **Offline classification performance using the full 40-trial training set.** Accuracies are reported as % ($\pm$ SD). Bit nAUC is the normalized area under the bit-wise learning curve.

Model	Bit Acc.	Bit nAUC	Trial Acc.
Ridge	77.0 $\pm$ 6.8	74.2	97.3 $\pm$ 3.0
btLDA	76.4 $\pm$ 6.1	73.6	96.9 $\pm$ 3.7
tsLDA	77.8 $\pm$ 6.5	73.9	98.4 $\pm$ 1.5
Early-fusion	77.8 $\pm$ 6.9	74.5	97.5 $\pm$ 3.7
Late-fusion	78.9 $\pm$ 6.4	75.9	98.4 $\pm$ 2.2

the models evaluated on the hold-out test set.

Learning: Notably, the *Late-fusion* model significantly outperformed all other models in both bit-wise nAUC ($p < .001$) and final bit accuracy ($p < .002$, Holm-Bonferroni corrected). Figure 1 demonstrates that all models achieved high trial-level accuracies given 40 training trials. While the *Late-fusion* model achieved the highest final trial accuracy (98.4% $\pm$ 2.2%), it did not differ significantly from the *tsLDA* ($p = .353$) or *btLDA* ($p = .096$) baselines at this terminal point after multiple comparison correction. However, the *Late-fusion* model demonstrated superior data efficiency, achieving a significantly higher trial-level nAUC than all baselines ($p < .01$, Holm-Bonferroni corrected).

We defined a performance plateau as the earliest point not significantly below peak performance (one-sided Wilcoxon signed-rank test, $p > 0.05$). The *Late-fusion* model converged the fastest, reaching 97.7% at just $T = 9$ trials. In contrast, other models required more data: *Early-fusion* plateaued at $T = 16$ (96.1%), *Ridge* and *btLDA* at $T = 17$ (96.2% and 95.8%), and *tsLDA* eventually plateauing at its peak accuracy of 98.4% at $T = 19$.

Decoding: When constrained to only 8 trials of training data (Figure 2A), *Late-fusion* achieved the highest performance, plateauing at $T = 3.42$ s (96.6%) with a final accuracy of 97.3% $\pm$ 3.5%. In this limited-data regime, *Late-fusion* significantly outperformed all other models across both decoding nAUC ($p < .01$) and final accuracy ($p < .05$, Holm-Bonferroni corrected). Under this same training constraint, the final accuracies of the other models were lower: 95.8% $\pm$ 4.6% for *tsLDA*, 94.2% $\pm$ 6.4% for *btLDA*, and 93.8% $\pm$ 9.9% for *Early Fusion*.

With the full 40-trial training set (Figure 2B), *Late-fusion* plateaued at $T = 3.42$ s (98.4%) with a final accuracy of 98.4% $\pm$ 2.2%. It maintained a higher nAUC than *Ridge*, *btLDA*, and *Early-fusion* ($p \leq .001$), though the difference against *tsLDA* was not significant ($p = .052$). Similarly, *Late-fusion*'s final accuracy did not significantly differ from *tsLDA* ($p = .353$) or *btLDA* ($p = .096$) after correction (see the final column of Table 1).

DISCUSSION

In the section, we address the key findings of the staged pipeline selection process and highlight the *Late-fusion* approach, which achieved the strongest performance.

The superior performance with the 20–65 Hz band may suggest that the subtle I-c-VEP stimulus induces a weaker

visual drive, leaving low-frequency endogenous alpha activity less suppressed. Consistent with Riemannian theory [12], *tsLDA* benefited from a broader spectrum due to its manifold metric, which is inherently robust to the high-variance noise that typically degrades Euclidean classifiers.

While the 81.4% trial accuracy of the narrowband 60 Hz classifier confirms the carrier's discriminative value, the superior performance of the broadband models demonstrates that lower frequencies contain informative evoked components, providing essential complementary features.

The success of the *Late-fusion* underscores the benefits of combining fundamentally different models. Because individual models can provide partially orthogonal views of the neural data, their errors are likely not fully correlated. Furthermore, regularization in the base models may prevent highly confident misclassifications. Consequently, when averaging them, the scores of misclassifications may sometimes be suppressed by the confident correct predictions of complementary models, especially when aggregating multiple robust pipelines.

Regarding data efficiency, the *Late-fusion* model peaked at 98.4% accuracy (40 training trials) but reached a statistical plateau at just nine trials (62.4 s training time; 97.7% accuracy), demonstrating that calibration time can be substantially reduced. Future work should extend this toward zero-training paradigms for maximum efficiency [9].

Using all 40-trials for training, the *Late-fusion* model yielded an 88.8% accuracy at the 1.17 s decoding mark. Assuming a 1.5–1.0 s gaze-shift interval, this translates to an offline simulated ITR of 34.0–41.9 bits/min. While competitive for a four-target interface, the 12 Hz m-sequence presentation rate bottlenecks the ITR, as the long bit duration increases the minimum time required for target discrimination. Future studies should explore higher presentation rates to reduce the required decoding time and maximize system throughput. Finally, generating and aggregating multiple predictions per bit—similar to the decoding strategy seen in the EEG2Code framework [8]—could further enhance classification robustness. Furthermore, as this study establishes that the underlying code patterns remain reliably decodable, future research should investigate various high-frequency carriers and phase-shifting techniques to further enhance signal decodability and system throughput.

CONCLUSION

In this study, bit-wise linear decoding strategies for the imperceptible code-modulated visual evoked potentials paradigm were selected through a staged, domain-informed calibration process and validated on an independent hold-out dataset recorded one hour post-training across 16 participants. Both ridge regression and regularized LDA models exceeded 97% mean trial-level accuracy, with a *Late-fusion* model, which aggregated the

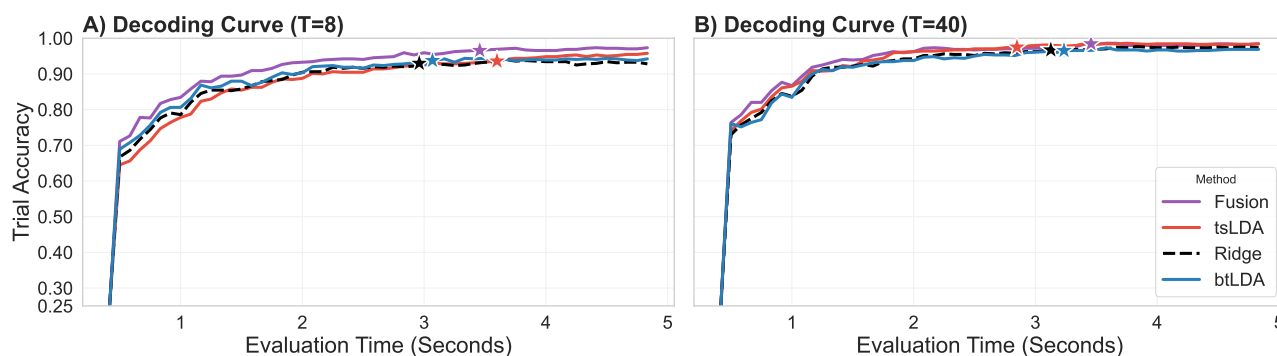


Figure 2: **Average offline decoding curves** as a function of evaluation time for $N = 16$ participants, 4 targets. **(A)** Model performance in a limited-data scenario, trained on only 8 trials (≈ 55 s of training). **(B)** Model performance trained on the full dataset (40 trials). The star markers denote the plateau point for each curve, defined statistically via a one-sided Wilcoxon signed-rank test as the earliest time point not significantly worse ($p > .05$) than the peak accuracy. The theoretical chance level is 25%.

outputs of independently trained base models, achieving a peak accuracy of 98.4%. These results establish that linear bit-wise decoding can provide a robust and computationally efficient foundation for high-performance, real-time I-c-VEP brain-computer interfaces.

ACKNOWLEDGEMENTS

This work was supported by the European Union's research and innovation program under the Marie Skłodowska-Curie grant agreement No 101118964, the "The Friends of the University Rhine-Waal - Campus Cleve" association, and the Dutch Research Council (NWO) (project number 024.005.022).

AUTHOR CONTRIBUTIONS

MAF: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft; MT, JT: Conceptualization, Methodology, Validation, Writing - review & editing, Supervision; IV: Funding acquisition, Writing - review & editing, Supervision.

REFERENCES

- [1] Wolpaw JR, R. Millán J del, Ramsey NF. Chapter 2 - Brain-computer interfaces: Definitions and principles. In: *Brain-Computer Interfaces*. Elsevier, 2020, 15–23.
- [2] Vialatte FB, Maurice M, Dauwels J, Cichocki A. Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives. *Progress in Neurobiology*. 2010;90(4):418–438.
- [3] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-Computer Interfaces Based on Code-Modulated Visual Evoked Potentials (c-VEP): A Literature Review. *Journal of Neural Engineering*. 2021;18(6):061002.
- [4] Azadi Moghadam M, Maleki A. Fatigue factors and fatigue indices in SSVEP-based brain-computer interfaces: a systematic review and meta-analysis. *Frontiers in Human Neuroscience*. 2023;17.
- [5] Brickwedde M, Bezsudnova Y, Kowalczyk A, Jensen O, Zhigalov A. Application of rapid invisible frequency tagging for brain computer interfaces. *Journal of Neuroscience Methods*. 2022;382:109726.

- [6] Liu K et al. A high-frequency SSVEP-BCI system based on a 360 Hz refresh rate. *Journal of Neural Engineering*. 2023;20(4):046042.
- [7] Fodor MA, Volosyak I. Towards Visual-Fatigue-Free BCI with Imperceptible Visual Evoked Potentials (I-VEP). In: *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Oct. 2025, 3813–3818.
- [8] Nagel S, Spüler M. Asynchronous non-invasive high-speed BCI speller with robust non-control state detection. *Scientific Reports*. 2019;9(1).
- [9] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. *Journal of Neural Engineering*. 2021.
- [10] Blankertz B, Lemm S, Treder M, Haufe S, Müller, K-R. Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage*. 2011;56(2):814–825.
- [11] Sosulski J, Tangermann M. Introducing block-Toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain-computer interfaces. *Journal of Neural Engineering*. 2022;19(6):066001.
- [12] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass Brain-Computer Interface Classification by Riemannian Geometry. *IEEE Transactions on Biomedical Engineering*. 2012;59(4):920–928.
- [13] Ying J, Wei Q, Zhou X. Riemannian geometry-based transfer learning for reducing training time in c-VEP BCIs. *Scientific Reports*. 2022;12(1).
- [14] Rivet B, Souloumiac A, Attina V, Gibert G. xDAWN Algorithm to Enhance Evoked Potentials: Application to Brain-Computer Interface. *IEEE Transactions on Biomedical Engineering*. 2009;56(8):2035–2043.
- [15] Ramoser H, Müller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*. 2000;8(4):441–446.

Ethical Considerations for User-Centric iBCI Studies: A Thematic Review

A. Thomas.^{1, 3} and R. Patel.^{1, 3}, J. Podmore.^{2, 3}, R. Scherer.², T. Carlson.¹

¹Aspire Create, University College London, England, UK

²Brain-Computer Interfaces and Neural Engineering Laboratory, University of Essex, England,
United Kingdom

³UK BCI Consortium, England, United Kingdom

E-mail: uceerjp@ucl.ac.uk, alex.thomas@ucl.ac.uk

ABSTRACT:

Implanted brain-computer interfaces (iBCI) have been transformative for patients with paralysis, substantively restoring independent function in key domains. Despite this, the experimental nature and surgical complexity of implants means only 67 patients received long-term implants across published studies between 1998–2023. As the field grows, understanding what users actually want and need from these systems becomes critical to shaping future studies and applications. We performed a reflexive thematic review on the publicly available transcript of 'The Lived Experiences of BCI Users' panel, conducted with three implanted BCI users, identifying 6 key themes: Purpose, Support, Understanding, Connection, Restriction, and Ethics. For each theme we discuss the core ideas with supporting quotes, derive best practice suggestions for future user-centred studies, and compare findings to existing literature, providing insight into the participant experience for researchers without direct exposure to user studies.

INTRODUCTION

The field of implanted brain computer interfaces (iBCI) has seen development since the late 1950s [1], with the first studies being used to map the functions of both animal and human brains. More recent studies within the past 2 decades have focussed on using implanted devices to control assistive technologies, aimed at improving the lives of people suffering motor impairments as a result of amyotrophic lateral sclerosis, high level complete or incomplete spinal-cord injury (SCI) and paralysis after stroke [2]. Although these trials have achieved great successes in controlling prosthesis, decoding imagined speech and generating tactile feeling through cortical stimulation, the limited number of participants in these studies can direct the research involved down a purely neuroscientific or engineering route with the aim of benefitting the user with the broader involvement of the current users or potential user base [3–5]. In this paper, we perform a thematic review of a long form panel discussion with 3 members of the BCI Pioneers Coalition, on the importance and principles of user centric design in iBCI development and deployment [6].

In this paper we aim to answer the research question: what are the most important needs and ethical practices for participants in iBCI studies?

MATERIALS AND METHODS

This study adopts an interpretative, contextualist framework; ethical meaning is treated as socially constructed and context-dependent [7]. Reflexive thematic analysis [8] was selected accordingly. All authors are BCI researchers who facilitated the conversation, introducing potential precommitment to technological optimism. Reflexive discussion throughout analysis ensured themes remained grounded in participant accounts rather than researcher aspiration.

Panellists and Data Collection: Three implanted BCI users from the BCI Pioneers¹ [9] community participated in a semi-structured group conversation. All three were males with SCI: a participant with incomplete quadriplegia from C3–C5 injury implanted with a Utah Array; a participant with cervical SCI implanted with six Utah Arrays; and a participant with C5 tetraplegia whose Neuroport electrode array had since been explanted.

The discussion focused on lived experience with iBCIs, including perceptions of ethical risk, value, access, governance, and legitimacy. Questions and topics were jointly decided between the researchers and the panel members. Questions were open-ended and exploratory rather than hypothesis-driven, enabling the panellists to articulate their own priorities and conceptual framings. The full transcript is publicly available via Zenodo [6].

Analysis Procedure: A reflexive thematic analysis was conducted following Braun and Clarke's six-phase approach [8]. All authors first familiarised themselves with the transcript independently through repeated reading and analytic memoing. Initial semantic coding was conducted individually to capture explicit meaning within the data. Codes were then discussed collectively, and broader patterns of shared meaning were constructed through iterative dialogue. Themes were developed collaboratively, refined through further engagement with the transcript, and finalised during the writing process. Coding was not

¹<https://bcipioneers.org/>

treated as a reliability exercise; interpretation was treated as an active, reflexive process. Consistent with reflexive thematic analysis conventions [8], results and interpretive commentary are presented in an integrated format."

RESULTS

In total, 6 core themes were identified: *Purpose, Support, Understanding, Connection, Restriction, and Ethics*.

Purpose: A first core theme identified was the purpose of the implant to the trial participants, and how involvement in research had granted them a new found purpose in life. This purpose is extremely important to the panellists as it is a primary driving force in their willingness to join the study and their continued dedicated participation. One panellist highlighted:

"But the potential is giving us participants a new purpose in life, and that's bigger than anything".

This purpose was especially meaningful to the panellists as it allowed them to specifically work on the technology to help their own rehabilitation and engagement with the world, directly helping themselves.

"If you give that person, you're caring for a chance at doing something for themselves, it changes their whole life. That's what I would say about it"

This sense of purpose was not only restricted to the experiment, but further reinforced by allowing them to revive experiences from their past, previous work, and hobbies:

"I've been a musician my whole life—I sing in a punk band. I really wondered if there was any way I would be able to generate music directly from my neuronal activity".

This highlights the considerable impact on a trial member's life that these implanted devices and projects have, and their ability to provide a purpose not in the present, but connected to the human experience [10]. This aligns with findings from non-invasive BCI studies, in which participation similarly restored biographical continuity and meaningful occupation [3–5].

A further sense of purpose was felt in contributing to the advancement of the field for future users, with active contribution being as important to participants as personal benefit [4]. This altruistic motivation is consistent across invasive and non-invasive BCI user populations [3, 5]; however, a meaningful distinction exists. Non-invasive users framed participation primarily as contribution, whereas iBCI participants here framed it as transformation, a restoration of identity and life purpose that extended well beyond the experimental context.

"You have so much pulled away from your life of what you can't do. This is a fantastic op-

portunity to be able to contribute positively towards not only society in general but also helping individuals who are going through the same things that you're going through."

Notably, this quote emphasises the panellists' desire to help shape the trajectory of the technology they were implanted with.

Support: A core theme was the need for support throughout the process of the study. The panellists described the support as not only physical due to their condition, but also as forms of direct and indirect psychological support throughout the research process.

Notably, a prominent form of support described was not logistical but relational, moral support derived from researcher dedication.

"member of the team in particular, the leader, showed up at the hospital every day just to be there—not doing research, but just building a very real connection"

Another key support structure discussed by the panellists is the support provided by the family and carers.

"I wouldn't be able to participate were it not for my wife, my care partner, being fully engaged in the process."

This sentiment highlight not only the need for support from the families during the trial, but also how the experimental surgery and experiment process does not solely change the life of the user, but also their wider family [11, 12]. The need for this support is also echoed in a contrasting anecdotal quote recalled by a panellist regarding a family member of another participant:

"His mom said, "His brain's the only thing normal left on him, and I'm afraid I'm gonna lose that because they're gonna put electrodes on his brain"."

The panellists also highlighted that this support is also required and needs to be guaranteed for a duration of the study. Participants may be reluctant to get involved if they fear the company or lab may go bankrupt. This concern is consistent with researcher-reported consensus that post-trial care obligations extend beyond study completion [12].

"Also, what are the ethical considerations behind supporting individuals participating in clinical trials, and what happens long-term for these devices if a company goes out of business or researchers just decide to stop doing the research?"

Continuous support is highly desirable for all future user-centred iBCI studies. This is reinforced by Pham et al., who found 83% researcher agreement that post-trial care obligations persist beyond study completion, and 82%

agreement that researchers should develop a nuanced understanding of participants' daily lives including caregivers [12]. That these conclusions were reached independently by researchers, without any end-user input in that study, strengthens the validity of our participants' accounts.

Understanding: Understanding the lived realities of participants is central if BCI research is to meaningfully serve the individuals it is intended to benefit. Discussions with the panellists consistently emphasised that neurological injury, while profoundly altering physical capability, does not alter the fundamental identity of the individual. As one panellist expressed,

"Yeah, this is a profoundly altering condition, but this person is still funny, still a pain in the ass, still whatever."

This observation highlights a persistent tension in the way participants can be framed within research contexts. Participants are often discussed in technical or clinical terms, yet the individuals involved retain the same personalities, social roles, and identities that existed prior to injury. In this sense, the panellists described successful BCI research not purely as a technical achievement, but as work that recognises the continuity of personhood beyond disability [5].

The panellists also described how neural stimulation and sensory feedback delivered through BCIs can rapidly become embodied. Rather than being experienced as an external technological input, these signals can integrate into existing perceptual frameworks. One participant noted that

"the stimulation was that close to real life that I didn't even have to think twice about how I perceived it."

These accounts suggest that when stimulation closely approximates natural sensory patterns, the resulting perception is not consciously interpreted as artificial. Instead, it is incorporated into the participant's existing sensory model of the world. From a design perspective, this reinforces the importance of developing systems that align with natural sensorimotor processing rather than introducing signals that must be cognitively interpreted. This process of sensory incorporation mirrors accounts from non-invasive users who described assistive devices becoming extensions of their own body image [5]. The discussions also highlighted the relational dimension of BCI research. The panellists repeatedly emphasised that productive collaboration depends not only on technical competence, but on researchers developing an understanding of the physiological and social realities associated with severe disability. As one panellist described,

"being sensitive to the weirdnesses of a quad[riplegia]'s physiological experience... that level of humanity is important."

These realities include fatigue, fluctuating bodily states, care routines, and other constraints that shape how and

when participation in research is possible. The panellists framed this awareness as a prerequisite for building trust and sustaining long-term engagement within research programmes.

Another theme concerned public understanding of BCI technologies. The panellists argued that demonstrations of BCI capability, particularly those that appear playful or expressive, are often undervalued within academic contexts. One panellist commented that

"scientists say it's gimmicky, but that's where people see how your science could be applied in real life."

From the participant perspective, these demonstrations play an important role in communicating the tangible potential of the technology to wider audiences. They translate otherwise abstract technical progress into applications that are immediately legible outside the laboratory. The panellists also raised concerns about representation and inclusivity within clinical trials. If participant cohorts are narrow or unrepresentative, the resulting systems risk being optimised for only a subset of potential users. One panellist noted,

"We need to make sure we're including a diverse group of individuals within these trials so we're not building models that are one size fits none."

This is directly supported by Pham et al.'s inclusivity principle [12] and by Kübler et al., who demonstrated that non-invasive BCI systems developed without sufficient user diversity failed to translate into daily use for the majority of end-users, a problem they termed the "translational gap" [13].

Finally, the panellists reflected on the emotional consequences of trial conclusion. One participant described returning to everyday limitations as

"a little bit of that loss all over again."

Kögel et al. documented this structurally: non-invasive users described losing their "job", their team, and their regular outings when studies ended, framing post-trial withdrawal not merely as psychological loss but as deprivation of social participation [5]. Gröbler et al. went further, identifying post-trial withdrawal as an established ethical concern, with professionals describing users left feeling useless after weeks of feeling important [4]. Taken together, these perspectives illustrate that BCI research cannot be understood purely in technical terms. The development of these systems is embedded within a broader human context involving identity, embodiment, social relationships, and the lived experience of disability. Recognising and engaging with these dimensions is therefore not peripheral to BCI innovation, but fundamental to ensuring that technological progress translates into meaningful outcomes for the individuals involved.

Connection: A core theme that was focused on by all of the panellists was the connection formed between the

research team and the participants throughout the trial process. One panellist felt as if the research team became a new family throughout the research process.

“You learn about who’s in each other’s lives. You can call them anytime. They become your friend, your family”

This code encapsulates the deep personal nature of participating in an iBCI study [4, 5], and the emotional involvement that is key to the studies success. This connection was not only felt solely with the research team carrying out the specific study. A consistent feature of the discussion was all of the panellists felt connection with the wider research community, whether that would be meeting new people at international events, or as they described:

I’m just really grateful that there are all you scientists out there helping figure these out for all the different ailments for all of us

Our iBCI panellists noted that this broader connection provided not only a sense of purpose, but a sustained source of morale drawn from the collective dedication of a global research community.

Further, connection to the research team and experience also has a positive effect on the participant, even through the difficult periods of the study, including the surgery structure as mentioned previously in support.

Restrictions: A significant portion of the discussion also focused on the current restrictions placed on users with iBCI technologies, including ownership of the equipment and access to the data.

This theme captures how iBCI participation remains institutionally bounded, resource-intensive, and only partially translated into meaningful daily use. Rather than being positioned as personal technologies, current systems are experienced as infrastructures controlled by laboratories, institutions, and trial protocols.

The panellists described limited authority over their own neural outputs.

One noted:

I wanted to use it in a recording, and [University] said, “No, we own that.”

Such accounts position data ownership and decision-making power with institutions rather than users, reinforcing asymmetries between participant contribution and institutional control which can differ globally [14]. Autonomy was further constrained by the structure of clinical trials. As one participant reflected:

The only time I’m actively engaged with my BCI is when I’m in the lab... I can’t hook in my system on my own. I really wish I could.

Even with implanted systems, use remains episodic and researcher-mediated, limiting individual benefit beyond experimental sessions.

Participation itself was described as socio-economically selective. Intensive schedules and care needs function as implicit gatekeepers:

“I wouldn’t be able to participate were it not for my wife... Most quad[riplegic]s require a fairly intense level of personal assistance.”

Access is therefore shaped not only by medical eligibility but by financial resources, care giving capacity, and logistical stability [3, 11].

The panellists also emphasised incomplete translation into everyday life. One articulated the aspiration that BCI should enable

“being able to do something that allows you to do what you want to do when you want to do it... and have the BCI really be just a data pathway, nothing more, nothing less.”

This contrasts with current systems that require specialist oversight and custom technical support.

Finally, scaling access was framed as ethically inseparable from privacy protection:

Making sure these devices are accessible in a way that you’re not giving up too much and not taking advantage of individuals with disabilities...

Together, these accounts reveal that iBCIs are presently experienced less as autonomous assistive tools and more as governed research devices, with meaningful access contingent on institutional structures and personal resources [3, 13]. A critical distinction from non-invasive literature applies here: Kübler et al. found that non-invasive users were blocked by hardware barriers (setup burden, electrode preparation, transport) [13], whereas the restrictions our panellists faced were institutional and ownership-based. Having already accepted surgical risk, the subsequent denial of data access and independent use represents a comparatively more acute ethical problem.

Ethical Implementation: Lastly, the panellists’ call for proactive ethical structuring of iBCI research and deployment. It spans anticipatory governance, neural data sovereignty, structured openness for collective benefit, protection of human agency, embodied surgical risk, and co-design grounded in lived experience.

Panellists described current regulatory frameworks as lagging behind technological capability, positioning governance as a precondition for responsible innovation rather than a corrective after harm occurs:

“If regulators aren’t thinking about the implications... we’re really going to have problems.”

Ethical responsibility was framed as distributed across regulators, institutions, and researchers [12], particularly in embedding privacy and ownership into present data practices.

Neural data were treated as uniquely sensitive and temporally unstable: information collected today may become

more revealing as decoding advances. This generated concerns around long-term ownership and secondary use, as one participant reflected:

“if someone would get that [data] 5 years from now and have more advanced decoding AIs, would they be able to decode my thoughts?... how do we give the user ownership of that data?”

At the same time, some advocated structured data sharing to accelerate progress given small participant populations, revealing an ongoing tension between collective benefit and durable user control [14]. This tension is quantified in the researcher literature: Naufel and Klein found 58% of BCI researchers supported giving participants access to their raw neural data post-study, yet also felt individuals should be limited in their freedom to donate or sell it, reflecting an institutional possessiveness over data that our panellists directly experienced [14]. Separately, Pham et al. found 90% of researchers agreed BCI data should be shared more widely [12], yet no mechanism for user-controlled sharing was proposed in either study.

The panellists also highlighted the moral weight of invasive intervention. Surgical risk, limited longitudinal evidence on cognitive effects, and the psychosocial impact of device removal extend ethical responsibility beyond initial consent and trial completion.

“I wish there were more data available on... what are the cognitive effects of the surgery... there aren’t really good longitudinal studies... Is there any potential for negative effects?”

Finally, co-design was presented as a structural requirement rather than optional consultation, ensuring that devices enhance rather than constrain agency [3, 12, 15].

“We can make sure we’re developing devices with lived experience and users in mind... that’s really going to make a difference”

Overall, iBCI participation was experienced within conditions of regulatory uncertainty, evolving data power, and incomplete longitudinal knowledge, necessitating anticipatory, participant-centred ethical frameworks.

DISCUSSION

After a full reflexive thematic analysis, we identified six core themes. Purpose emerged as a primary driver of participation, converging with Kögel et al. and Grübler et al. in showing altruism and biographical continuity as central motivations [4, 5], however a meaningful distinction emerged, namely that non-invasive participants described participation as meaningful and our iBCI panellists described it as reaffirming both identity and purpose. This difference likely reflects the greater restorative capacity of invasive systems, but also the higher personal stakes involved where participants had accepted surgical

risk, making their investment in the research fundamentally different in kind. Future research teams would benefit greatly from recognising this, and from including participants in the research process as much as possible to produce a development cycle targeting the outcomes that matter most to the people living with these devices.

By extension, support in the manner of clinical, familial, and relational plays a key role in participants’ ability to contribute. The dedication of researchers beyond a purely clinical role is an often overlooked source of support. Pham et al. found 82% of BCI researchers agreed that developing a nuanced understanding of end users’ daily lives, including their caregivers, was a professional responsibility [12], yet 83% also agreed post-trial care obligations are rarely formalised; a gap our panellists felt acutely. The effect of trial involvement on the wider family is equally important: Versalovic et al. found that for some users, concerns about social stigma and the burden of adjustment led loved ones to question whether functional benefit outweighed the cost [11]. The panellists in our review described their research teams as new families, a finding corroborated on the researcher side by Grübler et al. [4] and mirrored in non-invasive populations [5], suggesting that this depth of relational bond is a consistent feature of long-term BCI participation rather than incidental to it.

Understanding was our broadest theme, and arguably the one with the most immediate implications for research practice. Research teams must keep in mind that participants retain the same personalities, social roles, and identities throughout the study. The injury changes physical capability, not personhood. When this is overlooked, participants risk being reduced to data sources rather than collaborators which may corroborate some of the restrictions and ethical concerns that emerged. Equally, participant-defined goals risk being deprioritised in favour of publishable outcomes, a tension that is structural rather than intentional. Notably the incentives of academic research, sometimes do not naturally align with what users most want from their devices, something one must bear in mind when coproducing with end-users. Representative recruitment compounds this, if trial cohorts remain narrow, the resulting systems will reflect only a subset of potential users. Pham et al.’s inclusivity principle and Kübler et al.’s translational gap framework both point to the same conclusion: systems built without diverse user input tend not to generalise [12, 13].

Restrictions on out-of-lab use and data access remain a significant and underaddressed unmet need. The frustration our panellists expressed wanting their BCI to be simply “a data pathway” available whenever they needed it appeared almost verbatim in Peters et al.’s non-invasive cohort [3], suggesting this is a field-wide failure rather than a problem specific to invasive systems. What makes it more acute here is context: our panellists had already accepted the risk of permanent surgery, yet still found themselves unable to use their devices independently or access their own neural data. Kübler et al. showed that

even hardware barriers alone were sufficient to prevent non-invasive BCI adoption in daily life [13]; our panellists faced institutional barriers on top of these. Access to participation itself is further constrained by care networks and personal resources, meaning the population currently shaping iBCI development may not be representative of those who will eventually need it most [11].

The panellists raised urgent concerns about data regulation at a moment when decoding technology is accelerating rapidly in adjacent fields. Naufel and Klein found that 58% of BCI researchers supported giving participants access to their raw neural data, yet the same researchers felt individuals should be limited in their freedom to donate or sell it [14], a institutional possessiveness that our panellists encountered directly. Separately, Pham et al. found 90% of researchers agreed BCI data should be shared more widely, yet only 54% agreed end users should be part of the research team [12]. That tension, wanting the benefits of user data without full user participation, is precisely what our panellists were pushing back against. Critically, no end users participated in the Pham et al. survey; the present study provides the counterpart that the researcher-survey literature lacks. Our panellists were clear that co-design is not optional consultation but a structural requirement, a position supported consistently across the non-invasive literature [3, 13].

Limitations: This review is based on a single publicly available transcript involving three panellists, limiting both sample size and data diversity. All authors were involved in facilitating the original conversation, introducing potential researcher influence on what was voiced. Findings are not generalisable but are intended to inform and prompt further participant-centred inquiry. Further, all three panellists were male, US-based, English-speaking, and participated in trials at well-resourced institutions with substantial caregiver support. This relatively narrow demographic profile likely reflects the structural selectivity of iBCI trial participation, rather than the broader potential user population. Notably, the panellists themselves acknowledged this selectivity: one panellists described iBCI trial participants as "just the lucky, crazy ones who want to participate.". Findings should therefore be read as reflecting the experiences of highly engaged, resource-supported iBCI users, and may not capture the perspectives of those currently excluded from participation by the very barriers the themes identify.

CONCLUSION

We completed a reflexive thematic review uncovering six themes; Purpose, Support, Understanding, Connection, Restriction, and Ethics. Identifying positive practices and improvements in iBCI studies defined by users themselves. These findings can inform future study designs and provide insight into the participant experience for researchers not directly involved in user-centred work.

REFERENCES

- [1] Patil AC TN. Implantable neurotechnologies: A review of micro- and nanoelectrodes for neural recording. *Medical & Biological Engineering & Computing*. 2016;54:23–44.
- [2] Patrick-Krueger KM, Burkhart I, Contreras-Vidal JL. The state of clinical trials of implantable brain–computer interfaces. *Nature Reviews Bioengineering*. 2025;3:50–67.
- [3] Peters B et al. Brain-computer interface users speak up: The virtual users’ forum at the 2013 international brain-computer interface meeting. *Archives of Physical Medicine and Rehabilitation*. 2015;96:S33–S37.
- [4] Grüber G et al. Psychosocial and ethical aspects in non-invasive eeg-based bci research—a survey among bci users and bci professionals. *Neuroethics*. 2013;7(1):29–41.
- [5] Kögel J, Jox RJ, Friedrich O. What is it like to use a bci? – insights from an interview study with brain-computer interface users. *BMC Medical Ethics*. 2020;21:2.
- [6] Burkhart I et al. The lived experiences of bci users, defining the gold standard in human-centred bci research: Principles, progress, and perspectives from those who use it. Jan. 2026. [Online]. Available: <https://doi.org/10.5281/zenodo.18347790>.
- [7] Schwandt T. Constructivist, interpretivist approaches to human inquiry. *Handbook of Qualitative Research* Thousand Oaks, California: Sage. 1994.
- [8] Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 2006;3(2):77–101.
- [9] The bci pioneers coalition. Feb. 2026. [Online]. Available: <https://bcipioneers.org/>.
- [10] Hammell KW. Dimensions of meaning in the occupations of daily life. *Canadian Journal of Occupational Therapy*. 2004;71(5):296–305.
- [11] Versalovic E, Diamond M, Klein E. “re-identifying yourself”: A qualitative study of veteran views on implantable bci for mobility and communication in als. *Disability and Rehabilitation: Assistive Technology*. 2022;17:807–814.
- [12] Pham MT, Goering S, Sample M, Huggins J, Klein E. Asilomar survey: Researcher perspectives on ethical guidelines for bci research. *Brain-Computer Interfaces*. 2018;4(5):97–111.
- [13] Kübler A et al. The user-centered design as novel perspective for evaluating the usability of bci-controlled applications. *PLOS ONE*. 2014;9:e112392.
- [14] Naufel S, Klein E. Brain–computer interface (bci) researcher perspectives on neural data ownership and privacy. *Journal of Neural Engineering*. 2020;17:16039.
- [15] Koontz A, Duvall J, Johnson R, Reissman T, Smith E. “nothing about us without us:” engaging at users in at research. *Assistive Technology*. 2022;34(5):499–500.

EEG2EMG-Net: A Deep Learning Framework for Continuous EEG-to-EMG Estimation for Real-World Applications

D. Marcos-Martínez^{1,2,3}, E. Santamaría-Vázquez^{1,2,3}, V. Martínez-Cagigal^{1,2,3}, K. Kostoglou⁴, G. R. Müller-Putz^{4,5}, R. Hornero^{1,2,3}

¹Biomedical Engineering Group, University of Valladolid, Valladolid, Spain

²Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Valladolid, Spain

³ Valladolid Health Research Institute (IBioVALL), Valladolid, Spain

⁴Institute of Neural Engineering, Graz University of Technology, Graz, Austria

⁵BioTechMed Graz, Graz, Austria

E-mail: diego.marcos.martinez@uva.es

ABSTRACT: Estimating muscle activity from invasive neural recordings has been researched for decades with encouraging results. Such brain-computer interfaces (BCIs) are promising, as they offer an alternative pathway for motor control in patients with spinal cord injury. However, this application remains largely unexplored with non-invasive techniques, such as electroencephalography (EEG). Current non-invasive approaches typically rely on within-subject training, requiring both input (EEG) and output (electromyography, EMG) signals for calibration. This strategy is impractical for paralyzed patients unable to produce the required EMG. To address this, we developed EEG2EMG-Net a deep learning framework for continuous EEG-to-EMG estimation tailored for real-world clinical use. Our approach introduces a subject-adaptation strategy using synthetic EMG templates, eliminating the need for patient-specific muscle recordings during calibration. Results from 20 healthy participants show that the fine-tuning method allows accurate estimation of the EMG (pearson correlation coefficient: 0.69 ± 0.32) while suppressing spurious activations during resting-state. These findings open new research avenues for non-invasive BCI alternatives aimed at assisting motor function in individuals with severe impairments.

INTRODUCTION

Brain-computer interfaces (BCIs) hold significant potential in the field of neurorehabilitation. By decoding the kinetic and kinematic parameters of movements from neural activity, BCIs provide a pathway to re-establish functional motor control in individuals with paralyzed or weakened muscles. While a kinematic BCI decodes brain signals into control commands such as position or velocity [1], a kinetic BCI aims to estimate the muscle activation patterns that would result from voluntary motor intention [2]. This kinetic approach is particularly valuable as the predicted muscle activation can serve as an input

to stimulate muscles through functional electrical stimulation (FES), yielding the desired movement.

The continuous estimation of muscle activity from neural recordings for neurorehabilitation applications has been researched in the BCI field for decades. This kinetic decoding has been primarily explored using invasive techniques, demonstrating that high-fidelity muscle activation patterns can be extracted from the primary motor area (M1) to reconstruct movements in non-human primates [3, 4] and patients with spinal cord injury (SCI) [2]. In addition, arm and finger electromyography (EMG) signals have been decoded using electrocorticographic (ECoG) signals in humans [5].

Electroencephalography (EEG) is an appealing recording technique, as its non-invasive nature avoids the inherent risks associated with the surgeries required for invasive recording. However, EEG signals are characterized by limited spatial resolution and a low signal-to-noise ratio (SNR) due to volume conduction. Accordingly, BCI aimed at predicting muscle activity for neurorehabilitation applications have not been as extensively explored within the non-invasive domain as its invasive counterparts. Despite these challenges, EEG offers high temporal resolution, which is pivotal for capturing the fast, non-stationary dynamics of motor intent [6] and the corresponding muscle activation patterns. By leveraging a wide range of decoding strategies, from classical linear estimators to advanced signal processing and deep learning (DL) algorithms, EEG to EMG decoding is emerging as a compelling strategy for non-invasive neurorehabilitation devices.

In the literature, early works aimed at predicting muscle activity from various upper limb muscles relied on linear estimators, often fed by EEG spectral features. These linear estimators include multivariate linear regression [7, 8], sparse linear regression [7], Wiener filters [9] and generalized linear models [10]. These studies demonstrated the feasibility of decoding muscle activity in a continuous manner from non-invasive signals. However, the

use of linear approaches prevented the estimators from capturing the complex and inherent non-linear association between EEG and EMG signals. Consequently, these models were vulnerable to outliers and struggled to decode complex movements. To address such limitations, recent studies have focused on non-linear methods, including the Unscented Kalman Filter [11] and various DL architectures [8, 12]. These studies yielded superior performance over linear estimators, with the CNN-LSTM architecture in [12] standing out as the best EEG-based muscle activity estimator to date.

Despite these promising advances in non-invasive muscle activity decoding, several limitations limit their clinical translation for real-world neurorehabilitation. First, existing linear and non-linear estimators typically rely on a pure within-subject training strategy. This requires synchronous EEG and EMG data from the user to calibrate the model. However, in a clinical scenario, patients with SCI or paralysis are often unable to produce the necessary ground-truth EMG signals, making these calibration-dependent models impractical. Furthermore, several approaches rely on non-causal estimation by mapping concurrent EEG and EMG windows [12]. In practice, this introduces a significant algorithmic latency, as the system must process the entire window before generating an output. These accumulated delays prevent real-time execution, which is critical for the proper integration of sensory feedback and the overall efficacy of closed-loop BCI systems.

The aim of this exploratory study was to design and validate a DL-based muscle activity estimator that addresses a critical real-world constraint: the lack of ground-truth EMG signals in paralyzed patients. We developed a cross-subject model trained to estimate muscle activation patterns (EMG envelopes) using only EEG signals as input. To enhance performance, we introduce a novel fine-tuning strategy that adapts the pre-trained model to the individual's neural dynamics using synthetic EMG templates, thus eliminating the need for users to generate these signals. To the best of our knowledge, this is the first EEG-based BCI designed to operate exclusively on past neural information (i.e., causal), which makes real-time implementation feasible, while enabling accurate cross-subject muscle activity estimation with minimal and clinically realistic adaptation.

MATERIALS AND METHODS

Dataset and experimental protocol: A total of 20 healthy participants (aged 25.5 ± 4.8 years, 12 males, 8 females) participated in this study. All gave informed consent prior to the experiment. Each participant completed a single session consisting of a five-minute eyes-open resting-state recording, followed by active ME trials. The recordings were performed using the "Cross recorder" application within the MEDUSA© ecosystem [13]. The active ME recordings comprised 10-s 60 trials, during which participants were required to actively ex-

tend their right wrist and return it to the neutral position within a random 1 to 2-second interval. To ensure synchronization, a rotation cue was employed, consisting of a rotating cross overlaid on a stationary reference. Participants were instructed to initiate the movement exactly when the rotating cross aligned with the stationary one.

EEG and EMG signals were recorded using an actiCHamp system (*Brain Products GmbH, Germany*) at a sampling rate of 500 Hz. EEG data were acquired from 60 active electrodes distributed across the scalp following the international 10-05 system, referenced to the amplifier's internal common reference. The ground electrode was placed at FPz. Simultaneously, EMG activity was recorded using two passive electrodes in a bipolar configuration, placed over the extensor digitorum muscle of the right forearm.

Signal processing: EEG signals were processed using a band-pass filter (0.1–60 Hz) and a 50 Hz notch filter. Afterwards, signals were re-referenced to the common average reference (CAR). In parallel, EMG signals were high-pass filtered above 20 Hz and full-wave rectified by taking the absolute value. The EMG envelopes were then extracted via a 4 Hz low-pass filter [14]. Then, both signals were downsampled to 100 Hz to reduce computational cost. To maintain synchronization, the group delay introduced by the filtering stages was calculated, revealing a relative shift of 11 samples (110 ms) of the EMG envelopes with respect to the EEG signals. This latency was compensated in the subsequent signal processing step. The continuous recordings were segmented into 2-s epochs. An automatic artifact rejection method based on signal statistics was applied, followed by a visual inspection. Finally, muscle activity epochs were baseline-corrected by subtracting the 2nd percentile of the amplitudes and normalized to the 90th percentile of the ME-related EMG envelopes [14].

Training dataset construction: To prepare the data for the model, the previously segmented 2-s epochs were subdivided using a sliding window approach. Specifically, each EEG epoch was divided into smaller input examples of L ms, with a stride of 10 ms between consecutive windows. To ensure the causality of the prediction, each EEG window included only past and current neural dynamics ($X[t-L, t]$). The corresponding target was defined as an EMG envelope window of H ms ($Y[t+p, t+p+H]$), where p represents the physiological delay between the neural and muscle signals [15]. This procedure yielded a comprehensive dataset of input-target pairs that ensures causal, continuous decoding independent of trial structures. The parameters L , p and H were subsequently tuned to optimize the estimator's performance. Each input example was standardized using the z-score. This ensures that standardization only takes into account the statistics of the present and past samples.

Deep learning-based estimator: The architecture of our EMG estimator, EEG2EMG-Net, was based on EEG-Inception [16], optimized for continuous and causal muscle activity prediction. While retaining the multi-scale

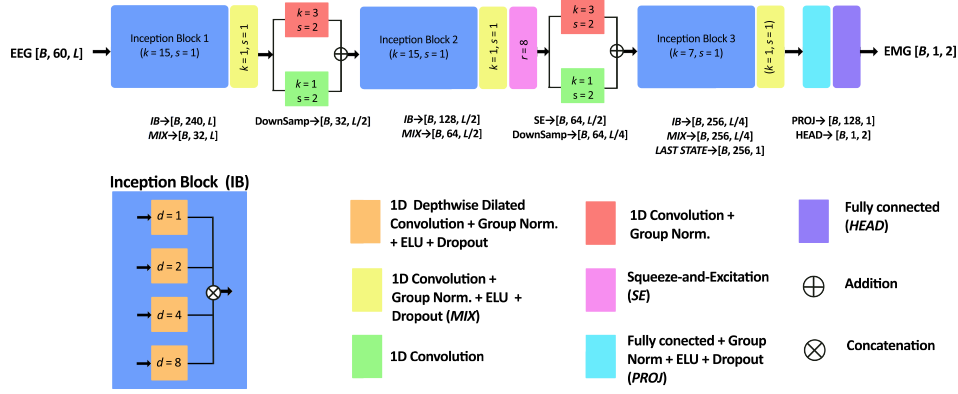


Figure 1: Overview of EEG2EMG-Net: Parameters k, s, d , and r denote kernel size, stride, dilation, and the Squeeze-and-Excitation reduction ratio, respectively. Tensor dimensions are expressed as $[B, C, L]$ (Batch, Channels, Length).

spatio-temporal extraction capabilities of the original model, EEG2EMG-Net introduces several key structural modifications. First, strict causal padding was implemented before each Inception and downsampling blocks to ensure real-time compatibility. This restricts the receptive field to past samples and prevents future-data leakage during prediction. Second, we employed dilated convolutions with a fixed kernel size to achieve an exponentially larger receptive field and high temporal resolution without increasing the model parameters. Furthermore, temporal downsampling was performed via strided convolutions rather than fixed pooling layers, enabling the model to learn complex, data-driven strategies for information synthesis. Additionally, Squeeze-and-Excitation (SE) blocks and residual connections were integrated to enhance feature extraction and training stability. Finally, the output stage was redesigned to map the latent features onto a regression horizon of H muscle activity samples. To achieve this, we extracted the last temporal step, referred to as the “Last State”, from the final convolutional block. This operation collapses the temporal dimension by selecting the most recent feature vector, which encapsulates the cumulative receptive field of the causal network, before passing it through the projection and fully connected layers for the final estimation. A detailed overview of the EEG2EMG-Net architecture and its implementation parameters is provided in Figure 1.

Hyperparameters optimization: The hyperparameters L , d , and H were optimized using a grid search within a five-fold cross-validation scheme. The search space for each parameter was defined as follows: $L \in \{200, 300, \dots, 600\}$ ms; $p \in \{20, 30, \dots, 50\}$ ms; and $H \in \{10, 30, \dots, 200\}$ ms. To ensure that our model was cross-subject, the data split was performed at the subject level: all samples belonging to a single participant were assigned exclusively to either the training or the validation set in any given fold. Specifically, in each iteration, 80% of the subjects were used for training, while the remaining 20% were reserved for validation. The model was trained with a batch size of 128, a learning rate of 0.001, and a dropout rate of 0.25. We used the

Mean Squared Error (MSE) between the real and the predicted muscle activity as loss function. Following the grid search, the optimal parameters for EMG decoding were identified as $L = 500$ ms, $p = 20$ ms, and $H = 20$ ms.

Training procedure: The training procedure comprised two sequential stages: a Leave-One-Subject-Out (LOSO) cross-validation to establish a subject-independent baseline, followed by a novel subject-adaptive phase. During each LOSO iteration, the model was trained using data from 19 subjects (15 for training and 4 for validation). The hyperparameters and loss function were kept identical to those described in the previous section. To prevent overfitting, an early stopping was performed if the validation loss failed to improve for 10 consecutive epochs.

Following the LOSO evaluation, we introduced a novel subject-adaptation process. While fine-tuning typically allows a pre-trained model to account for the inherent inter-subject variability of neural signals, its implementation in clinical settings is often limited by physiological constraints. In this sense, in a real-world clinical scenario, patients with SCI can generate motor intent EEG signals, but lack the ability to produce the corresponding ground truth EMG activity during ME attempts. Consequently, these signals are typically unavailable for subject-specific estimator adaptation. To overcome this limitation, we simulated the missing muscle activity by generating synthetic EMG templates derived from the healthy cohort. Specifically, the corresponding EMG trials of the remaining 19 subjects were temporally aligned to compute a grand average and its standard deviation. During the fine-tuning phase, the real EMG targets of the test subject were substituted with these synthetic templates. To introduce natural physiological variability and prevent the model from collapsing to a single static prediction, random noise (scaled by the computed standard deviation) was added to the grand average. This synthetic dataset generation was applied exclusively to the training and validation sets used during fine-tuning. The adaptation process was evaluated using a five-fold cross-validation scheme, where the data were partitioned into 80% for training/validation and 20% for testing. This ap-

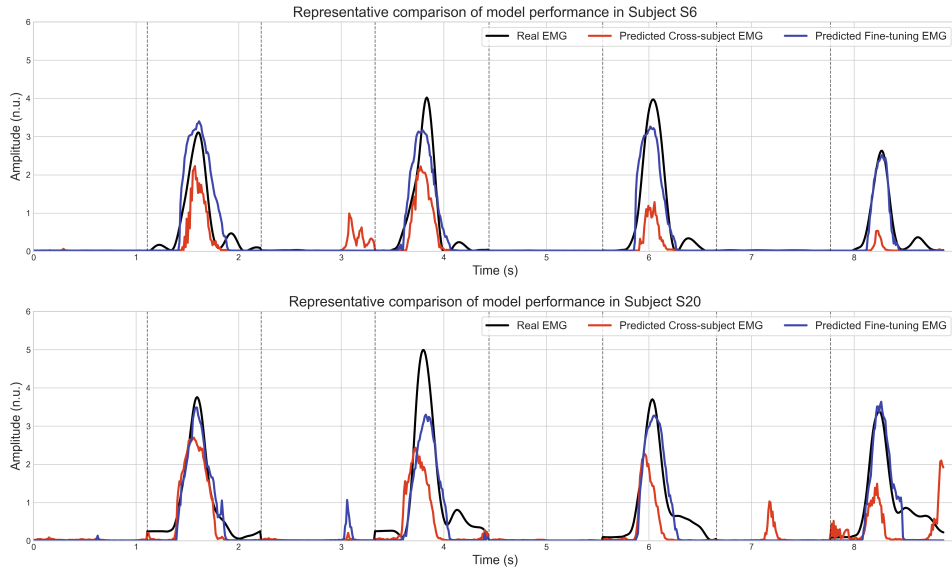


Figure 2: Representative EMG estimation examples for subjects S6 (top) and S20 (bottom). Temporal profiles show the real EMG envelope (black) compared against the estimations from the cross-subject model (red) and the subject-adaptive fine-tuned model (blue). Resting-state and motor execution epochs are delimited by vertical dashed lines.

proach ensures a comprehensive evaluation by allowing every trial to be utilized for both optimization and performance assessment. Importantly, as the input examples were generated using a highly overlapping sliding window, a standard random split would result in data leakage. Therefore, the data split was strictly performed at the event level (trial-wise), ensuring that all sliding windows originating from the same ME or resting-state event were grouped together in the same fold.

Evaluation metrics: To evaluate the estimator, condition-specific metrics were applied. During resting-state periods, performance was assessed using root mean squared error (RMSE) to identify spurious activations in the absence of motor intent. For the active ME condition, both Pearson correlation coefficient (PCC) and RMSE were employed. PCC enables to evaluate morphological fidelity while RMSE provides a complementary measure of magnitude accuracy relative to the real muscle contractions.

RESULTS

Representative estimation examples under the different training conditions are shown in Figure 2. Note that the different reconstructed epochs (delimited by dashed vertical lines) are the result of concatenating the output of several consecutive EEG inputs. Furthermore, Figure 3 provides a comprehensive comparison of the estimator’s performance across the cross-subject (red) and subject-adaptive (blue) modes. To evaluate the impact of the fine-tuning process, a paired one-tailed Wilcoxon signed-rank test was performed for each subject, with the False Discovery Rate controlled via the Benjamini-Hochberg (FDR-BH) procedure ($\alpha = 0.05$). Our analysis revealed that subject-adaptation significantly enhanced estimator

performance across the entire cohort for all evaluated conditions. During ME trials, the fine-tuning process yielded a statistically significant improvement in both RMSE and PCC for all 20 subjects. Importantly, this enhancement extended to the resting-state condition, where RMSE values were also significantly reduced for every participant. Furthermore, this superior performance was consistently maintained in the group-level analysis. These performance metrics, averaged across trials and subjects, are summarized in Table 1.

Table 1: Muscle activity estimation performance across subjects and trials (mean $\pm$ standard deviation). Lower RMSE and higher PCC indicate better performance.

Metric	Cross-subject	Fine-tuned
RMSE (Rest)	0.138 $\pm$ 0.185	0.072 $\pm$ 0.149
RMSE (ME)	1.108 $\pm$ 0.345	0.821 $\pm$ 0.385
PCC (ME)	0.443 $\pm$ 0.392	0.694 $\pm$ 0.324

DISCUSSION

Our findings, aligned with previous research in the field [8, 11, 12], proved that EEG signals encode sufficient information to directly estimate muscle activity from cortical dynamics. Notably, the performance of the cross-subject estimation demonstrated the ability of DL-based estimators to capture generalized neural-to-motor mappings shared across individuals.

Regarding the proposed subject-adaptation, our results showed that the novel approach significantly enhanced estimator performance. Specifically, RMSE values significantly decreased in the 20 subjects. Low RMSE values in estimating resting-state muscle activity are key for clinical safety, as unintended stimulation in the absence

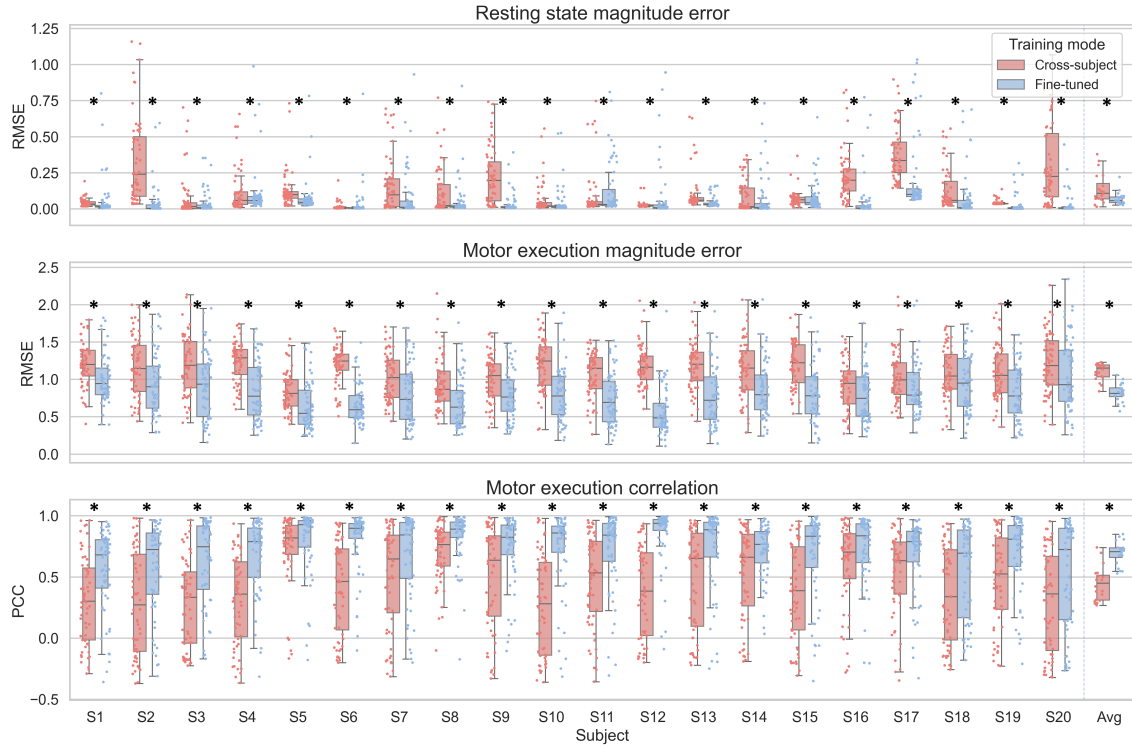


Figure 3: Performance comparison between cross-subject training and subject-adaptive fine-tuning. Distributions per subject are shown for resting-state RMSE (top), motor execution RMSE (middle), and Pearson Correlation Coefficient (PCC, bottom). The last column (AVG) presents the group-level performance averaged across all subjects. Boxplots indicate the median and interquartile range, with individual points representing single trials. Statistical significance between training modes was assessed per participant using a paired one-tailed Wilcoxon signed-rank test; asterisks denote significant differences after False Discovery Rate (FDR-BH) correction ($p < 0.05$).

of motor intent must be strictly avoided. Furthermore, despite the use of synthetic EMG templates instead of real muscle recordings, the fine-tuning process enabled the estimator to better capture the temporal coupling between neural motor intent and muscle activation. This was evidenced by a significant increase in PCC values during ME trials, indicating a better alignment between the ground truth and the estimated envelopes. However, despite yielding a significant reduction in RMSE during ME trials, the magnitude error after the fine-tuning process remained relatively high. This is a likely result of the fact that synthetic templates do not account for individual-specific EMG magnitudes. Thus, these RMSE results may reflect the intrinsic amplitude mismatch between the generalized templates and the participant’s actual muscle activity.

When compared to state-of-the-art EMG estimators, our subject-adapted model achieved performance levels comparable to existing EMG estimators. For instance, Sburlea *et al.* (2021) [11] reported a median PCC of 0.20 across 33 grasp conditions, with a maximum of 0.35. Similarly, Amiri *et al.* (2025) [12] and Kumarasinghe *et al.* (2021) [10] achieved mean PCC values ranging from 0.50 to 0.65 across multiple muscle groups and one specific movement. However, it should be noted that, unlike previous studies, our results were obtained without rely-

ing on the participants’ actual EMG signals. This highlights the potential of the proposed approach for clinical applications where ground-truth EMG is unavailable. Furthermore, to the best of our knowledge, no previous research has incorporated a resting-state condition in this context. We believe that accounting for the absence of motor intent from EEG signals is fundamental for ensuring safety and reliability in self-paced real-world applications.

Despite these promising results, we acknowledge several limitations in the present study. Due to the exploratory nature of this research, the current analysis was restricted to a single movement type and a single EMG channel. Consequently, future work should evaluate the proposed approach in more complex scenarios, such as multi-joint movements or high-density EMG recordings. Scaling this framework to multiple movements would represent a significant step toward practical motor control. Furthermore, while the proposed adaptation strategy was implemented using data from healthy participants, its application to clinical populations requires further validation. Although research indicates that patients with SCI engage similar motor cortical areas during motor attempts [17], it is also well-documented that post-injury cortical reorganization can alter these activation patterns [17]. Therefore, the real-world feasibility of our synthetic

template-based adaptation must be tested with SCI patients to ensure the estimator remains robust despite the specific EEG patterns induced by the injury.

CONCLUSION

To the best of our knowledge, this is the first study to explore the feasibility of an EEG-based muscle activity estimator specifically designed to adapt to real-world requirements. In this regard, we developed a causal and continuous DL model capable of estimating muscle activity without relying on a pure within-subject training strategy, as explored in previous studies. We found that while the cross-subject estimator achieved relatively good performance, it was significantly improved by following a novel subject-adaptation approach. By not relying on the real muscle activity of the testing subject, our fine-tuning approach could meet the requirements for use with SCI patients. These results pave the way for future BCI-based FES systems in neurorehabilitation.

ACKNOWLEDGMENTS

This work was supported by ‘0124_EUROAGE_MAS_4_E project’, co-financed by the EU through the Interreg VI-A Spain-Portugal Programme (POCTEP) 2021-2027, and by the VA140P24 project, funded by the Regional Government of Castilla y León (Consejería de Educación) and EU-ERDF. Support has also been provided by the Centre for Biomedical Research in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN, CB19/01/00012) through the Carlos III Health Institute, co-financed with ERDF funds.

REFERENCES

- [1] Müller-Putz GR *et al.* Feel Your Reach: An EEG-Based Framework to Continuously Detect Goal-Directed Movements and Error Processing to Gate Kinesthetic Feedback Informed Artificial Arm Control. *Frontiers in Human Neuroscience*. 2022;Volume 16 - 2022.
- [2] Rizzoglio F *et al.* From monkeys to humans: observation-based EMG brain-computer interface decoders for humans with paralysis. *Journal of Neural Engineering*. 2023;20(5):056040.
- [3] Moritz CT, Perlmutter SI, Fetz EE. Direct control of paralysed muscles by cortical neurons. *Nature*. 2008;456:639–642.
- [4] Ethier C, Oby ER, Bauman MJ, Miller LE. Restoration of grasp following paralysis through brain-controlled stimulation of muscles. *Nature*. 2012;485:368–371.
- [5] Flint RD *et al.* Extracting kinetic information from human motor cortical signals. *NeuroImage*. 2014;101:695–703.
- [6] Pfurtscheller G, Neuper C. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*. 1997;239(2):65–68.
- [7] Yoshimura N, DaSalla CS, Hanakawa T, Sato M aki, Koike Y. Reconstruction of flexor and extensor muscle activities from electroencephalography cortical currents. *NeuroImage*. 2012;59(2):1324–1337.
- [8] Kawase T, Yoshimura N, Kambara H, Koike Y. Controlling an electromyography-based power-assist device for the wrist using electroencephalography cortical currents. *Advanced Robotics*. 2017;31(1-2):88–96.
- [9] Paek AY, Brown JD, Gillespie RB, O’Malley MK, Shewokis PA, Contreras-Vidal JL. Reconstructing surface emg from scalp eeg during myoelectric control of a closed looped prosthetic device. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2013, 5602–5605.
- [10] Kumarasinghe K, Kasabov N, Taylor D. Brain-inspired spiking neural networks for decoding and understanding muscle activity and kinematics from electroencephalography signals during hand movements. *Scientific Reports*. 2021;11(1).
- [11] Sburlea A, Butturini N, Müller-Putz G. Predicting emg envelopes of grasping movements from eeg recordings using unscented kalman filtering. In: Proceedings of the Annual Meeting of the Austrian Society for Biomedical Engineering 2021. Verlag der Technischen Universität Graz, 2021, 71–74.
- [12] Amiri G, Shalchyan V. Decoding muscle activity via CNN-LSTM from 3D spatiotemporal EEG. *Computer Methods and Programs in Biomedicine*. 2025;271:108983.
- [13] Santamaría-Vázquez E *et al.* MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. *Computer Methods and Programs in Biomedicine*. 2023;230:107357.
- [14] Ma X, Rizzoglio F, Bodkin KL, Miller LE. Unsupervised, piecewise linear decoding enables an accurate prediction of muscle activity in a multi-task brain computer interface. *Journal of Neural Engineering*. 2025;22(1):016019.
- [15] Ibáñez J, Zicher B, Burdet E, Baker SN, Mehring C, Farina D. Peripheral neural interfaces for reading high-frequency brain signals. *Nature Biomedical Engineering*. 2025;9(9):1391–1402.
- [16] Santamaría-Vázquez E, Martínez-Cagigal V, Vaquerizo-Villar F, Hornero R. EEG-Inception: A Novel Deep Convolutional Neural Network for Assistive ERP-Based Brain-Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2020;28(12):2773–2782.
- [17] Hotz-Boendermaker S *et al.* Preservation of motor programs in paraplegics as demonstrated by attempted and imagined foot movements. *NeuroImage*. 2008;39(1):383–394.

Neural and Clinical Variability in Immersive VR-BCI Training: A Case-Series Study with Stroke Patients

Madalena Valente¹, Ines Oliveira^{1,2}, Jean-Claude Fernandes¹, Carla Mendes Pereira³, Ana Isabel Almeida³, Patricia Figueiredo¹, Athanasios Vourvopoulos¹

¹Institute for Systems and Robotics - Lisboa, Bioengineering Department, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

²Physiotherapy Department, Alcoitão Rehabilitation Medicine Center, Alcabideche, Portugal

³Health School, Polytechnic University of Setúbal, Setúbal, Portugal

E-mail: madalena.valente@tecnico.ulisboa.pt

ABSTRACT: Immersive virtual reality-based brain-computer interface (VR-BCI) systems have emerged as promising tools for post-stroke motor rehabilitation, providing embodied feedback and task-specific training. However, reliable neurophysiological predictors of clinical responsiveness remain unclear. This case-series study investigated whether longitudinal modulation of event-related desynchronization (ERD) during VR-BCI training predicts upper-limb motor improvement in individuals with sub-acute and chronic stroke. Five participants completed 12 VR-BCI sessions over four weeks, with clinical assessments conducted pre- and post-intervention. ERD was extracted using individualized frequency bands, and linear mixed-effects models were applied to estimate subject-specific ERD trajectories, including baseline magnitude and session-to-session change. These features were then examined in relation to motor recovery. Robust Alpha ERD was observed across participants, indicating consistent sensorimotor engagement. However, ERD progression across sessions did not significantly predict clinical gains, consistent with prior findings. Finally, BCI accuracy also showed substantial inter-subject variability and was not directly associated with motor improvement.

INTRODUCTION

Stroke remains a major cause of long-term disability worldwide, frequently leading to persistent upper-limb motor impairments that compromise functional independence [1]. Electroencephalography (EEG)-based brain-computer interfaces (BCIs) offer a unique approach by directly translating neural activity into real-time feedback or control signals [2].

Most motor rehabilitation BCIs rely on motor imagery (MI) or motor observation (MO) to modulate sensorimotor rhythms, eliciting event-related desynchronization (ERD) in the Alpha and Beta bands [2, 3]. ERD reflects task-related activation of motor cortical networks and constitutes a core feature for both decoding and neurofeedback paradigms. Meta-analyses indicate

that BCI-assisted interventions can yield moderate-to-large improvements in upper-limb motor function after stroke [4]. However, beyond aggregate clinical effects, the BCI field increasingly recognizes substantial inter-individual variability in neural modulation and controllability. More recent critiques argue that such variability often reflects system design and neural diversity rather than user deficits [5], reinforcing the need for personalized and adaptive BCI approaches [6]. Psychological and neurophysiological predictors of sensorimotor rhythms (SMR)-BCI performance further support the relevance of individual differences in baseline neural patterns and learning dynamics [7].

Integration of immersive virtual reality (VR) with BCI training introduces an ecologically rich feedback environment that can enhance embodiment, motivation, and sensorimotor coupling [8]. By closing the sensorimotor loop in a first-person context, VR-BCI systems may reinforce residual motor networks through contingent feedback. Nevertheless, substantial inter- and intra-subject variability in ERD magnitude, modulation stability, and online BCI accuracy remains a persistent challenge [9, 10]. Although stronger and more lateralized ERD patterns, particularly in the ipsilesional hemisphere, have been associated with improved motor outcomes and neuroplastic adaptation after stroke [11, 12], it remains unclear which neural features best characterize responsiveness during longitudinal VR-BCI interventions. In particular, the relative contributions of baseline ERD magnitude, session-to-session modulation dynamics, and online BCI accuracy to overall training responsiveness remain unclear.

While our previous work examined EEG-based predictors of recovery in a larger pilot cohort [13], the present study focuses specifically on neural and clinical variability within an independent case-series sample undergoing immersive VR-BCI rehabilitation. Here, we emphasize subject-specific ERD trajectories, inter-individual differences in BCI controllability, and their exploratory relationship with motor outcomes.

In this context, the present study investigates individ-

Table 1: Demographic and Clinical Profile of Participants with Stroke. FMA refers to the Fugl-Meyer Assessment for the upper limb (maximum value = 66). Δ FMA is the change between pre- and post-FMA scores, with * indicating minimal clinically important difference (MCID). MPS = months post-stroke.

ID	Sex	Age	Type	MPS	Side	FMA		
						Pre	Post	Δ
P01	F	29	Hem.	17	R	18	19	1
P02	M	61	Isch.	1	L	4	6	2
P03	M	54	Hem.	64	L	31	33	2
P04	M	62	Isch.	1	R	40	50	10*
P05	F	75	Isch.	27	R	17	17	0

ual learning dynamics during a 12-session VR-BCI intervention in sub-acute and chronic stroke. We focus on subject-specific ERD trajectories and online BCI accuracy to characterize neural controllability and variability across sessions. By examining the relationship between ERD-derived features, BCI performance, and motor outcomes, our objective is to contribute to the identification of neural signatures associated with the responsiveness of VR-BCI and to inform the development of more adaptive and personalized neurorehabilitation systems.

MATERIALS AND METHODS

Participants: Five individuals with stroke in the sub-acute and chronic phases were recruited at the Centro de Medicina de Reabilitação de Alcoitão (CMRA), Portugal. Mean age was 56 years (SD = 17), 3 male and 2 female. Two participants were affected by hemorrhagic stroke and three by ischemic stroke (Table 1). All participants provided their written informed consent in accordance with the Declaration of Helsinki. This study was approved by the Ethics Committee of the CMRA, with approval no.: 2025/007. Finally, all participants continued their usual therapy; no additional rehabilitation was provided in the study context. Harms were defined as any adverse events or unintended effects related to study procedures (e.g., discomfort, dizziness, anxiety, technical intolerance) and were monitored non-systematically through self-report and investigator observation.

Clinical Outcome Measures: The primary clinical outcome measure included the Fugl-Meyer Assessment for upper extremity (FMA-UE) [14, 15], and all participants with stroke were assessed pre- and post-intervention. The FMA is a standardized Likert-scale assessment widely used to evaluate motor function recovery following stroke-induced hemiplegia and has a maximum score of 66 points [14]. It assesses multiple domains, including motor function, sensation, range of motion, and joint pain. A minimal clinically meaningful improvement following treatment is typically defined as an increase of +4 to +7 points, for chronic stroke, and +10 for sub-acute stroke [16, 17].

Experimental Design: The experimental protocol for this study consisted of a 4-week intervention comprising 12 VR-BCI training sessions. Clinical evaluations were conducted at two distinct time points: (1) before the in-

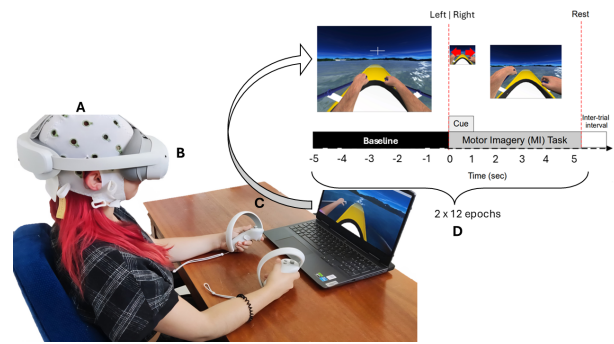


Figure 1: Experimental Setup: A. EEG cap for the System of 32 electrodes; B. HMD for VR feedback; C. VR controllers; D. MI protocol illustrating the epoch length and visual feedback.

tervention (pre); and (2) immediately after completing the intervention (post). EEG data were recorded using the Liveamp 32 EEG amplifier (Brain Products GmbH, Gilching, Germany), with a 32-channel setup and a sampling rate of 500 Hz. Visual feedback was delivered through the Pico 4 (Resolution: 2160x2160 per-eye; Refresh Rate: 90 Hz; field of view (FoV): 104° horizontal, 103° vertical; 6 degrees of freedom (DoF) motion tracking) standalone VR system (Figure 1B).

The VR-BCI training task used NeuRow, a first-person perspective upper-limb MI and MO paradigm in immersive VR [18]. The protocol involved MI-MO training using embodied feedback rendered via a Head-Mounted Display (HMD). Task-oriented VR environments have been recommended as effective approaches to support ecological validity and patient engagement in MI-based BCI rehabilitation protocols [19].

The training procedure comprised two phases: (1) Calibration — participants performed cued left- vs right-hand MI synchronized with avatar rowing actions. Each session included 12 randomized trials per hand, with 5 s of baseline and 5 s of MI (Figure 1D). A randomized inter-trial interval (1.25–1.5 s) was applied to prevent stimulus anticipation and carry-over effects. EEG data from calibration were used to extract spatial and spectral features through a common spatial patterns (CSP) filter (4 filters used, in the frequency band of 7 – 25 Hz), which trained a linear discriminant analysis (LDA) classifier, a widely adopted approach for MI-based BCIs [20]. (2) Online training — the trained model then classified MI patterns in real time, enabling participants to control the virtual boat through MI of the cued arm, with the same trial number and duration as in training, and as implemented in previous studies using similar protocols [13, 21]. The online BCI phase was not initiated on the first intervention day, providing participants with an initial familiarization period to adapt to the system and develop an effective neurofeedback control. Patients P01 and P02 started the online BCI phase in session 9, while the rest started in session 7. For each one of the 12 sessions, the number of runs (calibration plus online) was between 2 and 4. Every session where we performed online training, the CSP filters and LDA classifier were trained with

calibration data from the same session.

EEG Data Analysis: For post-hoc analysis, the EEG signals were processed using MATLAB R2023a (The MathWorks, MA, United States) and EEGLAB toolbox v2023.1. The EEG data preprocessing from the training sessions involved band-pass filtering between 1 and 40 Hz; then, for correct continuous noisy data and for rejecting bad channels, we applied the Artifact Subspace Reconstruction (ASR) method, in two steps, first to identify and interpolate bad channels, and then to reject bad segments. EEG data were subsequently downsampled to 125 Hz. Next, an Independent Component Analysis (ICA) was performed to remove remaining artifactual components in the EEG signals. ICA was computed on data after rejecting bad segments, and the resulting weights were applied to the data before segment rejection. We used the ICLabel [22] method for automatically removing any component classified as "muscle" or "eye" artifact (probability values: $\geq 90\%$). The average number of independent components removed was 3. Further, we performed re-referencing to the common average and then rejected bad trials. Finally, the data were segmented into epochs corresponding to left-hand and right-hand trials, and then time-frequency analysis was performed through the event-related spectral perturbation (ERSP). Since the ERSP values were in decibels, we transformed them into ERD as a percentage decrease, based on the following formula:

$$ERD(\%) = (10^{ERSP/10} - 1) \times 100\% \quad (1)$$

Moreover, since ERD magnitude can differ substantially both across participants and within the same participant over time, we adopted an automatic individualized frequency-band approach to improve the precision of ERD estimation [13]. This strategy enables the analysis to more accurately reflect subject-specific neural oscillatory patterns, particularly given the spectral shifts and altered rhythmic activity commonly observed after stroke recovery [23]. Rather than applying the standard 8-12 Hz Alpha range, we defined a custom frequency window between 6-14 Hz for each participant, electrode, task, and trial. The full implementation of this approach is described further in Valente et al. [13] and it is publicly accessible online ¹.

The ERD was extracted from all 32 electrodes, using the same method. Our analysis focused primarily on C3 and C4, as they are positioned over the motor and somatosensory cortices [24], they are also the most commonly analysed in the literature, facilitating comparisons with previous studies.

Statistical Analysis: All analyses were performed in MATLAB R2023a (The MathWorks, MA, United States). Linear mixed-effects models (LMEs) were used to analyze neurophysiological data, accounting for repeated measures and inter-individual variability, including random intercepts per participant. Statistical significance was set at two-tailed $p < 0.05$.

¹<https://github.com/LaSEEB/Individualized-ERD>

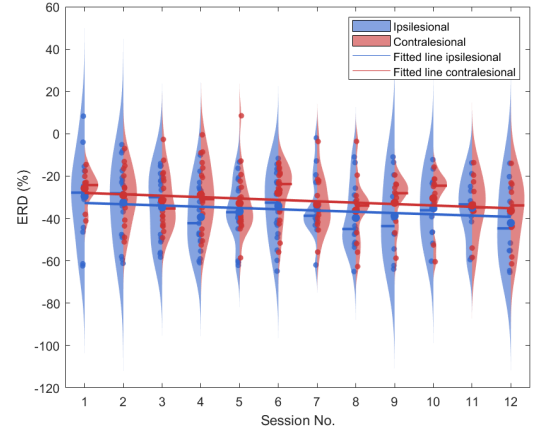


Figure 2: Ipsilesional and contralateral ERD distributions across VR-BCI training sessions for each participant. For each session, each participant had between 2 and 4 runs, which we use here, instead of only one value per participant per session. Fitted lines represent linear regression trends.

To examine the relationship between longitudinal predictors (ERD progression across training sessions) and cross-sectional clinical outcomes (ΔFMA), we implemented a two-stage modeling approach, used in previous studies [23, 25]. LME models were selected because they provide robust regression estimates while accounting for subject-specific variability.

In the first stage, an LME model was fitted to estimate individual ERD trajectories (from the affected side) over the intervention period using the following formula:

$$\text{mean_erd} \sim \text{sessions} \times \text{stroke_type} + (\text{sessions} \mid \text{subjects}) \quad (2)$$

This model included time (sessions) and subjects as random effects, allowing subject-specific variability in baseline ERD (intercept) and ERD progression over time (slope). The extracted intercept and slope represent each participant's baseline ERD level and the rate of change in ERD throughout the intervention.

In the second stage, the extracted ERD *intercepts* and *slopes* were then incorporated into a linear regression model to predict clinical motor recovery, measured by ΔFMA :

$$\text{delta_fma} \sim \text{slope} \times \text{intercept} \quad (3)$$

RESULTS

Clinical Outcomes: Considering the upper-limb impairment level according to FMA-UE scores in pre-assessment, participants P03 and P04 presented mild impairment, and participants P01, P02, and P05 presented severe impairment. At post-assessment, one participant (P04) showed an improvement ($\Delta FMA = 10$), corresponding to the minimal clinical importance difference for the sub-acute population. In contrast, the remaining participants demonstrated improvements ranging from 0 to 2, below significant values (Table 1).

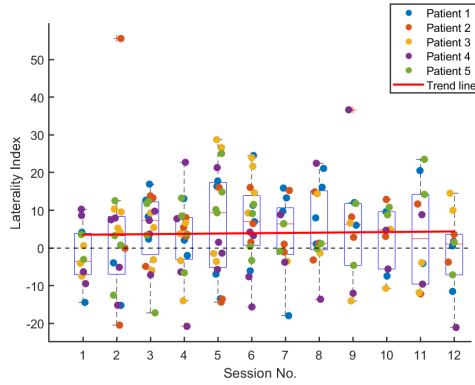


Figure 3: ERD LI across VR-BCI training sessions for each participant. Fitted line represents linear regression trend.

ERD Dynamics Across Sessions: ERD was consistently observed across sessions in both ipsilesional ($y = -0.595x - 32.088$) and contralesional ($y = -0.658x - 27.276$) sensorimotor regions (Figure 2). ERD values remained strongly negative throughout the intervention, typically ranging between approximately -20% and -60%, indicating sustained suppression of sensorimotor rhythms during motor imagery training.

Laterality Index Analysis: In terms of ERD hemispheric patterns, LI values showed substantial variability across sessions and participants, with distributions centered around zero (Figure 3). Linear trend analysis revealed a small positive slope ($y = 0.079x + 3.441$), suggesting a mild shift toward ipsilesional dominance.

Linear Mixed-Effects Modeling: Finally, to formally examine longitudinal ERD dynamics, the LME model revealed a significant negative intercept ($\beta = -37.97$, $p < 0.001$), indicating strong overall ERD suppression across participants during the intervention. However, session progression was not statistically significant ($\beta = 0.07$, $p = 0.89$), suggesting that ERD magnitude remained relatively stable across training sessions. Stroke type (ischemic vs hemorrhagic) also did not significantly influence ERD trajectories ($\beta = 7.34$, $p = 0.46$), and the interaction between session and stroke type was not significant ($\beta = -0.07$, $p = 0.93$). Random-effects estimates indicated substantial between-subject variability in baseline ERD ($\sigma^2 = 9.95$), whereas variability in session-related slopes was comparatively small ($\sigma^2 = 0.61$) (Table 2). A second-stage linear regression analysis examined whether ERD baseline magnitude (intercept) and ERD progression (slope) predicted motor recovery (ΔFMA). The model explained 27% of the variance in ΔFMA but was not statistically significant ($R^2 = 0.27$, $F = 0.123$, $p = 0.935$), indicating that neither baseline ERD nor longitudinal ERD modulation reliably predicted clinical improvement in this cohort (Table 3).

BCI Accuracy: BCI accuracy showed marked inter-subject variability across sessions (Figure 4). Participant P02 achieved the highest performance, reaching near-ceiling accuracies in the later sessions (95.8–100%). In contrast, participants P01 and P03 remained close to chance level throughout the intervention (approximately

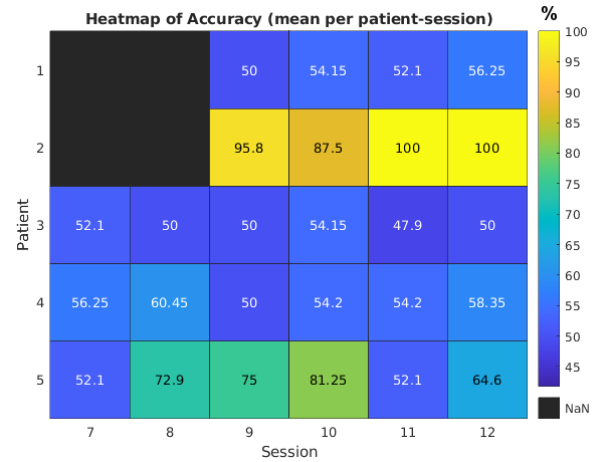


Figure 4: Heatmap of mean BCI accuracy per online session for each participant. For sessions with two runs, accuracy represents the mean across runs.

50–56%). Participants P04 and P05 demonstrated intermediate performance, with accuracies generally ranging between 50–70% and showing moderate fluctuations across sessions. The highest accuracy values occurred in sessions in which participants appeared to combine MI with slight overt motor intent or minimal limb movement. Such behavior can increase the separability of sensorimotor EEG patterns and consequently improve classification performance. Finally, the relationship between high BCI accuracy (ΔAcc) and clinical improvement (ΔFMA) showed no significant relationship ($r = 0.06$, $p = 0.92$).

DISCUSSION

This case-series study extends our prior randomized pilot by testing the same ERD extraction and mixed-effects trajectory framework in an independent five-patient cohort [13]. Consistent with the larger study, ERD engage-

Table 2: Fixed and random effects estimates from the LME Model analyzing ERD progression over sessions and stroke types.

Fixed Effects	β (95% CI)	t	p
Intercept	-37.97 [-50.53, -25.41]	-5.97	1.4×10^{-8}
Sessions	0.07 [-0.99, 1.14]	0.14	0.89
Stroke Type	7.34 [-12.42, 27.09]	0.73	0.46
Sessions $\times$ Type	-0.07 [-1.74, 1.59]	-0.09	0.93
Random Effects	Var (95% CI)		
Subject (Intercept)	9.95 [4.71, 21.02]		
Session (Slope)	0.61 [0.13, 2.86]		
Residual	12.85 [11.52, 14.34]		

Table 3: Second-stage regression linking ERD features to ΔFMA .

Predictor	β	t-statistic	p-value
Intercept	14.39	0.51	0.70
ERD Slope	-35.73	-0.49	0.71
Baseline ERD	0.24	0.38	0.77
Slope $\times$ Baseline	-0.93	-0.48	0.71
Model Fit			
$R^2 = 0.27$			$p = 0.94$

ment was robust, but longitudinal ERD progression did not emerge as a reliable predictor of motor gains. Further, previous work investigating ERD dynamics during stroke recovery, reported that patients with favorable recovery trajectories exhibited a progressive increase in ipsilesional ERD modulation accompanied by a decrease in contralesional activity during the acute phase after stroke [26]. In contrast, the present cohort of chronic stroke participants showed relatively stable ERD magnitudes across sessions, with only a mild shift toward ipsilesional dominance reflected in the LI. These differences may reflect the later stage of recovery in our cohort, where neural engagement patterns may already be stabilized and less sensitive to longitudinal modulation during training.

Clinical Implications: From a clinical perspective, only one participant (P04) achieved a clinically meaningful improvement in FMA-UE ($\Delta FMA = 10$), exceeding the minimal clinically important difference. Taking into account the higher baseline assessment of participant P04 ($FMA - UE = 40$), corresponding to a lower impairment level, these findings suggest that differences in baseline motor function may have influenced clinical outcomes. These results are consistent with prior BCI rehabilitation studies where baseline FMA-UE ≤ 30 has been associated with poorer recovery outcomes [27].

ERD Dynamics During VR-BCI Training: Across sessions, VR-BCI training consistently elicited robust ERD over sensorimotor areas. ERD values remained strongly negative throughout the intervention, indicating reliable engagement of motor-related cortical networks during the training task. However, ERD magnitude showed minimal longitudinal modulation. LME model indicated that ERD remained relatively stable across sessions, suggesting that participants were able to engage sensorimotor rhythms early in the intervention rather than progressively increasing ERD magnitude over time.

Hemispheric Variability: Analysis of the LI revealed heterogeneous hemispheric recruitment patterns. Although a slight shift toward greater ipsilesional dominance was observed across sessions, LI values remained widely distributed around zero, indicating that both hemispheres contributed to motor imagery engagement. Such variability is consistent with prior studies showing that stroke recovery may involve either ipsilesional reactivation or compensatory contralesional recruitment depending on lesion characteristics and recovery stage ([12, 26]).

Longitudinal ERD Trajectories: The two-stage modeling approach further indicated that ERD features alone were not sufficient to explain clinical improvement. While baseline ERD and slope explained a modest portion of the variance in motor recovery, the regression model was not statistically significant. These findings suggest that simple ERD magnitude or linear progression metrics may not capture the complex neural adaptations underlying functional recovery in VR-BCI interventions.

BCI accuracy and recovery outcomes: BCI accuracy also showed substantial variability across participants. Notably, decoding performance did not correspond to

clinical improvement, reinforcing the dissociation between BCI classification accuracy and functional recovery outcomes ([28, 29]).

Together, these findings emphasize the importance of considering inter-individual neural variability when interpreting BCI rehabilitation outcomes, and the importance of personalized and patient-centered VR-BCI rehabilitation paradigm [19].

Limitations: Several limitations should be acknowledged. First, the small sample size limits statistical power and the generalizability of the findings. Further, heterogeneity in stroke chronicity and baseline motor impairment likely contributed to the variability observed in both neural and behavioral outcomes. Future studies with larger cohorts and longer longitudinal protocols will be necessary to better characterize intra-individual neural dynamics and their relationship with motor recovery during VR-BCI rehabilitation.

CONCLUSION

This case-series study investigated neural and clinical variability during immersive VR-BCI rehabilitation in individuals with chronic stroke. The results show that VR-BCI training reliably elicited sensorimotor ERD, confirming consistent engagement of motor-related cortical networks during motor imagery. However, longitudinal ERD modulation did not significantly predict motor recovery, and substantial inter-individual variability was observed in both neural activity and clinical outcomes. These findings suggest that simple ERD magnitude or linear ERD progression may not fully capture the neural mechanisms underlying functional recovery in VR-BCI interventions. Instead, recovery may involve more complex neurophysiological adaptations beyond mean ERD measures. Overall, this work highlights the importance of considering neural variability when interpreting ERD-based biomarkers and supports the need for more personalized approaches in VR-BCI neurorehabilitation.

ACKNOWLEDGMENTS

This research was supported by the Fundação para a Ciência e Tecnologia (FCT) through the LARSyS grant (DOI:10.54499/LA/P/0083/2020; 10.54499/UIDP/50009/2020; 10.54499/UIDB/50009/2020), and the NOISyS project (DOI: 10.54499/2022.02283.PTDC).

REFERENCES

- [1] Feigin VL et al. World stroke organization (wso): Global stroke fact sheet 2022. *International Journal of Stroke*. 2022;17(1):18–29.
- [2] Neuper C, Scherer R, Wriessnegger S, Pfurtscheller G. Motor imagery and action observation: Modulation of sensorimotor brain rhythms during mental control of

- a brain-computer interface. *Clinical Neurophysiology*. 2009;120(2):239–247.
- [3] Pfurtscheller G, Aranibar A. Event-related cortical desynchronization detected by power measurements of scalp eeg. *Electroencephalography and Clinical Neurophysiology*. 1977;42(6):817–826.
- [4] Cervera MA et al. Brain-computer interfaces for post-stroke motor rehabilitation: A meta-analysis. *Annals of Clinical and Translational Neurology*. 2018;5(5):651–663.
- [5] Thompson MC. Critiquing the concept of bci illiteracy. *Science and engineering ethics*. 2019;25(4):1217–1233.
- [6] Perdakis S, Millan JdR. Brain-machine interfaces: A tale of two learners. *IEEE Systems, Man, and Cybernetics Magazine*. 2020;6(3):12–19.
- [7] Jeunet C, N’Kaoua B, Lotte F. Advances in user-training for mental-imagery-based bci control: Psychological and cognitive factors and their neural correlates. *Progress in brain research*. 2016;228:3–35.
- [8] Vourvopoulos A, Jorge C, Abreu R, Figueiredo P, Fernandes JC, Badia S Bermúdez i. Efficacy and brain imaging correlates of an immersive motor imagery bci-driven vr system for upper limb motor rehabilitation: A clinical case report. *Frontiers in human neuroscience*. 2019;13:244.
- [9] Brain EBS. Intra-and inter-subject variability in. *Inter-and Intra-subject Variability in Brain Imaging and Decoding*. 2022:8722.
- [10] Lotte F, Larrue F, Mühl C. Flaws in current human training protocols for spontaneous brain-computer interfaces: Lessons learned from instructional design. *Frontiers in Human Neuroscience*. 2013;7.
- [11] Ray AM, Figueiredo TDC, López-Larraz E, Birbaumer N, Ramos-Murguialday A. Brain oscillatory activity as a biomarker of motor recovery in chronic stroke. *Human Brain Mapping*. 2020;41(5):1296–1308.
- [12] Remsik AB et al. Ipsilesional mu rhythm desynchronization correlates with improvements in affected hand grip strength and functional connectivity in sensorimotor cortices following BCI-FES intervention for upper extremity in stroke survivors. *Frontiers in Human Neuroscience*. 2021;15:725645.
- [13] Valente M, Branco D, Badia S Bermúdez i, Fernandes JC, Figueiredo P, Vourvopoulos A. Eeg-based predictors of motor recovery during immersive vr-bci rehabilitation. *Scientific Reports*. 2026.
- [14] Gladstone DJ, Danells CJ, Black SE. The fugl-meyer assessment of motor recovery after stroke: A critical review of its measurement properties. *Neurorehabilitation and Neural Repair*. 2002;16(3):232–240.
- [15] Woodbury ML, Velozo CA, Richards LG, Duncan PW. Rasch analysis staging methodology to classify upper extremity movement impairment after stroke. *Archives of physical medicine and rehabilitation*. 2013;94(8):1527–1533.
- [16] Page SJ, Fulk GD, Boyne P. Clinically important differences for the upper-extremity fugl-meyer scale in people with minimal to moderate impairment due to chronic stroke. *Physical Therapy*. 2012;92(6):791–798.
- [17] Narayan Arya K, Verma R, Garg R. Estimating the minimal clinically important difference of an upper extremity recovery measure in subacute stroke patients. *Topics in stroke rehabilitation*. 2011;18(sup1):599–610.
- [18] Vourvopoulos A, Ferreira A, Badia SB i. Neurow: An immersive vr environment for motor-imagery training with the use of brain-computer interfaces and vibrotactile feedback. In: *International Conference on Physiological Computing Systems*. 2016, 43–53.
- [19] Oliveira I, Russo M, Almeida AI, Vourvopoulos A, Mendes Pereira C. Recommendations for combining brain-computer interface, motor imagery, and virtual reality in upper limb stroke rehabilitation: Qualitative participatory design study. *JMIR Rehabilitation and Assistive Technologies*. 2025;12:e71789.
- [20] Lotte F et al. A review of classification algorithms for eeg-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering*. 2018;15(3):031005.
- [21] Vourvopoulos A, Fleury M, Blanco-Mora DA, Fernandes JC, Figueiredo P, Badia S Bermúdez i. Brain imaging and clinical outcome of embodied vr-bci training in chronic stroke patients: A longitudinal pilot study. *Brain-Computer Interfaces*. 2024;11(4):193–209.
- [22] Pion-Tonachini L, Kreutz-Delgado K, Makeig S. Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*. 2019;198:181–197.
- [23] Ray A et al. Brain oscillatory activity as a biomarker of motor recovery in chronic stroke. *Human Brain Mapping*. 2020;41(5):1296–1308.
- [24] Pfurtscheller G, Lopes da Silva F. Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [25] Afonso M, Sánchez-Cuesta F, González-Zamorano Y, Romero Muñoz JP, Vourvopoulos A. Investigating the synergistic neuromodulation effect of bilateral rtms and vr brain-computer interfaces training in chronic stroke patients. *Journal of Neural Engineering*. 2024;21.
- [26] Tangwiriyasakul C, Verhagen R, Rutten WL, Putten MJ van. Temporal evolution of event-related desynchronization in acute stroke: A pilot study. *Clinical neurophysiology*. 2014;125(6):1112–1120.
- [27] Wu Q et al. Analysis of prognostic risk factors determining poor functional recovery after comprehensive rehabilitation including motor-imagery brain-computer interface training in stroke patients: A prospective study. *Frontiers in Neurology*. 2021;12:661816.
- [28] Ang KK et al. A large clinical study on the ability of stroke patients to use an eeg-based motor imagery brain-computer interface. *Clinical EEG and Neuroscience*. 2011;42(4):253–258.
- [29] Prasad G, Herman P, Coyle D, McDonough S, Crosbie J. Applying a brain-computer interface to support motor imagery practice in people with stroke for upper limb recovery: A feasibility study. *Journal of neuroengineering and rehabilitation*. 2010;7(1):60.

Learning brain connectivity features for improving BCI performance

M. Pavaux^{1, 2}, F. de Vico Fallani¹, D. La Rocca²

¹Paris Brain Institute - ICM, Sorbonne Universite, CNRS, Inria, Inserm, AP-HP, Pitie Salpetriere
Hospital, Paris, France

²Thales cortAix Labs, Palaiseau, France

E-mail: marion.pavaux@icm-institute.org

ABSTRACT: Brain computer interfaces (BCIs) rely on the decoding of information embedded in brain signals for applications ranging from neurofeedback to device control. While power spectral features are widely used in motor imagery-based BCIs, recent evidence suggests that considering network interactions between sensors could improve performance. In practice however, connectivity between different brain signals is impaired by the high variability of classical estimators for short noisy segments. To address this issue, we propose a deep learning framework trained to recover ground-truth imaginary coherence from complex simulated signals. Here, we show that our framework learns to reconstruct smooth coherence spectra, designed as Gaussian mixtures. Evaluation on an electroencephalography dataset across three motor execution and imagery tasks, shows that connectivity matrices derived from the proposed method consistently outperform those obtained with Welch's estimator. These results demonstrate the feasibility of using deep learning to obtain stable informative coherence estimations in sparse and noisy contexts, supporting the use of functional brain networks for BCI applications.

INTRODUCTION

Modeling the brain as a network can provide highly informative insights in brain-computer interfaces [1]. The resulting interactions reflect the functional or statistical relationships between brain regions during different tasks. This concept is intuitive, and brain connectivity has long been central in network neuroscience, as it captures distributed coordination that cannot be inferred from local activity alone. For instance, recent advances have shown potential in using connectivity for motor imagery based Brain Computer Interfaces (BCIs), as including those connectivity features in mental states classification led to an increase in accuracy [2–4].

However, estimating connectivity from brain signals presents significant challenges. Classical models typically assume stationary long signals, whereas in EEG-based BCIs, signals are noisy and trial windows are short. Such biases can render connectivity estimates noisy and highly unstable across time windows, trials, or subjects, limiting their practical utility. Due to their variability, connectivity estimators need careful attention and exper-

tise to be used properly, which is why they have been less popular than power spectra counterparts [5].

In this study, we explore whether Deep Learning can optimize the estimation of connectivity features for short and noisy signals, by outperforming traditional approaches. Using a combination of temporal convolutions, an attention mechanism and synthetically driven supervision, where imaginary coherence was designed as a Gaussian mixture, we are able to better reproduce connectivity in simulated signals compared to Welch's method. We test our approach on motor imagery and execution tasks from the PhysioNet EEG Motor Movement/Imagery Dataset [6], and are able to obtain better classification performance with matrices derived from our method, compared to matrices derived from Welch's.

MATERIALS AND METHODS

Imaginary Coherence: Imaginary coherence measures phase-lagged coupling between two signals $x(t)$ and $y(t)$ while ignoring contributions close to zero-lag correlations caused by volume conduction [7]. It ranges from 0 to 1.

Method summary: The objective of this study is to estimate imaginary coherence in the 1–40 Hz range from short EEG segments. Because ground-truth coherence cannot be accessed in real recordings, we first train a deep learning model on simulated signal pairs for which imaginary coherence is analytically defined. The learned estimator is then evaluated on three motor execution or imagery tasks using the PhysioNet EEG Motor Movement/Imagery Dataset [6].

Synthetic data generation: Precise control over coherence values is required to supervise training. Signal pairs are therefore simulated according to the Fourier decomposition principle, allowing direct parameterization of amplitude and phase at each frequency (Eq. 1). Amplitudes $A_{x,f}$ and $A_{y,f}$ are sampled uniformly in $[0,1]$, and phases ϕ_f in $[0, 2\pi]$. Frequencies range from 1 to 40 Hz with a resolution of 0.05 Hz. Sampling rate is 100 Hz. A total of 10,000 pairs of signals are generated.

$$\begin{cases} x(t) = \sum_f A_{x,f} \sin(2\pi f t + \phi_f) \\ y(t) = \sum_f A_{y,f} \sin(2\pi f t + \phi_f + \Delta\phi(f)) \end{cases} \quad (1)$$

To ensure a smooth yet structured spectral maxima, the

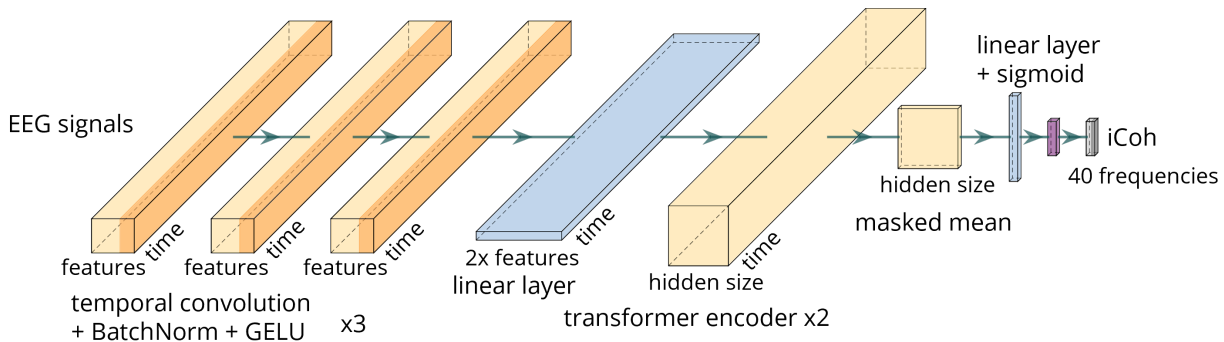


Figure 1: **Model architecture.** The architecture is composed of three temporal convolutional blocks, each followed by batch normalization and GELU activation. Features extracted from the two signals are then concatenated and projected into a higher dimensional space and processed by two Transformer encoder layers. The representations are averaged across the temporal dimension, and fed into a linear layer with 40 outputs. A sigmoid activation clips predictions between 0 and 1.

phase difference spectrum $\Delta\phi(f)$ is modeled as a Gaussian mixture in which Gaussians' widths σ_p range from 0.5 to 6 Hz, centers $f_{c,p}$ from 1 to 40 Hz, and amplitudes A_p from $-\pi$ to π . This formulation ensures smooth spectral lobes while preserving variability and sharp transitions.

Twenty percent of the simulated data are held out for validation. Standardization is performed on the signals using statistics computed on the training set. Furthermore, noise will be added during training and windows will be randomly cropped (see section Model training).

EEG data processing: Evaluation is conducted on the PhysioNet EEG Motor Movement/Imagery Dataset. Signals are band-pass filtered between 1 and 43 Hz using a 10th-order Butterworth filter ($f_s = 160$), epoched from 0 to 3 seconds after cue onset, and linearly detrended.

Model architecture: Local temporal patterns are first extracted using three convolutional layers (kernel size 3, 32 filters, Batch Normalization, GELU activation [8]). This stage captures short-range dependencies and yields 64 feature maps for each signal pair.

The features are then projected into a 512-dimensional embedding space and processed by two Transformer encoder layers (4 attention heads, hidden dimension 512), enabling the modeling of longer-range temporal interactions. Before entering the encoder, features undergo layer normalization and sinusoidal positional encoding. The temporal dimension is averaged prior to a final linear layer of size 40, corresponding to the frequency bins. A sigmoid activation constrains the predicted imaginary coherence to the [0,1] interval (Fig. 1).

Model training: The model is trained using L1 loss optimized with Adam (learning rate 10^{-5} , batch size 16). To promote robustness to noise and short segments, each simulated pair is augmented four times with additive Gaussian noise, with standard deviation linearly spaced between 0 and a randomly selected maximum between 1.8 and 2. In addition, signals are cropped four times using window lengths linearly spaced between a minimum randomly drawn between 100 and 120 samples, and 300 samples. These augmentations increase variability during

training and increase the effective batch size from 16 to 256.

Welch's computation: We compare our model to Welch's method, which estimates power spectra by averaging periodograms of overlapping windows before computing imaginary coherence. Using the Python SciPy library implementation, 3-second segments sampled at 160 Hz are analyzed with 1-second windows (160 samples), 0.9-second overlap, and a Hann window to reduce spectral leakage.

Classification: To evaluate the quality of our prediction, we compute for each 3-second epoch the pairwise upper triangle coherence matrix (2016 pairs, 1–40 Hz, 40 frequency bins) using both Welch's method and the proposed model. Each epoch corresponds to either a first condition (left hand closing, left hand closing imagination, or hands imagination) or second condition (right hand closing, right hand closing imagination or feet imagination). A logistic regression classifier with standard data scaling is trained under a subject-wise 5-fold cross-validation scheme. Balanced accuracy is used to assess performance.

RESULTS

Performance on simulated data: The model is first trained on simulated data, as described in Materials and Methods. The L1 validation loss reached 0.114. Since the analytical imaginary coherence spectrum is known in this setting, it is possible to directly assess spectral fidelity across frequencies and compare it with Welch's method. Coherence is estimated on 3-second validation segments. The model closely follows the analytical target across the full frequency range, whereas Welch's estimates exhibit higher noisiness (Fig. 2). This behavior is consistent with the known sensitivity of Welch's method to short segment length and windowing effects.

Validation on EEG-based BCI data: After confirming spectral accuracy on simulated data, we assess whether the learned estimator transfers to real EEG recordings from the PhysioNet EEG Motor Movement/Imagery

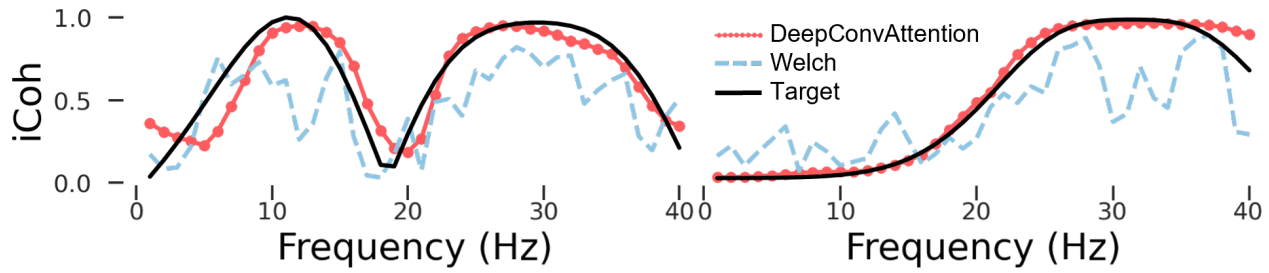


Figure 2: **iCoh spectrum in simulated data.** Predicted imaginary coherence is compared to Welch’s method on validation data. The proposed model better matches the analytical target across frequencies, while Welch’s estimates show increased noisiness.

Dataset, while real data was not seen during training. Connectivity matrices are computed across the two methods and averaged in the alpha band (8 to 13Hz) and the beta band (13 to 25 Hz). Fig. 3 displays those matrices for a representative subject (subject 1). Results show that the model recovers structured and spatially consistent connectivity patterns that align with established properties of phase-lagged interactions. Channel pairs located close to the diagonal exhibit weaker connectivity, reflecting reduced phase differences between neighboring electrodes. In contrast, more distant regions display stronger long-range interactions. Although absolute scales differ between methods, the spatial organization remains comparable (Fig. 3).

Classification performance on BCI tasks: The usability of our method is assessed using coherence matrices (2016 pairs, 1–40 Hz) as features for three BCI tasks from the PhysioNet’s EEG Motor Movement/Imagery Dataset: (i) left vs. right hand execution, (ii) left vs. right hand imagery, and (iii) hands vs. feet imagery. Using only a simple logistic regression classifier (Materials and Methods), we demonstrate that matrices derived from our method yielded superior classification performance across the three tasks (Tab 1). Both the minimum and maximum subject-level performances were higher with the proposed method. The inter-subject variability is comparable between approaches, with mean standard deviations of 0.101 for the proposed method and 0.103 for Welch.

DISCUSSION

Connectivity estimation in BCI remains constrained by non-stationarity and limited data length. Classical approaches rely on windowing and averaging to stabilize spectral estimates. Although those methods have been widely used, they still yield variable connectivity patterns as EEG data remain particularly challenging. Using ground truth imaginary coherence in noisy and sparse data settings however, encourages the model to learn a stable spectral representation. Transfer from simulation to EEG is enabled by carefully designing the source domain to reflect properties of real signals. Namely, resting-state EEG typically exhibits smooth spectral organization over extended recordings, motivating the use of Gaussian

mixtures to model phase differences. In addition, Gaussians’ parameters were chosen to simulate more complex phase portraits. Lastly, a dense set of sinusoids is used to approach the infinite frequency resolution of EEG signals. To further improve estimation quality on EEG data, a systematic investigation of the simulation parameters and their underlying statistical distributions is warranted. It would also be of interest to explore other more complex simulation frameworks, based on multivariate autoregressive (MVAR) models for instance, that can fully capture time-varying phase couplings [5].

Although power spectral density remains the dominant feature in motor imagery BCIs, the present work highlights the discriminative value of functional connectivity. Using a deliberately simple classifier, connectivity matrices derived from the proposed estimator consistently outperformed those obtained with Welch’s method. This suggests that improving the stability of connectivity estimation can directly enhance downstream decoding performance. However, further analysis needs to be conducted to assess how these connectivity features compare to power spectral density features, and how combining the two approaches could ultimately improve decoding performance.

Finally, the comparison is restricted to Welch’s estimator. Alternative short-window spectral methods, such as autoregressive approaches, should be considered in future work. Moreover, the framework could potentially be extended to learn other spectral or connectivity estimators, opening perspectives for broader applications.

CONCLUSION

Despite its potential, the use of functional connectivity and network analysis in BCIs remains a difficult challenge because of the variability of current estimators. Here, we demonstrate that training a deep learning architecture for estimating imaginary coherence provides more precise and less variable estimates, especially in short-time noisy windows, compared to Welch’s method. When applied to a 64-channel EEG-BCI dataset, the proposed framework consistently improves classification performance over standard Fourier-based estimation. These results support the feasibility of learning-based connectivity estimation and encourage further ex-

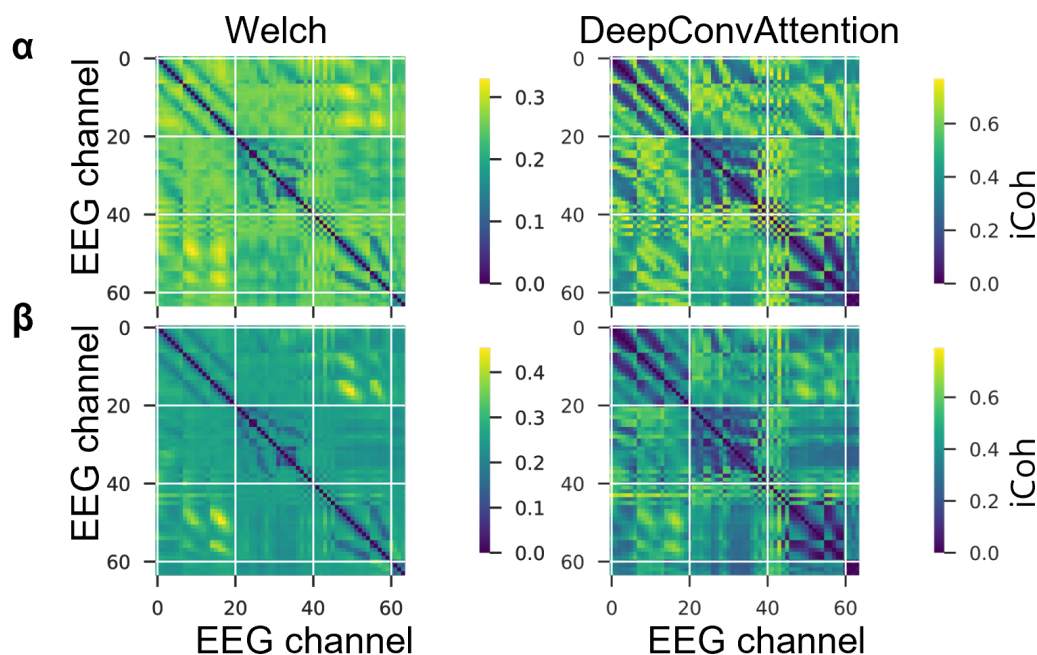


Figure 3: **iCoh connectivity matrices comparison in the alpha and beta bands for a representative subject.** Using PhysioNet’s EEG Motor Movement/Imagery Dataset, connectivity matrices are computed in subject 1 for Welch’s method and the proposed model. While the model was not trained on real EEG data, band-averaged coherence (alpha [8, 13] Hz, beta [13, 25] Hz) reveals consistent spatial patterns across methods, showing the transferability of our method to EEG data. Our method produces finer spatial resolution while preserving physiologically plausible organization.

Table 1: Balanced accuracy on three BCI tasks

Model/Task	Execution left vs right hand	Imagery left vs right hand	Imagery hands vs feet
Welch	0.564 ± 0.095	0.567 ± 0.082	0.556 ± 0.080
DeepConvAttention	0.678 ± 0.101	0.667 ± 0.106	0.603 ± 0.095

ploration of connectivity-driven representations in BCI decoding frameworks.

ACKNOWLEDGMENT

FdVF was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant Agreement N° 864729) and by the Agence Nationale de la Recherche under the France 2030 program (Grant Agreement ANR-23-IACL-0008).

REFERENCES

- [1] De Vico Fallani F, Bassett DS. Network neuroscience for optimizing brain–computer interfaces. *Physics of Life Reviews*. 2019;31:304–309.
- [2] Cattai T, Colonnese S, Corsi MC, Bassett DS, Scarano G, De Vico Fallani F. Phase/amplitude synchronization of brain signals during motor imagery BCI tasks. *arXiv:1912.02745 [q-bio]*. Dec. 2019. Accessed: Jan. 29, 2025. [Online]. Available: <http://arxiv.org/abs/1912.02745>.
- [3] Perinelli A, Assecondi S, Tagliabue CF, Mazza V. Power shift and connectivity changes in healthy

aging during resting-state EEG. *NeuroImage*. 2022;256:119247.

- [4] Gonzalez-Astudillo J, Cattai T, Bassignana G, Corsi MC, De Vico Fallani F. Network-based brain–computer interfaces: Principles and applications. *Journal of Neural Engineering*. 2021;18(1):011001.
- [5] De Vico Fallani F, Richiardi J, Chavez M, Achard S. Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. 2014;369(1653):20130521.
- [6] Schalk G, McFarland D, Hinterberger T, Birbaumer N, Wolpaw J. BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Transactions on Biomedical Engineering*. 2004;51(6):1034–1043.
- [7] Nolte G, Bai O, Wheaton L, Mari Z, Vorbach S, Hallett M. Identifying true brain interaction from EEG data using the imaginary part of coherence. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*. 2004;115(10):2292–2307.
- [8] Hendrycks D, Gimpel K. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415 [cs]*. Jun. 2023. Accessed: Feb. 27, 2026. [Online]. Available: <http://arxiv.org/abs/1606.08415>.

A Systematic Overview of EEG-Detectable Control-Related Mental States and Processes in HCI

Matthias Eidel¹, Tomko Settgast¹, Sarita Silveira Teja², Laurens R. Krol², Johanna Wagner²,
Andrea Kübler¹

¹ Institute of Psychology, Department of Psychology I, University of Würzburg, Würzburg,
Germany

² Zander Labs, Germany

E-mail: matthias.eidel@uni-wuerzburg.de

ABSTRACT: We report our endeavor to systematically map and **describe all EEG-detectable mental states and processes (MS)** with a focus on **Human-Computer Interaction (HCI)** from existing literature.

Methods: Literature was collected in two steps. First, we performed a large-scale query against three relevant literature databases to systematically retrieve all articles dealing with EEG-detectable MS within a **HCI context**. Second, we broadened our scope to encompass all MS without the restriction to HCI. We obtained and screened the full list of paradigms from the #EEGManyLabs consortium and performed a second database query, looking specifically for influential reviews and meta-analyses about EEG-detectable MS.

Results: We provide a limited overview, focusing on mental states relating to control processes, based on 1024 screened articles, thus far. Specifically, we describe the different MS within this category and their respective **elicitation paradigms** and **neural correlates**.

Significance: While originally created for the NAFAS project, this overview may serve as **a reference for anyone working with EEG-detectable mental states**.

INTRODUCTION

The present study is part of the NAFAS project (Neuroadaptivity for Autonomous Systems www.zanderlabs.com/nafas). An overview of this project was presented at the 9th Graz BCI Conference [1]. It encompasses a number of synergistic hard- and software developments aimed at developing safe, practical neurotechnological solutions that allow technology to adapt to and learn from ongoing human cognition, ranging from human-in-the-loop neuroadaptive HCI applications to autonomously or semi-autonomously operating neuroadaptive artificial intelligences. Human cognition can be represented by different **mental states or processes (MS)**, e.g., whether a person is attentive and concentrated, overburdened, surprised, or notices an error. Several MS have electrophysiological neural correlates that can be reliably detected using EEG or other tools, thus, allowing the decoding of the associated brain activity in real time [2]. While these correlates are generally associated with individual

neurophysiological processes, brain activity is complex, with a multitude of simultaneously active mental states and processes.

One goal of NAFAS is to define a **multidimensional mental state (MDMS)**. MDMS refers to a **combination of several distinctive dimensions** of maximally orthogonal, ‘atomic’ mental states and processes and is meant to reflect a more complex state than individual dimensions. Where traditional BCI applications usually deploy one classifier focused on one individual mental state or process, NAFAS proposes to deploy several classifiers in parallel, resulting in a multidimensional representation of a user’s mental state. With this information, HCI and AI systems, now with access to a comprehensive representation of a user’s current state, could **anticipate or react** to their needs and adjust their functionality accordingly, without the need for direct user input [3]. Such systems should be able to detect and interpret task-relevant mental processes from brain signals in real time.

To achieve this, NAFAS includes the following efforts:

- 1) The **identification of all EEG-detectable MS**. From this list, a subset of suitable, ideally non-redundant, maximally orthogonal dimensions of MS shall be selected to **define the MDMS**.
- 2) The **collection of EEG data** from a large and diverse sample using **standardized paradigms** to elicit the states and processes. With this, classifiers will be trained, enabling parallel decoding that requires no or little individual calibration.

To date, there is no exhaustive list of EEG-detectable MS, and available literature is often inconsistent in the definition, nomenclature and neurophysiological aspects of the respective MS. Thus, we describe our recent effort to **create a comprehensive list that categorizes all EEG-detectable mental states** and provides information on how they are **elicited** and **decoded**, based on a systematic literature screening.

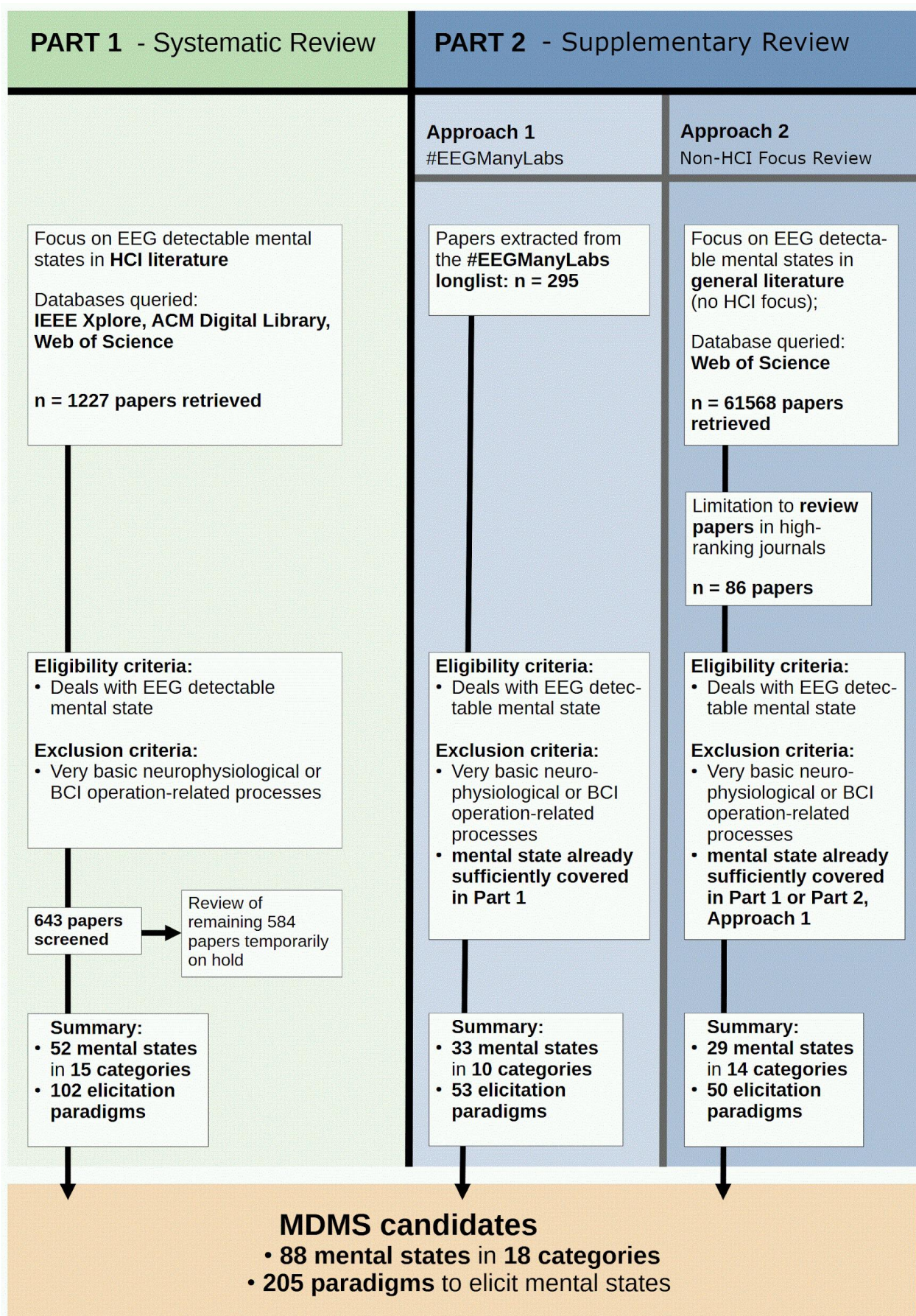


Figure 1: Overview of the literature retrieval and screening process.

MATERIALS AND METHODS

We retrieved literature using two different approaches: a systematic scoping review and a broader screening of existing literature, as delineated in **Figure 1** and described below:

Part I - Systematic Literature Review: This approach was loosely based on the PRISMA (preferred reporting items for systematic reviews and meta-analyses) guidelines for scoping reviews (www.prisma-statement.org) [4]. Here, we looked for EEG-detectable MS published explicitly within an **HCI context**. To match all papers on **mental states or processes** in relation to **HCI** and **EEG**, we defined the following logical search term:

((*mental OR psychological OR cognitive OR affective OR emotional OR brain*) AND (*state OR process*))
AND
(*human OR man*) AND (*computer OR machine OR technology OR robot OR artificial intelligence OR AI*)
AND (*interaction OR communication OR interface OR cooperation OR symbiosis*)
AND
(*EEG OR electroencephalography OR electroencephalogram*)

We queried against three highly relevant databases: The *IEEE Xplore Digital Library*, the *Web of Science* (all databases) & the *ACM Digital Library*.

The generic logical search term was adapted to match the syntax of the respective database. For IEEE Xplore and Web of Science, we restricted our search to title, abstract, and keywords. For ACM, “mental states” and similar were searched in the full text, whereas “EEG” (or similar terms) had to be mentioned in the abstract. The “HCI” term was not applied, but the search was restricted to records being classified as HCI-related via the ACM-distinct *Computing Classification System*.

Overall, this resulted in a list of 1227 hits after excluding duplicates via DOI.

Part II: Extended Literature Screening: To further broaden the scope, we extended the search to fit any EEG-detectable MS from the general neuroscientific literature, i.e., without the prior obligatory HCI context (hence, search terms such as ‘computer’, ‘interaction’ or ‘interface’ were removed from the query). Instead, we focused on **review papers** from the following **high-ranking journals**: *Nature Reviews Neuroscience*, *Trends in Neuroscience*, *Trends in Cognitive Science*, *Nature Reviews Neurology*, *Nature Communications*, *Nature Neuroscience*, *Nature Human Behavior*, *Neuron* (Cell Press), *Behavioral and Brain Sciences*.

Further, we obtained an existing list of **impactful EEG paradigms** (295 papers) nominated by active EEG researchers (the #EEGManyLabs project) [5]. This list was screened for additional MS paradigms. If a MS had already been discovered in Part I, we only added novel information, such as an elicitation paradigm not previously listed for the respective mental process.

Both approaches are summarized in **Figure 1**.

Information Extraction: All literature from these approaches was pooled and remaining duplicates were removed, resulting in 1608 research articles and reviews. Papers were subjected to a first screening to determine relevance and eligibility. Inclusion criteria were as follows: The MS must be detectable by EEG alone. Papers dealing with basic neurophysiological processes, e.g., sensory processing potentials such as visually evoked potentials, and BCI paradigms (e.g., motor imagery) were excluded. Further exclusion criteria were: Literature dealing with neurological or mental disorders, lack of an elicitation or detection paradigm, and inaccessible literature.

Included literature was subjected to a thorough review (currently ongoing), during which we extracted the most relevant information, in particular: The **name** of the MS **as provided by the authors**, the **elicitation** paradigm, the associated **neurophysiological** processes, the **decoding** technique, and the **psychological construct** [6] the authors tested and related to the MS.

RESULTS

We provide preliminary results from the first batch of screened papers (1024 at the present stage). At the time of writing, we have discovered **88 unique EEG-detectable mental states and processes** and **205 different elicitation paradigms**.

MS were clustered by potential use-case into 20 categories (numbers indicate items in the category):

Attention (13), *Mental Workload* (5), *Error* (4), *Vigilance* (5), *Cognitive Flexibility* (2), ***Control*** (10), *Prediction* (7), *Intention* (7), *Learning* (7), *Interest* (4), *Liking* (2), *Joint Attention* (4), *Understanding* (3), *Mental Stress* (2), (*higher*) *Perceptual Processes* (7), *Language* (5), *Motivation* (1), *Navigation/Orientation* (1) & *Emotion* (to be determined).

It should be noted that MS in the same category do not necessarily reflect the same neurophysiological process.

Emotion was by far the most prolifically researched topic, making up more than half of papers in our list (a detailed report of this category is in preparation). For the current report, however, we focus on the ten MS subsumed under **Control**, which we assume are highly relevant for BCI and HCI. (see **Tab. 1** for an overview). The states and processes of the Control category were: *Sense of Agency*, *Response Conflict*, *Response Inhibition*, *Cognitive Control*, *Metacognition*, *Strategic Control*, *Inaptitude*, *Capability*, *Decision Confidence*, *Decision Difficulty* and *Control of Attention*.

Notably, *Inaptitude* and *Capability* (e.g., [7]) were mentioned as MS in different papers. We assumed that they were at opposite ends of a spectrum of the same MS and thus merged them into one: *Inaptitude-capability*. Further, *Response Conflict*, *Response Inhibition*, while similar, were found as distinct MS (see Tab. 1 for definitions). Still, both were elicited with variations of established paradigms such as the Go/Nogo or Flanker tasks.

Table 1: Mental states and processes from the category **Control**.

Mental State or Process	Description	Experimental Paradigm, Elicitation	Tested Construct	Neural Correlates
Sense of Agency	“feelings of control of own actions” [8]	object manipulation in virtually-augmented environment		PSD, phase coherence (both alpha-frequency band)
Response Conflict	“co-activation of the representations for correct and incorrect responses” [9]	Go/Nogo	executive control, response monitoring	N2 prior to response, P300; Theta increase during conflict
		Stroop	executive control, response monitoring	N450 (Medial frontal & parietal negativity), posterior delta/theta
		(cued) Flanker task	resolving of response conflict	aperiodic exponent, theta-frequency band power
Response Inhibition	“suppression of a prepotent response.” [10]	Go/Nogo	executive control, response inhibition	N2, P3
		cued continuous performance test (Go/NoGo conditions)	response inhibition	spatial distribution of the P300 (hyperactivity in the right frontal lobe during NoGo condition)
		Flanker task with go/nogo conditions	inhibition/detection of inappropriate tendency to respond	N2
Cognitive Control	“the ability to actively maintain and use goal-directed information to regulate behavior in a task” [11]	AX-CPT (continuous performance test)	cognitive control, cognitive load	no consensus: e.g., 1) decrease in frontal alpha-frequency band power, 2) connectivity, 3) ERP
Metacontrol	emerging from the interplay of functional systems [12]	Simon Go/NoGo task	persistence-flexibility	aperiodic exponent and offset
Strategic Control	conscious selection of incompatible actions [13]	dynamic reward-learning task	strategic control during exploratory choices	frontal midline theta
Inaptitude-Capability	assumed to be degrees of the same dimension and merged	buy/sell decisions at the stock market		alpha-frequency band power (not significant)
Decision Confidence	subjective certainty after making a decision [14]	Eriksen flanker task, Circle discrimination task		ERN, Pe, P3, delta-, theta-frequency band power
Decision Difficulty	“inherent task difficulty” [15]	visual discrimination task	decision uncertainty	D220 [15] (difficulty-related ERP)
Control of Attention	“task-dependent attentional control settings that determine what sorts of stimulus properties will attract attention” [16]	visual search paradigm	suppression of attention capturing	distractor positivity (Pd)

Where possible, the MS are listed exactly as named in the original literature. Further, we provide the **mental construct** the authors intended to test with the respective paradigm, and the **neural correlates** measured to do so.

ERN = Error Related Negativity; ERP = Event Related Potential; Pe = Error Positivity; PSD = Power Spectral Density

Both time-locked (event-related) potentials (e.g., P300, N2) and continuous signal (e.g., power band changes/distribution) analysis were used to detect control-related MS. Notably, however, these neural correlates were unspecific. For example, the N2, P300 and the theta band were associated with more than one MS from this category. Similarly, several paradigms were used to elicit more than one state (e.g., Go/Nogo, Flanker, or discrimination tasks).

DISCUSSION

We present first results from our endeavor to create a comprehensive list of all described EEG-detectable mental states, including information on how they are elicited and decoded, based on existing literature. Once completed, this list may serve both as a **basis for defining a MDMS** as well as a **reference for any future EEG studies** related to mental states and processes.

A closer look at the *Control* category revealed that some neural correlates were not exclusive to a specific MS, which was expected due to the high degree of similarity of the states within a category, but will likely impair the precise discrimination between these MS. This may become more of an issue with neural correlates that are unspecific between different, unrelated categories, which may render the discrimination of the mental states difficult.

Screening the large body of scientific literature, we found that authors often conflated mental states, for example vigilance and attention (for a discussion of the conceptual differences see [17,18]). This might be due to a lack of common knowledge of existing definitions or an unawareness of the differences between states, which, in common language, may be used interchangeably. The availability of an overview such as the one presented here may already help to raise awareness of this issue. Once completed and backed up by the planned data collection, this project might also contribute to a better separation of some MS by including specific and discriminating patterns of neural correlates - if discovered - into their definition.

Finally, we recommend always confirming the elicitation of a MS by collecting subjective measures (e.g., via questionnaires), which studies often failed to do. Without such confirmation, there is no way to ensure that the correct MS is, in fact, elicited.

Limitations: Only Part I of our literature screening was strictly systematic. The list was supplemented with entries from other sources, such as the community of active researchers (#EEGManyLabs) to ensure a more comprehensive coverage.

CONCLUSION

Once completed, this report will provide an up-to-date and comprehensive summary of all EEG-detectable mental states, which is currently lacking. It may, thus, serve as an **overview of described MS, their elicitation and detection** for any scientist working with mental

states and EEG, particularly in the field of HCI. Further, the availability of this list may contribute to a **more consistent application of the terms** used to describe mental states, which are often conflated between existing pieces of literature. Although neural correlates were unspecific in some cases, in other cases it might prove useful to supplement definitions by including neural correlates as a distinguishing factor of the MS. This should be explored upon the completion of the project.

FUNDING

NAFAS is funded with €30 million by the German federal agency Agentur für Innovation in der Cybersicherheit - "Innovation for Cybersecurity" (Cyberagentur).

REFERENCES

- [1] Krol, L. R., & Zander, T. O. PROJECT NAFAS: ANNOUNCEMENT AND BRIEF OVERVIEW. In: Proceedings of the 9th Graz Brain-Computer Interface Conference 2024. Graz, 2024.
- [2] Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7), 523-534.
- [3] Krol, L. R., Andreessen, L. M., & Zander, T. O. (2018). Passive brain-computer interfaces: a perspective on increased interactivity. In *Brain-computer interfaces handbook* (pp. 69-86). CRC press.
- [4] McGowan, J., Straus, S., Moher, D., Langlois, E. V., O'Brien, K. K., Horsley, T., ... & Tricco, A. C. (2020). Reporting scoping reviews—PRISMA ScR extension. *Journal of clinical epidemiology*, 123, 177-179.
- [5] Pavlov, Y. G., et al. (2021). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex*, 144, 213-229.
- [6] Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health psychology review*, 11(2), 130-134.
- [7] Eskins, D., & Sanders, W. H. (2011). The multiple-asymmetric-utility system model: A framework for modeling cyber-human systems. In 2011 Eighth International Conference on Quantitative Evaluation of SysTems (pp. 233-242). IEEE.
- [8] Sun, W., Huang, M., Wu, C., & Yang, R. (2022). Sense of agency on handheld AR for virtual object translation. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 838-839). IEEE.
- [9] Donkers, F. C., & Van Boxtel, G. J. (2004). The N2 in go/no-go tasks reflects conflict monitoring not response inhibition. *Brain and cognition*, 56(2), 165-176.
- [10] Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: an event-related functional MRI study. *Proceedings of*

- the National Academy of Sciences, 96(14), 8301-8306.
- [11] Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in psychology*, 8, 1482.
- [12] Zhang, C., Stock, A. K., Mückschel, M., Hommel, B., & Beste, C. (2023). Aperiodic neural activity reflects metacontrol. *Cerebral Cortex*, 33(12), 7941-7951.
- [13] Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cerebral cortex*, 22(11), 2575-2586.
- [14] Selimbeyoglu, A., Keskin-Ergen, Y., & Demiralp, T. (2012). What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision. *Clinical Neurophysiology*, 123(6), 1158-1167.
- [15] Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *Journal of Neuroscience*, 26(35), 8965-8975.
- [16] Sawaki, R., & Luck, S. J. (2010). Capture versus suppression of attention by salient singletons: Electrophysiological evidence for an automatic attend-to-me signal. *Attention, Perception, & Psychophysics*, 72(6), 1455-1470.
- [17] Oken, B. S., Salinsky, M. C., & Elsas, S. M. (2006). Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical neurophysiology*, 117(9), 1885-1901.
- [18] van Schie, M. K., Lammers, G. J., Fronczek, R., Middelkoop, H. A., & van Dijk, J. G. (2021). Vigilance: discussion of related concepts and proposal for a definition. *Sleep Medicine*, 83, 175-181.

Finger abduction trajectory prediction from high-density ECoG

Axel Faes¹, Eva Calvo Merino¹, Mani Mirsaedi¹, Anaïs Van Hoylandt², Elina Keirse², Tom Theys², Marc M. Van Hulle¹

¹ KU Leuven - University of Leuven, Department of Neurosciences, Laboratory for Neuro- & Psychophysiology, B-3000 Leuven, Belgium

² KU Leuven - University of Leuven, Department of Neurosciences, Research Group Experimental Neurosurgery and Neuroanatomy, B-3000 Leuven, Belgium

E-mail: marc.vanhulle@kuleuven.com

ABSTRACT: A case study is conducted to demonstrate that not only finger flexion but also finger abduction trajectories can be decoded from high density electrocorticography (ECoG) recordings. Two patients, temporarily implanted with high-density ECoG grids as part of their clinical workup, performed three cued tasks: single finger flexion, finger abduction, and sign language alphabet gestures. The extended Block-Term Tensor Regression (eBTTR) model is trained on simultaneous ECoG and data glove recordings to predict continuous finger trajectories. Performance on the flexion task serves as a baseline for comparison with abduction decoding. The sign language task involves realistic, complex finger movements combining both flexion and abduction. We show that finger abduction trajectories can be decoded from ECoG with precision comparable to single finger flexion, even during sign language gestures that involve joint finger flexions and abductions. To our knowledge, this is the first demonstration that finger abduction trajectories can be decoded from ECoG.

INTRODUCTION

The field of Brain-Computer Interfaces (BCIs) has witnessed significant advancements in motor applications, particularly in restoring or replacing the function of lost or impaired body functions. Notable examples include the electrical stimulation of hand muscles and specific regions of the spinal cord involved in walking, as demonstrated by Lorach et al. [1]. Additionally, BCIs have proven effective in controlling prosthetic hands, arms, exoskeletons, and other effectors [2, 3, 4, 5]. In recent years, electrocorticography (ECoG) has garnered attention within the BCI research community. ECoG involves the placement of electrodes on the exposed cortical surface to record electrical activity. Compared to other implant types, ECoG offers a broader coverage of the cortical surface, enhanced long-term signal stability [6, 7], superior spatial resolution compared to electroencephalography (EEG) [8], and a more extensive spectral bandwidth with increased signal amplitude [9]. Various studies [10, 11, 12, 13, 14, 15, 16] have highlighted the potential of

ECoG signals from the primary motor cortex in decoding continuous arm trajectories, as they exhibit distinctive motor-related spatio-temporo-spectral patterns [17, 11, 18]. Nonetheless, decoding rapid and coordinated finger movements, such as object grasping and hand gestures, remains a challenge, necessitating further algorithmic developments. Pioneering work by Kubanek et al. [19] focused on continuous decoding of individual finger flexions and extensions using local motor potentials (LMP) and spectral amplitudes across five frequency bands.

Previous research has demonstrated successful decoding of finger flexion movements from ECoG signals [16, 20, 21, 22, 23, 24, 25]. Block-Term Tensor Regression (BTTR), as detailed in [26] and [27], was recently introduced for ECoG-based finger flexion decoding. The main algorithmic principle behind BTTR is Tucker decomposition. Block-Term refers to a deflation scheme that generates blocks with varying multilinear Tucker ranks. Originally, BTTR predicted single finger trajectories only, but extended BTTR (eBTTR) lifted this limitation so that coordinated finger movements could be predicted, including finger co-activations i.e. unintentional finger movements [28]. While, in principle, one could stack BTTR models of individual fingers, it will fall short in addressing the temporal overlap and spatial sparsity ECoG signals exhibit during coordinated finger flexions [29].

The aforementioned studies focused on finger flexions, which is useful for applications such as grasping a cup. However, when aiming for restoring hand dexterity, finger abduction becomes important as well. Although the accurate prediction of finger abduction trajectories from (high-density) ECoG could have a significant impact on the development of hand motor BCIs, it has received less attention.

Some studies tangentially refer to finger abduction as part of two-dimensional finger movements. In [30] an intracortical BCI was developed to decode attempted handwriting movements from micro-electrode implants in a paralyzed individual's hand knob area. Patients were instructed to imagine they were holding a pen and were writing on (ruled) paper. In [13] middle finger movements were decoded from ECoG to control a

cursor in two dimensions. These movements involve both finger flexion and abduction. However, the said studies do not directly investigate the decoding of finger abduction.

Hu and colleagues ([31]) investigated finger movement recognition via high-density surface electromyography (HD-sEMG). Their study involved 20 healthy subjects and, using a 22-sensor dataglove, had them perform various finger movements including finger flexion and finger abduction. They found that finger abduction could be decoded similarly to finger flexions.

In order to investigate finger abductions from high-density ECoG, a case study was conducted involving subjects performing three different tasks: a cued single finger flexion task, a cued sign language gesture alphabet task, and a cued finger abduction task. The cued single finger flexion task served as a reference to compare the performance of finger flexion decoding with finger abduction decoding. By analyzing the ECoG data, the study aimed to determine if finger abductions could be decoded with a similar level of accuracy as finger flexions. Furthermore, the sign language gesture alphabet task provided an opportunity to investigate the decoding of complex finger movements that involve both flexions and abductions. The dataset used in this study consisted of ECoG recordings from two patients who performed all gestures of the Flemish Sign

Language Alphabet, as well as explicit finger abduction movements. The eBTTR algorithm was employed as the decoding model [28]. The findings demonstrate that single finger abductions as well as complex hand gestures from the Sign Language Alphabet can be decoded from ECoG. The results indicate that finger abductions can be decoded with a similar level of precision as single finger flexions. This suggests that ECoG signals hold promise for predicting finger abduction trajectories, expanding the potential applications of BCIs for controlling hand prosthetics and exoskeletons and for promoting neurorehabilitation.

MATERIALS AND METHODS

Dataset and tasks

The dataset originates from two patients undergoing treatment of refractory neuropathic facial pain with epidural electrocorticography (ECoG) grids placed on part of the somatosensory and motor cortex. Ethics approval for this study was granted by the Ethics Committee Research UZ/KU Leuven (EC Research). The first patient had an epidural ECoG grid implanted on the right hand-knob area, while the second patient had a grid placed on the left subcentral gyrus and Broca area, and one on the right hand-knob area (refer to Table 1, and Figures 1-3).

Table 1: Patient Information Leuven dataset

	Subject 1	Subject 2
Age	58	41
Gender	Female	Female
Handedness	Right	Right
Region of pain	Trigeminal neuralgia (left)	Temporomandibular joint disorder, bilateral pain mostly in jaws.
Number of grids	1	2
Type of grid	Inc Medtronic Inc Specify® SureScan® MRI 5-6-5 grid (in-line spacing 4.5 mm, row spacing 1.0 mm, electrode size 1.5 mm x 4.0 mm)	2X Medtronic Inc Specify® SureScan® MRI 5-6-5 grid, in-line spacing 4.5 mm, row spacing 1.0 mm, electrode size 1.5 mm x 4.0 mm
Implanted hemisphere	Right	Left and right
ECoG Grid location	Medial frontoparietal, covering M1 and S1, face and lip area	Left: Lateral frontoparietal, inferior parietal cortex and Broca (Brodmann inferior area 44, inferior frontal gyrus pars opercularis left), Right: Medial frontoparietal, covering M1, S1 hand knob & PMDc
Total number of included electrodes	16	32
Number of electrodes over M1	3	Left: 2, Right: 3
Number of electrodes over S1	4	Left: 5, Right: 4

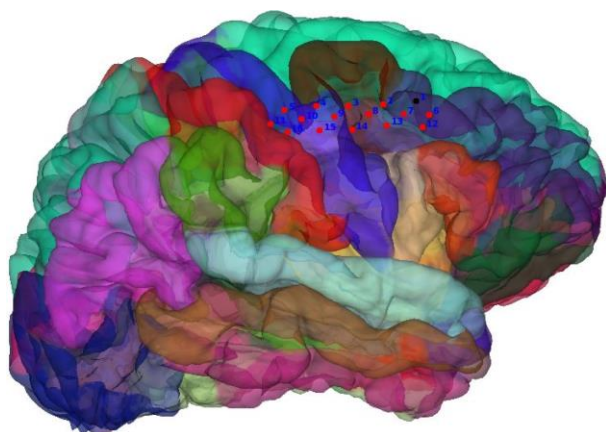


Figure 1: Visual representation of the ECoG grid placement on the right hand-knob area of subject 1.

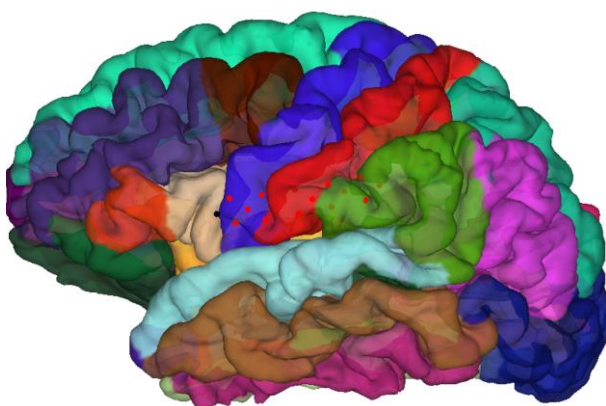


Figure 2: Visual representation of the ECoG grid placement on the left subcentral gyrus and Broca area of subject 2.

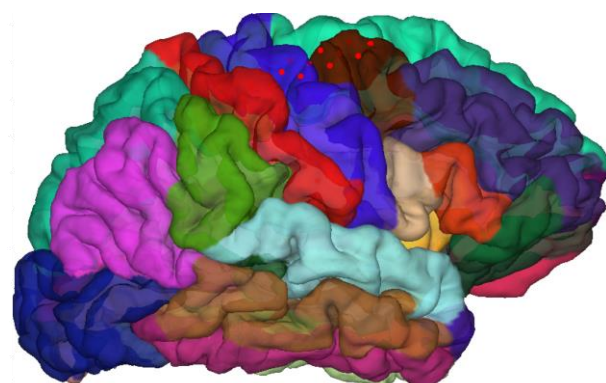


Figure 3: Visual representation of the ECoG grid placement on the right hand-knob area of subject 2.

Three tasks were performed (subject 2 participated in tasks 2 and 3 only). Task 1 involved single finger movements, in which case subject 1 received a cue to start flexing a specific finger with its trajectory measured using a 5DT 14 Ultra data glove (5DT Inc., Irvine, USA). On average, subject 1 was instructed to flex the prompted finger once for 1 to 2 seconds, followed by a rest period lasting 2 seconds. A total of 150 trials were conducted in a single session lasting 600

seconds, with 30 trials for each finger. In Task 2, both subjects were cued to replicate each of the 26 letters of the Flemish sign language alphabet, cued in randomized order using a hand avatar, with the finger trajectories being measured using the said data glove. Given the location of the experiment in Flanders (Belgium), the Flemish Sign Language alphabet was utilized. In Task 3, the two subjects were cued again using a hand avatar, to perform a specific finger abduction pattern, i.e., spreading of two adjoining fingers as wide as possible, again with the finger trajectories being recorded using the same data glove. A total of 120 trials were conducted in a single session lasting 480 seconds, with 30 trials for each abduction. On average, subjects were directed to perform the specified abduction (using an avatar) once for 1 to 2 seconds, followed by a rest period lasting 2 seconds.

Processing

Notch filters, centered at 50 and 100 Hz, were employed to eliminate power line noise. Subsequently, a visual inspection of the data was conducted to identify and eliminate faulty channels. A channel was flagged as problematic if it displayed an unstable or unchanging signal. Following this, Common Average Reference (CAR) re-referencing was applied by subtracting the average signal from all signals [32]. Finally, the ECoG signals were structured into a 4th-order ECoG tensor $\mathbf{X} \in \mathbb{R}^{\text{Samples} \times \text{Channels} \times \text{Frequencies} \times \text{Time}}$. In this representation, Samples corresponds to the size of vector $\mathbf{Y}$ (recorded from the data glove), and Channels reflects the number of electrodes remaining after the removal of faulty channels. Frequencies encompasses 16 components, including 15 gamma band powers and the local motor potential (LMP). The gamma band powers were extracted in the 60 to 130 Hz range (high-frequency bands) using 10 Hz frequency bands with 5 Hz overlap, by bidirectional third-order Butterworth band-pass filters [33]. The LMP was determined by extracting amplitudes in the 0.1 to 1.5 Hz range. Time comprises 10 instances derived from the most recent 1-second epoch, achieved by downsampling each of the 16 band-pass filtered ECoG signals to 10 Hz. The preprocessing steps are as follows:

- Normalize (i.e., z-score) data glove trajectories independently for each finger $\mathbf{Y} \in \mathbb{R}^{\text{Samples} \times \text{Fingers}}$.
- Apply epoch selection for each channel using the past second for band-pass filtering, followed by a downsampling to 10 Hz. This process yields the frequency and time dimensions for each sample and channel.
- Normalize (i.e., z-score) the tensor according to the channel and frequency bands of the training set, with the same parameters applied to the test set.
- Merge the channel matrices into $\mathbf{X} \in \mathbb{R}^{\text{Samples} \times \text{Channels} \times \text{Frequencies} \times \text{Time}}$.

The resulting preprocessed data is then utilized in extended Block Term Tensor Regression (eBTTR), a regression model specifically designed to accommodate the multilinear nature of human intracranial recordings

(see [28] for a complete description of the decoder). The model relies on recursive Tucker decomposition combined with automatic component extraction [27, 28]. A 5-fold cross-validation approach was used to determine the optimal number of Tucker blocks. For each of the three tasks, a separate eBTTR model is trained. This also allows for cross-testing, i.e. use an eBTTR model trained on task 1 (single finger) to decode task 2 (joint finger).

Model performance is expressed in terms of the Pearson correlation coefficient between the data glove recorded finger trajectories and the predicted ones. The test dataset was partitioned into five non-overlapping blocks. For assessing the significance of differences in average accuracies across various fingers and subjects, we employed the Wilcoxon signed-rank test (two-tailed) [34]; a result with a p-value smaller than 0.05 was considered significant.

RESULTS AND DISCUSSION

Per-task analysis

Table 2 summarizes the finger flexion results of Task 1 and 2. Note that Task 1 was performed by subject 1 only. Task 2 considers all gestures of the sign language alphabet, thus going beyond the selection of 4 gestures of the Utrecht experiment [35].

For Task 1, Subject 1 shows higher correlation coefficients for the thumb, index and ring fingers compared to the other fingers. For task 2, both subjects exhibit varying correlation coefficients across different fingers, with mostly higher coefficients for Subject 2.

Table 2: Pearson correlation coefficients of cued single finger movement and finger gesture trajectories.

Methods	Thumb	Index	Middle	Ring	Pinky
Single Finger subj.1	0.31±.05	0.45±.07	0.12±.04	0.23±.02	0.16±.01
Finger gesture subj.1	0.42±.05	0.32±.07	0.38±.04	0.23±.02	0.12±.01
Finger gesture subj.2	0.32±.05	0.46±.07	0.49±.04	0.32±.02	0.33±.01

Table 3 summarizes the finger abduction results of Task 3. The results reveal distinctive correlation patterns for specific finger pairs, with Subject 2 exhibiting overall stronger correlations, especially for the Middle-Ring and Ring-Pinky pairs. Comparing the Task 2 (finger gestures), the same fingers flexed individually also yield stronger correlations. It seems, thus, that finger flexions and finger abductions yield similar decoding accuracies. The exception is the Thumb as it features the lowest abduction decoding accuracy for both subjects but a good flexion decoding performance during finger gesture movements. We expect that this is due to the difficulty to accurately gauge Thumb abductions vs. flexions with the data glove (the abduction of the listed fingers pairs is measured by dedicated sensors unlike the thumb).

Table 3: Pearson correlation coefficients of cued finger abduction movement

Subject	Thumb-Index	Index-Middle	Middle-Ring	Ring-Pinky
Subj. 1	0.14±.05	0.32±.07	0.28±.04	0.22±.01
Subj. 2	0.04±.05	0.36±.08	0.37±.04	0.37±.01

Cross-task analysis

Since subject 1 performed both Task 1 and 2, a cross-task analysis was performed. Hereby, an eBTTR model was trained on the data of Task 1 and tested on Task 2 and vice-versa. The results are summarized in Table 4. The performance of the model trained on single finger movements is significantly lower compared to the performance of the model trained on joint finger movements (see Table 2).

Table 4: Pearson correlation coefficients of cued single finger movement and finger gesture trajectories (for Subject 1)

Methods	Thumb	Index	Middle	Ring	Pinky
Train T1/Test T2	0.22±.06	0.12±.02	0.21±.05	0.08±.02	0.09±.02
Train T2/Test T1	0.28±.03	0.27±.04	0.07±.01	0.21±.02	0.15±.01

CONCLUSION

We presented a comprehensive investigation into the prediction of finger abduction trajectories from high-density ECoG data. The study demonstrates the feasibility of decoding finger abductions using the eBTTR algorithm and shows a decoding performance that is comparable to finger flexion except for the thumb, possibly due to technical aspects of the used data glove. These findings could contribute to the development of hand prosthetics and exoskeletons and to hand neurorehabilitation.

ACKNOWLEDGEMENTS

This work is supported by research grants received from Horizon Europe's Marie Skłodowska-Curie Action (grant agreement No. 101118964), Horizon 2020 research and innovation programme under grant agreement No. 857375, the special research fund of the KU Leuven (C24/18/098), the Belgian Fund for Scientific Research -- Flanders (G0A4321N, G0C1522N, G031426N, 1102925N), and the Hercules Foundation (AKUL 043).

REFERENCES

- [1] Henri Lorach, Andrea Galvez, Valeria Spagnolo, Felix Martel, Serpil Karakas, Nadine Interling, Molywan Vat, Olivier Faivre, Cathal Harte, Salif Komi, et al. Walking naturally after spinal cord injury using a brain–spine interface. *Nature*, pages 1–8, 2023
- [2] Leigh R Hochberg, Mijail D Serruya, Gerhard M Friehe, Jon A Mukand, Maryam Saleh, Abraham H Caplan, Almut Branner, David Chen, Richard D Penn, and John P Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006
- [3] Jennifer L Collinger, Brian Wodlinger, John E Downey, Wei Wang, Elizabeth C Tyler-Kabara, Douglas J Weber, Angus JC McMorland, Meel Velliste, Michael L Boninger, and Andrew B

- Schwartz. High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet*, 381(9866):557–564, 2013
- [4] Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick Van Der Smagt, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375, 2012
- [5] B Wodlinger, JE Downey, EC Tyler-Kabara, AB Schwartz, ML Boninger, and JL Collinger. Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations. *Journal of neural engineering*, 12(1):016011, 2014
- [6] Wei Wang, Jennifer L Collinger, Alan D Degenhart, Elizabeth C Tyler-Kabara, Andrew B Schwartz, Daniel W Moran, Douglas J Weber, Brian Wodlinger, Ramana K Vinjamuri, Robin C Ashmore, et al. An electrocorticographic brain interface in an individual with tetraplegia. *PloS one*, 8(2):e55344, 2013
- [7] Ewan S Nurse, Sam E John, Dean R Freestone, Thomas J Oxley, Hoameng Ung, Samuel F Berkovic, Terence J O'Brien, Mark J Cook, and David B Grayden. Consistency of long-term subdural electrocorticography in humans. *IEEE Transactions on Biomedical Engineering*, 65(2):344–352, 2017
- [8] Kai J Miller, Larry B Sorensen, Jeffrey G Ojemann, and Marcel Den Nijs. Power-law scaling in the brain surface electric potential. *PLoS Comput Biol*, 5(12):e1000609, 2009
- [9] Richard J Staba, Charles L Wilson, Anatol Bragin, Itzhak Fried, and Jerome Engel Jr. Quantitative analysis of high-frequency oscillations (80–500 Hz) recorded in human epileptic hippocampus and entorhinal cortex. *Journal of neurophysiology*, 88(4):1743–1752, 2002
- [10] Zenas C Chao, Yasuo Nagasaka, and Naotaka Fujii. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey. *Frontiers in neuroengineering*, 3:3, 2010
- [11] Leigh R Hochberg Tonio Ball, Markus Kern, Isabella Mutschler, Ad Aertsen, and Andreas Schulze-Bonhage. Signal quality of simultaneously recorded invasive and non-invasive eeg. *Neuroimage*, 46(3):708–716, 2009
- [12] Jeroen CW Siero, Dora Hermes, Hans Hoogduin, Peter R Luijten, Nick F Ramsey, and Natalia Petridou. Bold matches neuronal activity at the mm scale: A combined 7 t fmri and ecog study in human sensorimotor cortex. *Neuroimage*, 101:177–184, 2014
- [13] Leigh R Hochberg Gerwin Schalk, Kai J Miller, Nicholas R Anderson, J Adam Wilson, Matthew D Smyth, Jeffrey G Ojemann, Daniel W Moran, Jonathan R Wolpaw, and Eric C Leuthardt. Two dimensional movement control using electrocorticographic signals in humans. *Journal of neural engineering*, 5(1):75, 2008
- [14] Takufumi Yanagisawa, Masayuki Hirata, Youichi Saitoh, Haruhiko Kishima, Kojiro Matsushita, Tetsu Goto, Ryohei Fukuma, Hiroshi Yokoi, Yukiyasu Kamitani, and Toshiki Yoshimine. Electrocorticographic control of a prosthetic arm in paralyzed patients. *Annals of neurology*, 71(3):353–361, 2012
- [15] Elizabeth A Felton, J Adam Wilson, Justin C Williams, and P Charles Garell. Electrocorticographically controlled brain-computer interfaces using motor and sensory imagery in patients with temporary subdural electrode implants: report of four cases. *Journal of neurosurgery*, 106(3):495–500, 2007
- [16] Alim Louis Benabid, Thomas Costecalde, Andrey Eliseyev, Guillaume Charvet, Alexandre Verney, Serpil Karakas, Michael Foerster, Aurélien Lambert, Boris Morinière, Neil Abroug, et al. An exoskeleton controlled by an epidural wireless brain-machine interface in a tetraplegic patient: a proof-of-concept demonstration. *The Lancet Neurology*, 18(12):1112–1122, 2019
- [17] Nicholas R Anderson, Tim Blakely, Gerwin Schalk, Eric C Leuthardt, and Daniel W Moran. Electrocorticographic (ecog) correlates of human arm movements. *Experimental brain research*, 223(1):1–10, 2012
- [18] David T Bundy, Mrinal Pahwa, Nicholas Szrama, and Eric C Leuthardt. Decoding three-dimensional reaching movements using electrocorticographic signals in humans. *Journal of neural engineering*, 13(2):026021, 2016
- [19] Kubanek, Kai J Miller, Jeffrey G Ojemann, Jonathan R Wolpaw, and Gerwin Schalk. Decoding flexion of individual fingers using electrocorticographic signals in humans. *Journal of neural engineering*, 6(6):066001, 2009
- [20] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (hops): a generalized multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673, 2012
- [21] Andrey Eliseyev and Tetiana Aksenova. Penalized multi-way partial least squares for smooth trajectory decoding from electrocorticographic (ecog) recording. *PloS one*, 11(5):e0154878, 2016
- [22] Rémi Flamary and Alain Rakotomamonjy. Decoding finger movements from ecog signals using switching linear models. *Frontiers in neuroscience*, 6:29, 2012
- [23] Nanying Liang and Laurent Bougrain. Decoding finger flexion from band-specific ecog signals in humans. *Frontiers in neuroscience*, 6:91, 2012.
- [24] Zuoguan Wang, Qiang Ji, Kai J Miller, and Gerwin Schalk. Decoding finger flexion from electrocorticographic signals using a sparse gaussian process. In 2010 20th International conference on pattern recognition, pages 3756–3759. IEEE, 2010
- [25] Ziqian Xie, Odelia Schwartz, and Abhishek Prasad. Decoding of finger trajectory from ecog using deep learning. *Journal of neural engineering*, 15(3):036009, 2018
- [26] Flavio Camarrone, Mariana P Branco, Nick F Ramsey, and Marc M Van Hulle. Accurate offline asynchronous detection of individual finger movement from intracranial brain signals using a novel multiway approach. *IEEE Transactions on Biomedical Engineering*, 68(7):2176–2187, 2020
- [27] Axel Faes, Flavio Camarrone, and Marc M Van Hulle. Single finger trajectory prediction from intracranial brain activity using block-term tensor regression with fast and automatic component extraction. *IEEE Transactions on Neural Networks and Learning Systems*, submitted 2022
- [28] A Faes and MM Van Hulle. Finger movement and coactivation predicted from intracranial brain activity using extended block-term tensor regression. *Journal of Neural Engineering*, 19(6):066011, 2022
- [29] Gang Pan, Jia-Jun Li, Yu Qi, Hang Yu, Jun-Ming Zhu, Xiao-Xiang Zheng, Yue-Ming Wang, and Shao-Min Zhang. Rapid decoding of hand gestures in electrocorticography using recurrent neural networks. *Frontiers in neuroscience*, 12:555, 2018
- [30] Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, 2021
- [31] Xuhui Hu, Aiguo Song, Jianzhi Wang, Hong Zeng, and Wentao Wei. Finger movement recognition via high-density electromyography of intrinsic and extrinsic hand muscles. *Scientific Data*, 9(1):373, 2022
- [32] Collin D Binnie. Clinical neurophysiology: EEG, paediatric techniques and applications, neurophysiology, special volume 2. Elsevier Health Sciences, 2003
- [33] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966
- [34] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992
- [35] Martin G Bleichner, Zachary V Freudenburg, Johannus Martijn Jansma, Erik J Aarnoutse, Mariska J Vansteensel, and Nick Franciscus Ramsey. Give me a sign: decoding four complex hand gestures based on high-density ecog. *Brain Structure and Function*, 221:203–216, 2016

EEG-Based Detection of Continuous and Discrete Hand Movements for Cursor Control

Markus R. Crell¹, Gernot R. Müller-Putz^{1,2}

¹Institute of Neural Engineering, Graz University of Technology, Graz, Austria

²BioTechMed Graz, Graz, Austria

E-mail: gernot.mueller@tugraz.at

ABSTRACT: Continuous cursor control for non-invasive brain-computer interfaces (BCIs) has recently received widespread attention. However, the usability of such systems for patients with e.g., neuromuscular diseases remains limited while a selection mechanism is not simultaneously provided. We therefore attempt to asynchronously detect and discriminate between continuous and discrete hand movements to enable intuitive cursor control and click possibilities for cursor BCIs. Our results show that discrimination between discrete and continuous motions is possible with high accuracy and a low rate of false positive detections. We propose that the introduced method can be applied to non-invasive cursor control BCIs to add a click functionality to the cursor movement.

INTRODUCTION

An increasing share of everyday communication occurs through digital platforms, where effective interaction depends on the ability to move towards and select targets on a screen. For people with neuromuscular conditions such as amyotrophic lateral sclerosis (ALS), the ability for voluntary movement declines, resulting in an incapability to execute such movements. Brain-computer interfaces (BCIs) offer a promising approach to restore computer access under such circumstances. Using invasive recordings, the control of a computer cursor has been achieved recently [1, 2] and demonstrated the importance of such systems for the patient. However, invasive BCIs require surgery, which may not be feasible for all patients and requires an extensive preparation period prior to the implantation. Non-invasive, EEG-based BCIs offer an alternative for these patients and can be implemented with substantially less preparation effort [3]. In the non-invasive field, prior work has demonstrated the successful typing of words using a simple hand-movement-based click [4, 5] or paradigms such as code-modulated visual evoked potentials [6].

The detection of discrete hand movements from EEG can be achieved from low-frequency neural correlates of motion, so called movement-related cortical potentials (MRCPs). MRCPs are changes in the EEG especially over

central motor regions that emerge prior to and during a movement. They show a gradual, negative shift beginning up to two seconds prior to movement onset, followed by a return to baseline [7]. Since the occurrence of MRCPs is independent of visual cues, these signals are well suited to detect self-paced motions. Prior work has demonstrated the successful asynchronous detection of MRCPs with the help of machine learning methods such as shrinkage linear discriminant analysis (sLDA) [4, 8, 9] or random forest classifiers [10].

Recent studies have also focused on the continuous control of computer cursors from non-invasive neural signals. These studies demonstrated that continuous motion and direction can be decoded successfully from electroencephalography (EEG) data [11–14]. While some studies demonstrated successful decoding of continuous movement from low-frequency components [14, 15], other studies showed high frequencies to encode continuous movement speed [16, 17].

Specifically, changes in oscillatory components relative to (continuous) movement occur in mu and beta frequency bands. Such changes in oscillatory components relative to a specific event have been described as event-related (de-)synchronization (ERD/S). During movement planning and execution, oscillatory power in the mu and beta bands shows event-related desynchronization (ERD) that persists until the movement ends [18, 19]. At termination, this desynchronization is followed by a transient increase in beta synchronization before returning to baseline, a phenomenon known as post-movement beta synchronization. Prior studies have also identified this pattern for continuous hand movements and utilized it to decode continuous hand movement [20]

Besides recent advances for cursor control BCIs, the practical use of continuous movement decoding for cursor control remains limited when systems provide only continuous cursor movement without a reliable selection mechanism. While dwell-time selection is possible based on cursor stopping, in everyday usage, dwell-time selection often slows communication. Based on prior work [4, 8, 16], we therefore investigate the discriminability between continuous and discrete hand motions to enable simultaneous cursor control and selection through clicks.

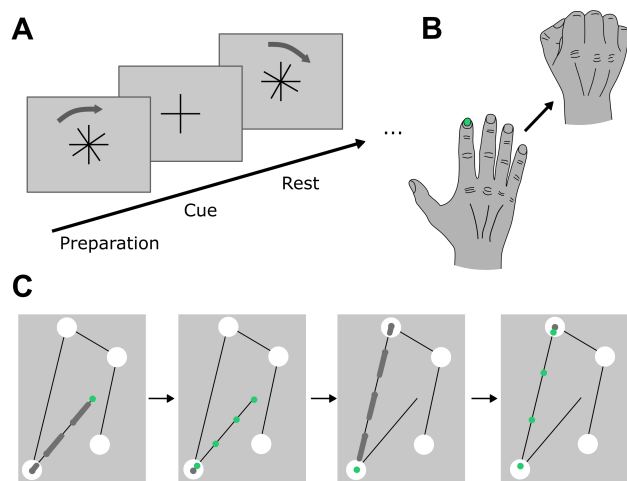


Figure 1: Paradigms and hand motion for calibration data collection. (A) Rotational cross used to collect discrete movement data. The overlap of the rotational cross with the static cross (cue) prompts the participant to execute the specified hand movement. (B) The hand movement that is executed whenever the two crosses overlap. (C) Continuous movement paradigm. The green cursor is controlled by the movement of the index finger of the participant. The grey cursor prompts the movement between the white dots (targets).

We aim to achieve a similar rate of true positives and false positives for the click detection as well as a similar continuous movement detection accuracy for the combined decoding as for individual tasks. We hypothesize that a differentiation between discrete, short and continuous motions can be achieved by leveraging different frequency components of the EEG based on the prolonged ERD/S patterns during continuous movement compared to the occurrence of low-frequency MRCPs at the movement onset of discrete motions.

MATERIALS AND METHODS

Experimental Paradigm: Five participants (3 male, 2 female) with a mean age of 29.2 ± 3.4 years took part in two consecutive EEG measurements. At the beginning of each measurement, participants positioned their right hand on a table with the palm facing downwards.

The first paradigm (adopted from [8, 21]) consisted of a static fixation cross overlaid by a rotating cross (Figure 1 A). The rotating cross rotated at different speeds. Participants were instructed to execute a short hand motion whenever two crosses overlapped (cue). This motion consisted of a flexion of the fingers and the thumb of the right hand as well as the wrist, followed by a return to the initial resting position (Figure 1 B). The first paradigm comprised three runs, each of which had a duration of 334 s. During each run, participants executed 100 discrete hand motions. The inter-trial interval ranged between 2.5 and 4.75 s.

The second paradigm comprised six runs of ten continuous start-and-stop movements each. A start-and-stop

movement (Figure 1 C) was designed around a fixed number of targets with black lines between the targets showing the intended movement pattern. Targets and trajectories together formed the target-trajectory. The timing of the movement was cued by a grey cursor, while participants controlled a green cursor that was mapped to the screen based on a tracked marker on the participants index finger. Both the grey and the green cursor started in the middle of the screen. When the grey cursor moved towards the first target position, participants waited until the grey cursor had reached the first target. The grey cursor paused its movement at the target and the participant moved the green cursor along the black line towards the first target. The grey cursor then continued towards the next target, during which the participants waited at the previous position. This procedure was repeated until the grey and green cursor reached the last target, after which a different target-trajectory was displayed. Participants started each trial at a fixed starting position and returned to that position as soon as a new target-trajectory was displayed. The number of targets (three to seven) as well as their position on the screen was pseudo-randomly selected, resulting in differing trial durations. However, the target-trajectories were similar for each participant. On average, trials had a duration of 30.6 ± 5.8 s. Out of the duration of one trial, on average 18% were spent executing a continuous movement while the rest was spent waiting for the next motion.

Prior to the two paradigms, participants executed an eye-movement paradigm [22] during which they performed left/right and up/down eye movements. The data collected during this paradigm was used to eliminate eye artifacts from the EEG.

Data Acquisition and Feature Extraction: We measured the EEG of participants from 60 electrodes positioned according to a standard 10-10 montage. Electrooculography (EOG) was recorded from four electrodes positioned on the outer canthi of the eyes as well as above and below the left eye. EOG and EEG was recorded using a biosignal amplifier (BrainAmp, Brain Products GmbH, Gilching, Germany) at a sampling rate of 500 Hz. Participants were further equipped with a motion capture marker positioned on the index finger of the right hand. Movements of the marker were recorded at a rate of 30 Hz using a custom motion capture system.

EEG and EOG data was first bandpass filtered between 0.3 and 70 Hz (Butterworth filter, 4th order). Eye-artifacts were eliminated using the SGEYESUB algorithm [22] and EOG channels were subsequently removed from the data. To extract the movement onset of discrete hand motions and the start and stop of continuous movements, we extracted the movement speed of the index finger from the positions of the motion capture marker. We then applied a threshold to the speed and aligned the movement data to the neural data.

Two sets of features were extracted: low-frequency features were obtained by lowpass-filtering the EEG to 3.5 Hz (Butterworth filter, 4th order) and downsampling

the data to 10Hz. Broadband-frequency features were obtained by lowpass-filtering the EEG to 40Hz (Butterworth filter, 4th order) and downsampling the data to 125Hz.

Discrete Movement Detection (First Paradigm): The detection of discrete movements from the first paradigm focused on the detection of the movement onset of the discrete motion as obtained with the motion capture system. We attempted to detect movement onsets from windows of low-frequency features using an sLDA. To improve detection performance, we applied a hyperparameter search over the windowlength w_l and the additional detection delay d as described in [4]. For each participant, we calibrated a model for the detection of movement onsets on the data from the first paradigm. We selected the optimal windowlength and detection delay from $w_l \in [0.1; 2.0]$, $d \in [-\min(1.1, w_l); 0.0]$ based on the maximum Matthews correlation coefficient (MCC) [23] of the validation set. The MCC is a metric for the evaluation of binary classification and is robust to class-imbalance, as is often the case in asynchronous BCIs. Movement onset predictions were regarded as true positives if they occurred up to -0.2 s prior or up to 0.8 s after the actual movement onset. We further applied a refractory period of 1.0 s after each movement onset detection to eliminate accidental multiple detections. For each model calibrated during the hyperparameter search, the detection probability threshold p was tuned to allow for not more than 2.0 false positives per minute (FP/min) based on the validation set. We selected the highest p for which this condition was true. The final detection performance was evaluated on the test set by calculating the MCC, the true positive rate (TPR) and FP/min. To obtain a higher confidence on the test set performance, we applied a three fold cross-validation procedure. Specifically, we generated three folds of the same size where each fold was treated as training, validation or test set exactly once. The same three-fold cross-validation procedure was also applied to all subsequent models.

Continuous Movement Decoding (Second Paradigm): We decoded continuous movement from windows of the broadband-frequency features of the data from the second paradigm. The ground truth of the continuous movement was obtained from the movement speed of the right hand measured with the motion capture system. For this, we employed a CNN-LSTM neural network. The architecture of the CNN was based on the EEGNet architecture [24], which has been designed to extract relevant frequency bands and has been shown to enable speed decoding for continuous movements [16]. The inclusion of an LSTM has been shown to increase performance in EEG-based continuous movement tasks [25]. Similar to the discrete movement detection, we applied a three-fold cross validation procedure and hyperparameter search over the windowlength w_l , the additional detection delay d and the detection threshold p . We calibrated a model for each participant on the data from the second paradigm and selected the optimal hyperparameter

ters based on the MCC of the validation set. The final performance was evaluated on the test set by calculating the MCC, the TPR and the false positive rate (FPR). In contrast to the discrete movement detection, we did not apply a refractory period.

Combined Discrete and Continuous Movement Decoding: The detection of discrete and the decoding of continuous movements was achieved by applying the described procedures for both models to the data of both paradigms. Specifically, we applied each the discrete detection and continuous decoding on the separate datasets from the first and second paradigm. For the discrete movement detection, the labels of the continuous movement in the dataset from the second paradigm were eliminated. For the continuous movement decoding, the labels of the discrete motions in the dataset from the first paradigm were eliminated. For the combined continuous and discrete models, the complete three-fold cross-validation procedure and hyperparameter search was repeated for both models on the combined datasets and each participant. We finally applied the optimal models to the same test sets and evaluated the performance of the combined predictions. In addition to the refractory period of the discrete detection, we also applied a refractory period to the continuous movement decoding after each discrete movement onset detection. Specifically, we inhibited any continuous movement detection through a refractory period of 3.0 s after the detection of a discrete movement onset.

We compared the performance of the discrete movement detection on the data from the first paradigm with the performance on the combined data and similar for the continuous movement decoding. Additionally, we evaluated the rate of FP/min and the MCC, TPR and FPR of the combined decoding approach solely on the parts of the data corresponding to the first and second paradigm, so that we were able to directly compare the performance.

RESULTS

When calibrating models for the discrete movement detection on the data from the first paradigm and evaluating the model on data from the first paradigm (*discrete movement - single decoding*), we achieved an average TPR of the discrete movement detection of $49.09 \pm 20.55\%$ (mean $\pm$ standard deviation) with an average rate of FP/min of 2.83 ± 0.48 . The mean MCC was 0.58 ± 0.18 . For the continuous movement decoding calibrated and evaluated only on data from the second paradigm (*continuous movement - single decoding*), the average MCC was 0.59 ± 0.10 . With the combined approach, models for the discrete movement detection that were calibrated and evaluated on the data from both paradigms (*discrete movement - combined decoding*) achieved an average TPR of $48.03 \pm 21.32\%$ (-1.06% compared to *discrete movement - single decoding*) with an average rate of FP/min of 1.92 ± 0.63 (-0.91 compared to *discrete movement - single decoding*) and an MCC of 0.53 ± 0.21 (-0.05 compared to *discrete movement - single decoding*).

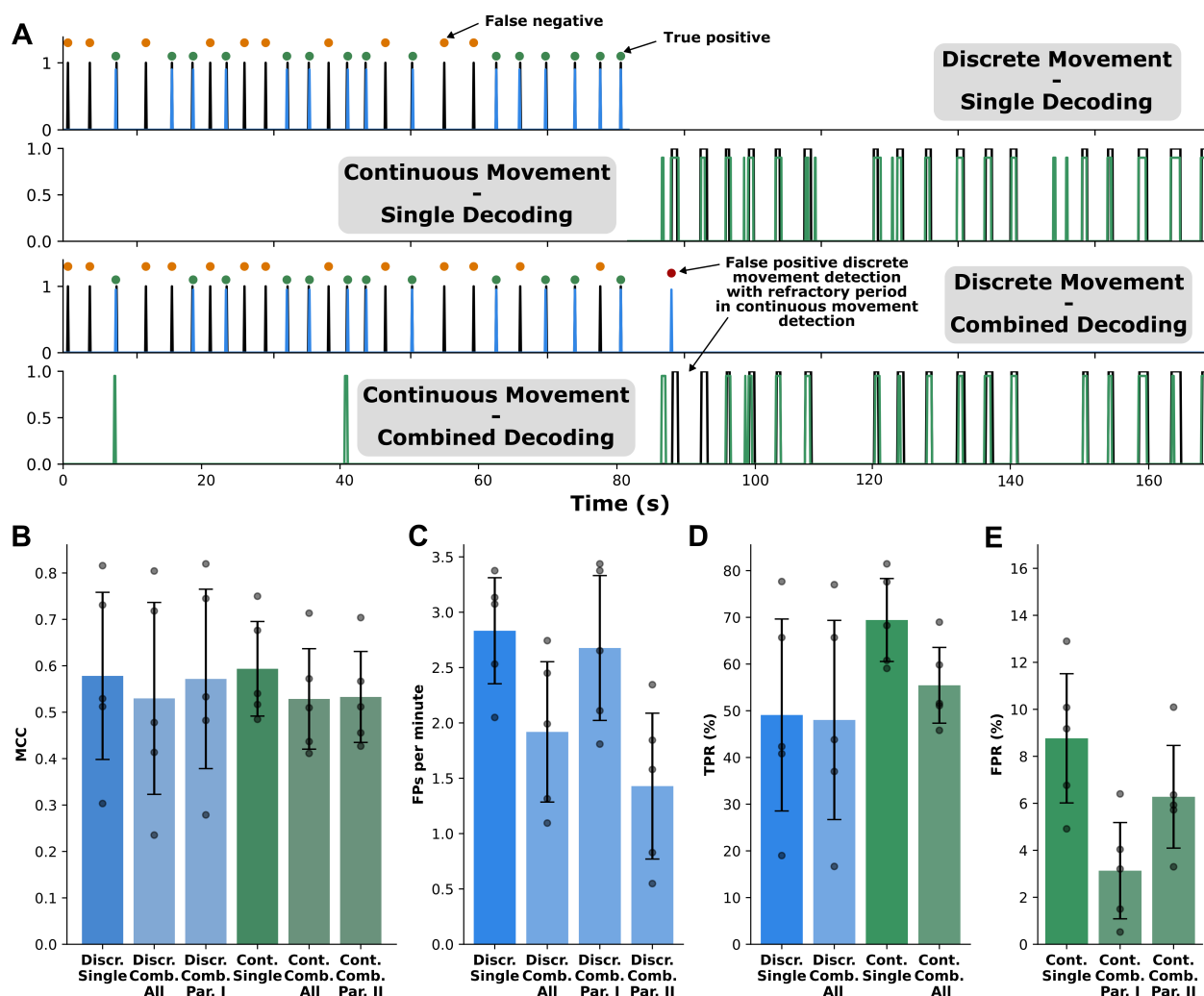


Figure 2: (A) Executed (black) and decoded (blue: discrete, green: continuous) discrete and continuous movement patterns. The upper graph shows parts of the data from the first paradigm with movement detections based on the optimal model calibrated only on the data from the first paradigm (discrete movement - single decoding). The second graph shows parts of the data from the second paradigm with continuous movements decoded based on the optimal model calibrated only on the data from the second paradigm (continuous movement - single decoding). The third and fourth graphs show the combined decoding output from the discrete and continuous models. Output from the discrete movement detection model is shown in the upper graph together with the executed motions while output from the continuous model is shown in the lower graph together with the executed movements. False positives (red) and false negatives (yellow) are shown for the discrete movement detection as dots. (B) MCCs for the different models (discrete and continuous) calibrated on the single or combined data and evaluated on the different paradigms. Black dots indicate values for individual participants. (C) Rate of FP/min for the discrete model calibrated and evaluated on the different paradigms. (D) TPR of the discrete and continuous model calibrated and evaluated on the different paradigms. (E) FPR of the continuous model calibrated and evaluated on the different paradigms.

ing). When evaluating these models only on the data from the first paradigm (*discrete movement - combined decoding - paradigm I*), we achieved an average rate of FP/min of 2.68 ± 0.65 (-0.15 compared to *discrete movement - single decoding*) and an MCC of 0.57 ± 0.19 (-0.01 compared to *discrete movement - single decoding*). The rate of FP/min of the *discrete movement - combined decoding - paradigm II* was on average 1.43 ± 0.66 .

For the continuous movement decoding, models that were calibrated and evaluated on the data from both paradigms (*continuous movement - combined decoding*)

achieved an average MCC of 0.53 ± 0.11 . When evaluating these models only on the data from paradigm two (*continuous movement - combined decoding - paradigm II*), we achieved an average MCC of 0.53 ± 0.10 (-0.06 compared to *continuous movement - single decoding*).

For the *continuous movement - single decoding* model, the TPR, i.e. the rate of actual movement that was identified as movement, was $69.43 \pm 8.88\%$, while the FPR, i.e. the rate of rest that was falsely identified as movement, was $8.77 \pm 2.75\%$. For the *continuous movement - combined decoding - paradigm I*, the FPR was

$3.13 \pm 2.05\%$. For the *continuous movement - combined decoding - paradigm II*, the TPR was $55.45 \pm 8.11\%$ (-13.98% compared to *continuous movement - single decoding*) while the FPR was $6.28 \pm 2.18\%$ (-2.49% compared to *continuous movement - single decoding*).

Figure 2 A shows parts of the paradigms and decoding results for one participant. The first two graphs show the detected discrete motions (blue) and the executed movements (black) for the first paradigm (*discrete movement - single decoding*) and the detected continuous movements (green) and executed movements (black) for the second paradigm (*continuous movement - single decoding*). The lower graphs show the decoding results of the combined decoding approach split into the detected discrete movements during both paradigms (*discrete movement - combined decoding*) and the detected continuous movements during both paradigms (*continuous movement - combined decoding*). The reported results for the different models and evaluation modes are shown in Figure 2 B for the MCC, Figure 2 C for the rate of FP/min, Figure 2 D for the TPR and Figure 2 E for the FPR.

DISCUSSION

The simultaneous detection of discrete and continuous hand movements for the utilization in cursor BCIs is of great relevance for the usability of cursor BCIs in motor-impaired patients. The current study shows that the asynchronous detection and differentiation of continuous and discrete movement is possible with only slight decreases of the performance compared to detecting only discrete or continuous movement.

Specifically, the TPR of the discrete movement onset detection was reduced minimally in the combined decoding approach. Similarly, the MCC was reduced only slightly for the discrete detection of movement. However, there was a substantial decrease in detected false positives per minute when evaluating both paradigms together, which was largely driven by a low rate of false positives during the second paradigm. This indicates that the model rarely detected any discrete motions during continuous movements.

The decoding of continuous movement demonstrated a higher reduction in the MCC. Specifically, the TPR dropped by nearly 14% when training the model on data from both paradigms. Simultaneously, the FPR also decreased. The occurrence of falsely decoded continuous movement was especially low during the first paradigm, indicating that discrete motions were rarely mistaken as continuous movements.

Overall, the detection performances of both the discrete and continuous model were only slightly reduced when trained and evaluated on data from both paradigms compared to the single paradigms. False detection rates remained sufficiently low in the corresponding original paradigms and were even reduced during the other paradigm, i.e., the rate of false positive detections of discrete motions in the second paradigm was even lower

than that in the first paradigm and similar for the continuous decoding.

Prior studies have shown that the rate of false positives that was achieved during the discrete motion detection is suitable for spelling five-letter words in a row-column scanner [4]. Thus, we assume that the false positive rates would not hinder the participant during the application of the cursor BCI. Further, the introduced refractory periods of 1 s for discrete movements and 3 s of inhibited continuous movement after a detected discrete movement are sufficiently low to allow for high-performance control of computers.

CONCLUSION

The presented results demonstrate that the asynchronous detection of continuous hand movement and discrete hand motions can be achieved simultaneously with sufficient differentiation between the actions. The low rates of false positive detections of discrete motions during continuous movement and vice versa demonstrates the feasibility of the proposed method. Thus, the proposed method could be utilized in a cursor control BCI to move the cursor and make selections on a screen without needing to hover over the target. We believe that these results are an important step towards a fully functional cursor control BCI that can be applied in patients with neuromuscular conditions.

Limitations of the current study mainly concern the sample size of participants and the separated data collection. Since data for the discrete movements was collected separately from the continuous movements, it is not certain that the decoder performances can be transferred to real-world scenarios. Future work should therefore focus on a paradigm which combines continuous movement and clicks. Further, the online application of the method was not evaluated. This is an important step to identify further issues and test whether the method is suited for an online BCI for patients and will be pursued in future work.

ACKNOWLEDGEMENTS

This project is funded by the European Union's HORIZON-EIC-2021-PATHFINDER CHALLENGES program under grant agreement No 101070939 and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00198.

REFERENCES

- [1] Wilson GH *et al.* Long-term unsupervised recalibration of cursor-based intracortical brain-computer interfaces using a hidden Markov model. *Nature Biomedical Engineering*. 2025;1–19.
- [2] Deo DR, Willett FR, Avansino DT, Hochberg LR, Henderson JM, Shenoy KV. Brain control of bimanual movement enabled by recurrent neural networks. *Scientific Reports*. 2024;14(1):1598.

- [3] Edelman BJ *et al.* Non-invasive Brain-Computer Interfaces: State of the Art and Trends. *IEEE Reviews in Biomedical Engineering*. 2024;PP(99):1–25.
- [4] Crell MR, Kostoglou K, Suwandjieff P, Cruz JRD, Müller-Putz GR. A non-invasive, MRCP-based BCI for online communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2026;PP(99):1–1.
- [5] Scherer R *et al.* Thought-based row-column scanning communication board for individuals with cerebral palsy. *Annals of Physical and Rehabilitation Medicine*. 2015;58(1):14–22.
- [6] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): a literature review. *Journal of Neural Engineering*. 2021;18(6):061002.
- [7] Shibasaki H, Hallett M. What is the Bereitschaftspotential? *Clinical Neurophysiology*. 2006;117(11):2341–2356.
- [8] Crell MR, Kostoglou K, Sterk K, Müller-Putz GR. A novel paradigm for fast training data generation in asynchronous movement-based BCIs. *Frontiers in Human Neuroscience*. 2025;19:1540155.
- [9] Ofner P, Schwarz A, Pereira J, Müller-Putz GR. Upper limb movements can be decoded from the time-domain of low-frequency EEG. *PLoS ONE*. 2017;12(8):e0182578.
- [10] Liu D *et al.* EEG-Based Lower-Limb Movement Onset Decoding: Continuous Classification and Asynchronous Detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2018;26(8):1626–1635.
- [11] Lu X *et al.* Electroencephalography Enables Continuous Decoding of Hand Motion Angles in Polar Coordinates. *Cyborg and Bionic Systems*. 2026;7:0469.
- [12] Mondini V, Sburlea AI, Müller-Putz GR. Towards unlocking motor control in spinal cord injured by applying an online EEG-based framework to decode motor intention, trajectory and error processing. *Scientific Reports*. 2024;14(1):4714.
- [13] Pulferer HS, Ásgeirsdóttir B, Mondini V, Sburlea AI, Müller-Putz GR. Continuous 2D trajectory decoding from attempted movement: across-session performance in able-bodied and feasibility in a spinal cord injured participant. *Journal of Neural Engineering*. 2022;19(3):036005.
- [14] Mondini V, Kobler RJ, Sburlea AI, Müller-Putz GR. Continuous low-frequency EEG decoding of arm movement for closed-loop, natural control of a robotic arm. *Journal of Neural Engineering*. 2020;17(4):046031.
- [15] Kobler RJ, Sburlea AI, Mondini V, Hirata M, Müller-Putz GR. Distance- and speed-informed kinematics decoding improves M/EEG based upper-limb movement decoder accuracy. *Journal of Neural Engineering*. 2020;17(5):056027.
- [16] Crell MR, Müller-Putz GR. Handwritten character classification from EEG through continuous kinematic decoding. *Computers in Biology and Medicine*. 2024;182:109132.
- [17] Lv J, Li Y, Gu Z. Decoding hand movement velocity from electroencephalogram signals during a drawing task. *BioMedical Engineering OnLine*. 2010;9(1):64.
- [18] Pfurtscheller G, Silva FLd. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [19] Pfurtscheller G, Solis-Escalante T. Could the beta rebound in the EEG be suitable to realize a “brain switch”? *Clinical Neurophysiology*. 2009;120(1):24–29.
- [20] Crell MR, Müller-Putz GR. Detection of motion termination from EEG during the execution of continuous hand movement. *Proceedings of the 9th Graz Brain-Computer Interface Conference 2024*. 2024:139–144.
- [21] Suwandjieff P, Müller-Putz GR. EEG Analyses of visual cue effects on executed movements. *Journal of Neuroscience Methods*. 2024;410:110241.
- [22] Kobler RJ, Sburlea AI, Lopes-Dias C, Schwarz A, Hirata M, Müller-Putz GR. Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals. *NeuroImage*. 2020;218:117000.
- [23] Chicco D, Jurman G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21.
- [24] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.
- [25] Chen YF *et al.* Continuous Bimanual Trajectory Decoding of Coordinated Movement From EEG Signals. *IEEE Journal of Biomedical and Health Informatics*. 2022;26(12):6012–6023.

Riemannian approaches for ECoG-based Motor Imagery BCI decoding using spatial covariance matrices : a comparative study

M. Etienne¹, F. Martel¹, S. Bonnet², T. Aksenova¹

¹Univ. Grenoble Alpes, CEA, LETI, Clinatec, Grenoble, France

²Univ. Grenoble Alpes, CEA, LETI, Grenoble, France

E-mail: mathis.etienne@cea.fr

ABSTRACT:

Symmetric positive definite (SPD) covariance matrices have demonstrated strong potential for motor imagery (MI) decoding in EEG-based brain-computer interfaces (BCIs), yet their application to electrocorticography (ECoG) data remains largely unexplored.

In this study, we evaluate and compare multiple classification approaches leveraging the Riemannian geometry of SPD matrices for ECoG-based MI decoding, across two paradigms: upper limb (3-class) and lower limb (3-class) motor imagery. Spatial covariance matrices were estimated from signals filtered in two task-specific frequency bands and used as input features. Three families of classifiers were compared: tangent space mapping (TSM) combined with machine learning models (LR, LDA, RSW-NPLS, LightGBM) and a multilayer perceptron (MLP); classifiers operating directly on the Riemannian manifold (MDM, FgMDM); and an SPD-aware neural network (SPDNetBN). Models were trained and tested offline on data from two implanted patients using lag-corrected balanced accuracy as the performance metric. FgMDM and TSM+RSWPLS achieved the highest performance for the upper limb paradigm (up to 91.6% and 89.5%, respectively), while SPDNetBN demonstrated superior accuracy for the lower limb paradigm (up to 69.4%).

INTRODUCTION

A brain-computer interface (BCI) is a technology that translates a person's brain activity into commands that control external devices. It typically consists of sensors that record neural activity, a signal processing pipeline to decode the recorded signal (often called a decoder), and one or more actuators. BCIs hold significant promise for assistive and rehabilitation applications, particularly for individuals with severe motor disabilities or neurodegenerative diseases..

Electroencephalography (EEG) remains the most widely used recording method for BCIs because of its low cost, non-invasive nature and ease of use. In contrast, electrocorticography (ECoG), although more invasive, offers superior spatial resolution, higher signal-to-noise ratio, broader frequency coverage, and greater long-term signal stability. These advantages make ECoG particularly

well suited for decoding complex motor imagery tasks, including fine movements of the upper and lower limbs [1, 2].

Traditionally, BCI decoding pipelines incorporate a feature extraction step that provides discriminative information to downstream classifiers. Commonly used features include, among others, band-power estimates in sensorimotor rhythms (e.g., μ , β and γ bands), event-related desynchronization/synchronization (ERD/ERS) patterns or time-frequency representations.

Since 2010, a new family of features based on symmetric positive definite (SPD) covariance matrices has gained prominence in EEG-based motor imagery (MI) classification, as they naturally capture spatial correlation patterns across channels [3]. Importantly, SPD matrices lie on a Riemannian manifold endowed with a non-Euclidean geometry, which should be taken into account during decoder learning. This geometric perspective has led to the development of dedicated classification algorithms that respect the manifold structure, demonstrating improved performance over conventional Euclidean methods [4].

Several methodological approaches exist for leveraging this Riemannian geometry: classifiers may operate directly on the manifold using appropriate geometric metrics, SPD matrices can be projected into a tangent space, where conventional machine learning algorithms or neural networks can be applied, or specialized neural network architectures can be designed to preserve manifold structure throughout processing.

While SPD matrix-based approaches have become increasingly popular for EEG, their application to ECoG remains largely unexplored. Unlike EEG, ECoG electrodes are placed directly on the cortical surface, providing more localized coverage and closer proximity to neural sources. These differences in spatial scale may influence the structure of SPD matrices and, consequently, the effectiveness of Riemannian-based analyses.

To date, only methods employing tangent space mapping (TSM) have been evaluated with ECoG data. These include covariance matrices projected via TSM and subsequently decoded with machine learning models such as logistic regression [5], neural networks (LSTM) [6], and approaches combining TSM-projected covariance matrices with more conventional spectro-temporal features where decoding is performed by classical machine learn-

ing models (linear model, Random Forest, Support Vector Machines) or gradient boosting trees (lightGBM) [7]. The tasks addressed in these studies range from hand motor imagery [5] (tetraplegic patient) to actual hand movements [6] (able-bodied monkeys) and fingers movements [7] (able-bodied humans). Manifold-based models and SPD neural network architecture have yet to be explored for ECoG data.

To address this gap, we evaluate several of these approaches on ECoG data acquired during upper and lower limb motor imagery tasks performed by one paraplegic and one tetraplegic patient. We investigate three categories of methods: [i] tangent space mapping (TSM) combined with conventional machine learning classifiers (logistic regression (LR), linear discriminant analysis (LDA)), N-way partial least squares (NPLS), and gradient boosted trees (lightGBM)[7]; [ii] classifiers operating directly on the Riemannian manifold (Minimum Distance to Mean (MDM) and Fisher Geodesic MDM (Fg-MDM)); and [iii] neural network approaches, including both SPD-aware architectures (SPDNetBN) and a conventional network applied to tangent space projections (multilayer perceptron (MLP)). To do so, we compared the decoding performance of each approach using lag-corrected balanced accuracy across two datasets (upper limb and lower limb conditions), each comprising three MI states.

MATERIALS AND METHODS

Experimental data:

Data were collected from two subjects: one tetraplegic (S1) enrolled in the BCI & Tetraplegia clinical trial (ClinicalTrials.gov identifier: NCT02550522), and one paraplegic (S2) enrolled in the STIMO-BSI clinical trial (ClinicalTrials.gov identifier: NCT04632290). Both participants were implanted with two wireless WIMAGINE devices [1] positioned over sensorimotor cortices for long-term epidural electrocorticography (ECoG) recording. Additionally, S2 received an epidural electrical stimulation (EES) pulse generator targeting spinal cord regions involved in leg movement control [2].

Participants performed asynchronous motor imagery tasks during experimental sessions: upper limb movements for S1 (3 classes : idle, left hand translation, right hand translation) and walking for S2 (3 classes : idle, left hip, right hip). Based on the decoder predictions, S1 could control a two-arm neuroprosthetic exoskeleton. S2 could control gait movements through stimulation of the nerves originating in the spinal cord using epidural stimulation. Two excerpts from the recording of cues and on-line predictions are shown in Fig. 1. The WIMAGINE devices recorded ECoG signals at 585 Hz sampling rate. Sessions used for training and testing are resumed in Table 1.

Feature extraction:

We employed spatial covariance estimators, which are widely established in the BCI literature [4]. Covariance

Table 1: Summary of sessions used for training and testing (number of classes, number of sessions and duration).

Paradigm	Classes	Train sessions	Test sessions
S1 (upper limb)	3	3 (191 min)	6 (242 min)
S2 (lower limb)	3	4 (15 min)	21 (125 min)

matrices were computed from filtered ECoG signals to capture spatial correlation patterns across distinct frequency bands. For each trial, signals from selected electrodes were bandpass-filtered using two fifth-order Butterworth filters to isolate frequency-specific neural activity.

Covariance matrix estimation For each trial and spatial configuration, the spatial covariance matrix was estimated separately for each frequency band. Given the filtered signals $\mathbf{X}_k \in \mathbb{R}^{n_c \times T}$ for band $k \in \{1, 2\}$, where n_c denotes the number of electrodes and T the number of time samples, the block covariance matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$ with $m = 2n_c$ is defined as:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 \end{pmatrix}. \quad (1)$$

Each block $\mathbf{C}_k$ is estimated using the Ledoit-Wolf shrinkage estimator [8], which provides a regularized estimate balancing the sample covariance with a scaled identity target:

$$\mathbf{C}_k = (1 - \rho_k)\mathbf{S}_k + \rho_k \frac{\text{tr}(\mathbf{S}_k)}{n_c} \mathbf{I}_{n_c}, \quad (2)$$

where $\mathbf{S}_k = \frac{1}{T-1} \mathbf{X}_k \mathbf{X}_k^T$ is the sample covariance matrix for band k , $\mathbf{I}_{n_c}$ is the identity matrix, and $\rho_k \in [0, 1]$ is the shrinkage parameter estimated as described in [8]. We also investigated a spatial frequency covariance estimate computed across all m channels jointly, allowing non-zero inter-frequency covariance between the two bands. Number of time samples per trial was fixed at $T = 590$ (1 second of data) for S1 and $T = 236$ (400 ms of data) for S2, based on previous studies [1, 2].

Classification:

Importantly, covariance matrices yield SPD matrices that lie on a differentiable Riemannian manifold, where standard Euclidean operations such as arithmetic mean and Euclidean metrics are inappropriate and can lead to sub-optimal classification performance [9].

There are three ways to handle this, taking into account the geometry of the covariance matrix space: use classification algorithms that operate directly on the Riemannian manifold, project the covariance matrices onto an Euclidean tangent space in order to apply conventional classification algorithms, or design specific neural networks that manipulate SPD matrices.

Classification within the Riemannian manifold Riemannian manifold geometry-based classifiers operate directly on the manifold of Symmetric Positive Definite (SPD) matrices by leveraging Riemannian metrics to compute distances and class means. These include the

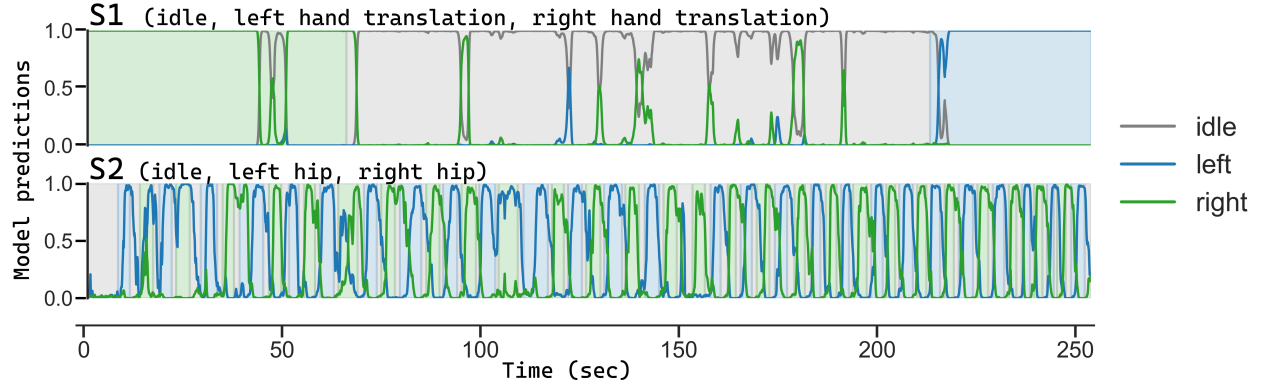


Figure 1: Online prediction excerpt from two sessions (S1, top; S2, bottom). The background shading indicates the desired cue, while the lines represent the model predictions. During each motor cue, S1 performed continuous 3D movement tasks with an exoskeleton [1]. The duration of the cues is therefore longer. For clarity, the display of the idle prediction is disabled for S2.

Minimum Distance to Mean (MDM) and its variant, the Fisher Geodesic MDM (FgMDM) [3]. These classifiers have been shown to be particularly effective for BCI applications, as they inherently respect the geometric structure of covariance matrices [4]. The distance employed in this study is the Affine Invariant Riemannian Metric (AIRM) distance, which operates on the manifold of symmetric positive definite (SPD) matrices. The AIRM distance between two SPD matrices C_1 and C_2 is defined as:

$$\delta_R(C_1, C_2) = \left\| \logm(C_1^{-1/2} C_2 C_1^{-1/2}) \right\|_F, \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\logm(\cdot)$ represents the matrix logarithm.

The MDM classifier assigns each trial to the class whose geometric mean is closest in terms of AIRM distance.

The geometric mean $\bar{C}$ of a set of SPD matrices $\{C_n\}_{n=1}^N$ is computed by minimizing the sum of squared Riemannian distances :

$$\bar{C} = \arg \min_C \sigma(C), \quad \sigma(C) = \sum_{n=1}^N \delta_R^2(C, C_n) \quad (4)$$

Since no closed-form solution exists, the geometric mean is estimated iteratively [9]. An example of the geometric mean for S2 is given in Figure 2.

Classification on tangent space To enable the use of conventional machine learning algorithms that operate in Euclidean space, Tangent Space Mapping (TSM) was applied [9]. TSM projects covariance matrices from the manifold of SPD matrices onto a tangent space of the manifold at a reference point, typically chosen as the geometric mean of the training covariances.

The tangent space feature vector t_i for a given covariance matrix C_i is obtained by computing the logarithmic map at the reference point:

$$t_i = \text{vech} \left[\logm \left(\bar{C}^{-1/2} C_i \bar{C}^{-1/2} \right) \right], \quad (5)$$

where $\text{vech}(\cdot)$ half-vectorizes the upper triangular part of the symmetric matrix into a feature vector of dimension $m(m+1)/2$, preserving ℓ_2 norm.

In this study, multiple state-of-the-art classifiers were applied to the tangent space features. Classical machine learning approaches included Linear Discriminant Analysis (LDA) and Logistic Regression (LR), an effective PLS model for classifying ECoG data in real time (Recursive Sample Weighted – N-way Partial Least Squares (RSW-NPLS) [10]), and an ensemble of gradient boosting trees (lightGBM) [7]. A multilayer perceptron (MLP) with two hidden layers (32 and 16 units) was also evaluated as a simple deep learning baseline. Hyperparameters were tuned on a validation set comprising 20% of the training data.

Classification with SPD Neural Networks Additionally, we investigated SPDNet [11] with Riemannian batch normalization (BN) [12], a specialized neural network architecture designed to operate directly on SPD matrices. SPDNet incorporates Riemannian geometry throughout its layers, utilizing bilinear mapping and rectification operations that preserve the manifold structure. The network concludes with a tangent space mapping layer using the identity matrix as the reference point, projecting features into Euclidean space for final classification. This approach combines the geometric awareness of Riemannian methods with the representational capacity and non-linearity of deep learning.

Hidden Markov Model A Hidden Markov Model (HMM) was applied to all the classifier outputs to improve prediction robustness and temporal consistency [1].

Feature configuration:

Frequency band selection Frequency band selection was guided by analysis of class distinctiveness across the spectrum [13]. This approach quantifies the separability between motor imagery classes by computing the Riemannian class distinctiveness for each frequency

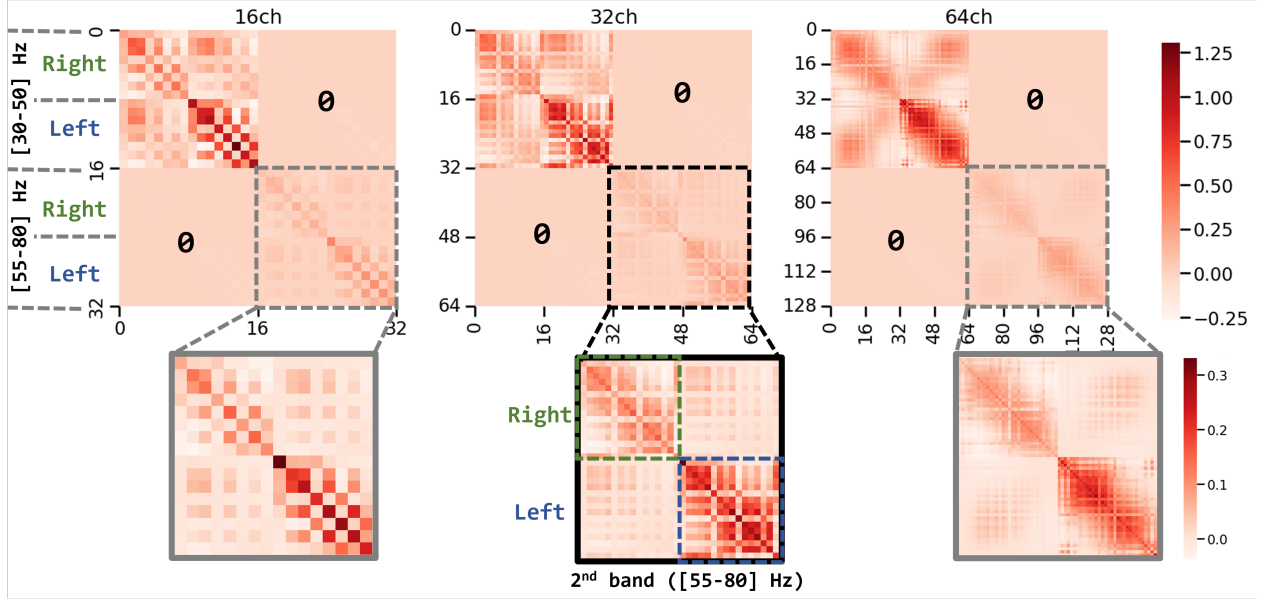


Figure 2: Mean covariance matrices from S2 training data (all classes) for 16, 32, and 64 channels. Each block matrix shows intra-band covariance (top-left: low-frequency; bottom-right: high-frequency). Within each diagonal block: right implant (top-left), left implant (bottom-right), and inter-implant covariances (top-right/bottom-left).

band, identifying bands where spatial covariance matrices (SCMs) differ most significantly between conditions. For a given frequency band f_i , the class distinctiveness between two classes A and B is defined as:

$$\text{classDis}(f_i) = \frac{\delta_R(\bar{C}_A(f_i), \bar{C}_B(f_i))}{\sigma_A(f_i) + \sigma_B(f_i)}, \quad (6)$$

where $\bar{C}_A(f_i)$ and $\bar{C}_B(f_i)$ are the Riemannian means of the SCMs computed from signals band-pass filtered in the frequency band $[f_i - 5 \text{ Hz}, f_i + 5 \text{ Hz}]$, for classes A (resting state) and B (active state, motor classes combined), respectively. The term $\delta_R(\cdot, \cdot)$ denotes the AIRM distance (Eq. 3), while $\sigma_A(f_i)$ and $\sigma_B(f_i)$ correspond to the Riemannian mean absolute deviation within each class and frequency band (defined in Eq. 4).

Higher values of $\text{classDis}(f_i)$ indicate greater discriminability between classes at specific frequency bands.

Based on the class distinctiveness analysis (Figure 3), conducted using only two training sessions, two broader frequency bands were selected for each subject: $[10, 45]$ Hz and $[55, 75]$ Hz for S1, $[30, 50]$ Hz and $[55, 80]$ Hz for S2. These bands are selected for their discriminability and coherence with event-related desynchronization (ERD) below 50 Hz and event-related synchronization (ERS) above 50 Hz observed in ECoG data [14].

Spatial electrode configurations Each WIMAGINE implant contains 64 electrodes arranged in an 8×8 grid, but only 32 of them were activated during the recording period used in this study, due to hardware limitations. Three spatial configurations were evaluated to assess the trade-off between spatial resolution and feature dimension: full electrode coverage with 64 channels (32 per

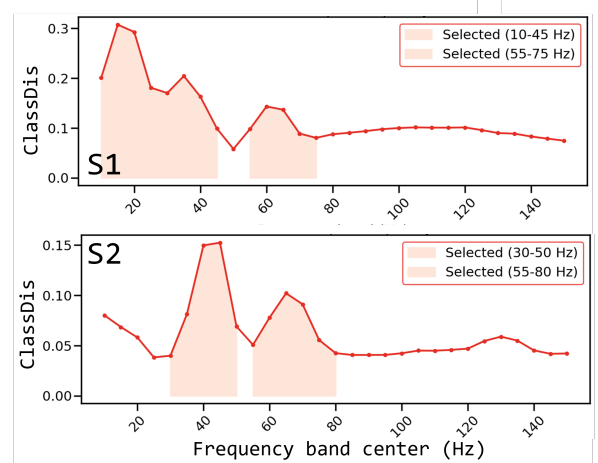


Figure 3: Class distinctiveness across frequencies for S1 and S2 training data.

implant) and spatially subsampled montages with 32 and 16 channels (Figure 4). The electrode subsets were selected randomly while preserving spatial distribution (reduction performed in areas of 4 close neighbours, 1 over 4 for 16, 2 over 4 for 32). The results are averaged over 50 sub-selections per configuration.

Performance evaluation:

We evaluated decoding performance using lag-corrected balanced accuracy (lcBA). For each session, we estimated the latency as the lag maximizing the cross-correlation between labels and predictions, averaged across motor states. Predictions were then shifted by this lag before computing balanced accuracy (average per-class recall).

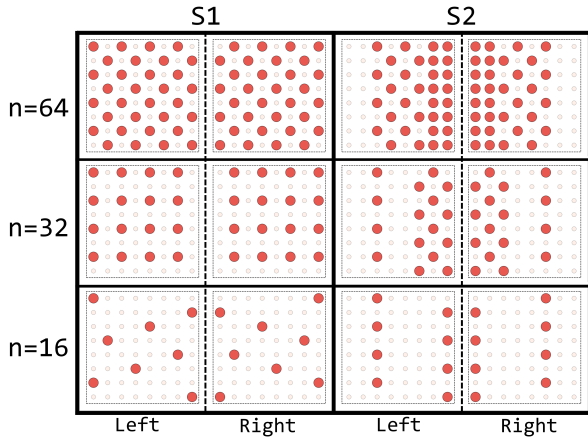


Figure 4: Example of electrode configurations with 64, 32, and 16 channels. Both participants were implanted with two 50 mm WIMAGINE implants (right and left hemispheres), each containing 64 recording electrodes. Red circles indicate active electrodes.

RESULTS

Decoding performance was evaluated using lag-corrected balanced accuracy across models, channel numbers (16, 32, and 64), and tasks (upper and lower limb motor imagery), as summarized in Table 2.

For S1 (upper limbs), the highest performance at 16 channels was obtained with TSM+RSWPLS ($85.4\% \pm 3.6\%$), closely followed by FgMDM ($85.1\% \pm 3.3\%$). At 32 and 64 channels, FgMDM achieved the best results ($88.8\% \pm 2.4\%$ and $91.6\% \pm 1.7\%$, respectively), with TSM+RSWPLS remaining competitive ($88.3\% \pm 2.8\%$ and $89.5\% \pm 2.0\%$). Across all channel configurations, TSM+RSWPLS consistently outperformed other tangent-space-based classifiers (TSM+LR, TSM+LDA, TSM+LightGBM, and TSM+MLP). MDM yielded substantially lower accuracies (58.2% – 58.6%). SPDNetBN achieved moderate results, reaching $84.1\% \pm 1.8\%$ at 64 channels, but remained below FgMDM and TSM+RSWPLS for all configurations.

For S2 (lower limbs), TSM+RSWPLS achieved the highest accuracy at 16 channels ($68.9\% \pm 6.3\%$), closely followed by SPDNetBN ($68.8\% \pm 6.0\%$) and FgMDM ($68.7\% \pm 6.3\%$). Increasing the number of channels did not lead to systematic improvements across models. SPDNetBN yielded the best performance at both 32 and 64 channels ($69.4\% \pm 5.3\%$ and $68.1\% \pm 4.3\%$, respectively), and was the only model to remain stable with increasing channel count. TSM+RSWPLS decreased to $67.2\% \pm 5.7\%$ and $64.9\% \pm 5.2\%$. Other tangent-space-based classifiers and manifold classifiers remained less performing and decreased similarly with the increasing number of electrodes. TSM+LightGBM and TSM+MLP showed relatively stable but lower performances across channel counts, with best results of $60.2\% \pm 4.9\%$ and $60.6\% \pm 5.0\%$ at 32 electrodes, respectively.

The spatio-frequency covariance (with inter-frequency covariance) performance was largely comparable, with

an average difference of -0.5% across all classifiers and channel counts.

DISCUSSION

This offline study provides a comparative evaluation of tangent space mapping (TSM), intrinsic Riemannian classifiers, and SPD-aware neural networks for ECoG motor imagery decoding. The three types of classifiers showed strong potential.

For the lower limbs paradigm, some models show declining performance with increasing electrodes number, likely due to the task's simplicity rendering additional spatial information redundant. This is not observed for the upper limb paradigm. Indeed, during cue periods, the patient performed motor imagery tasks for continuous hand movement (3D). This may have complicated the classification, where the addition of electrodes should be beneficial.

The lower performance observed for paradigm S2 (lower limbs) compared to S1 (upper limb) is likely attributable to task duration. In S2, gait tasks alternated every two to three seconds, whereas S1 maintained a single state for over 30 seconds on average. Since transition periods are inherently harder to classify, more frequent transitions in S2 reduce overall performance.

In practice, compatibility with incremental and online learning is essential, as it enables closed-loop training in which patients receive real-time feedback during learning. This is achievable with some of the models evaluated here, but with adjustments (Table 3).

Furthermore, combining SPD matrices projected into tangent space with other types of features (e.g. time-frequency), here referred to as hybrid features, has been shown to improve decoding performance [7]. However, this is not possible natively for classifiers on the Riemannian manifold and for neural networks designed for SPD matrices (Table 3). These considerations should guide model selection. Among the evaluated approaches, RSW-NPLS offers the best trade-off across these criteria.

Limitations:

In this study, only one covariance matrix is used as feature. This was chosen for compatibility with MDM, FgMDM and SPDNetBN. Only two frequency bands were selected in order to limit the feature size. To improve performance, one can consider increasing the number of matrices and/or the number of frequencies studied. In this case, more complex architectures can be used, such as graph neural networks on SPD manifolds. Other types of matrices, such as cross-frequency covariance matrices, may also be considered [6].

Furthermore, this study was conducted on only two patients. Increasing the amount of data will enable the robustness of the proposed models to be validated.

Finally this work is an offline study. It therefore lacks the human factor, which could influence the result (closed-loop BCI). An online study of these models for real-time experiments is needed to confirm the observed results.

Table 2: Lag-corrected balanced accuracy (mean $\pm$ std) across models, channel numbers, and tasks (in %).

Participant	Channels	TSM+RSWPLS	TSM+LR	TSM+LDA	TSM + LightGBM	TSM+MLP	MDM	FgMDM	SPDNetBN
S1 (upper limbs)	16 (x50)	85.4 $\pm$ 3.6	79.0 $\pm$ 2.6	78.9 $\pm$ 2.3	80.4 $\pm$ 1.5	80.9 $\pm$ 2.2	58.3 $\pm$ 3.7	85.1 $\pm$ 3.3	79.7 $\pm$ 2.4
	32 (x50)	88.3 $\pm$ 2.8	81.3 $\pm$ 2.0	82.2 $\pm$ 1.8	82.5 $\pm$ 1.4	82.0 $\pm$ 2.1	58.6 $\pm$ 3.2	88.8 $\pm$ 2.4	82.7 $\pm$ 2.1
	64 (x1)	89.5 $\pm$ 2.0	81.1 $\pm$ 1.6	81.8 $\pm$ 1.5	83.7 $\pm$ 1.1	81.8 $\pm$ 1.7	58.2 $\pm$ 3.2	91.6 $\pm$ 1.7	84.1 $\pm$ 1.8
S2 (lower limbs)	16 (x50)	68.9 $\pm$ 6.3	65.5 $\pm$ 5.5	65.2 $\pm$ 5.2	59.2 $\pm$ 5.3	60.2 $\pm$ 5.4	60.7 $\pm$ 10.0	68.7 $\pm$ 6.3	68.8 $\pm$ 6.0
	32 (x50)	67.2 $\pm$ 5.7	62.5 $\pm$ 4.9	60.5 $\pm$ 4.3	60.2 $\pm$ 4.9	60.6 $\pm$ 5.0	58.1 $\pm$ 9.7	65.0 $\pm$ 7.0	69.4 $\pm$ 5.3
	64 (x1)	64.9 $\pm$ 5.2	57.9 $\pm$ 3.9	50.3 $\pm$ 2.7	59.8 $\pm$ 5.2	60.3 $\pm$ 4.4	53.7 $\pm$ 8.4	60.3 $\pm$ 7.4	68.1 $\pm$ 4.3

Table 3: Model compatibility with online/incremental learning (streaming updates without full data storage, at the cost of adaptations/approximations) and hybrid features.

Model compatibility with	Streaming / Incremental Learning	Hybrid Features
TSM+RSWPLS	Yes	Yes
TSM+LR	Yes	Yes
TSM+LDA	Yes	Yes
TSM+LightGBM	No	Yes
MDM	Yes	No
FgMDM	Limited	No
SPDNetBN	No	No
TSM+MLP	No	Yes

CONCLUSION

This study compare multiple classification techniques leveraging the geometry of SPD-based features for ECoG motor imagery BCI. The features used are spatial covariance matrices of signals from two frequency bands.

Machine learning classifiers applied after tangent space mapping, classifiers operating within the manifold, and neural networks designed for SPD matrices were compared. To this end, the models were trained and tested offline on two paradigms : upper limb and lower limb motor imagery classification. All three approaches yielded satisfactory results on at least one of the two paradigms. Additional experiments should be conducted to confirm these results, with more participants performing other motor tasks. Then the models will need to be adapted, if possible, to function online for closed-loop learning.

ACKNOWLEDGMENT

The authors would like to thank all the members of the BCI project at Cinatec (CEA-Leti, Grenoble Alpes University Hospital, and University Grenoble Alpes), composed of the multidisciplinary technical and clinical teams. We also thank all the members of the NeuroRestore team for conducting the STIMO-BSI clinical trial and collecting the data used in this study. This study was funded by CEA (CFR and Bluesky-EXPLORE+).

REFERENCES

[1] Benabid AL et al. An exoskeleton controlled by an epidural wireless brain-machine interface in a tetraplegic patient: A proof-of-concept demonstration. *The Lancet Neurology*. 2019;18(12):1112–1122.
[2] Lorach H et al. Walking naturally after spinal cord injury using a brain-spine interface. *Nature*. 2023;618(7963):126–133.

[3] Barachant A, Bonnet S, Congedo M, Jutten C. Riemannian geometry applied to BCI classification. In: *LVA/ICA 2010 - 9th International Conference on Latent Variable Analysis and Signal Separation*. Springer: Saint-Malo, France, Sep. 2010, 629–636.
[4] Lotte F et al. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *Journal of Neural Engineering*. 2018;15.
[5] Liu D et al. “Reclaiming Hand Functions after Complete Spinal Cord Injury with Epidural Brain-Computer Interface.” Accessed: May 13, 2025. [Online]. Available: <https://www.medrxiv.org/content/10.1101/2024.09.05.24313041v6> preprint.
[6] Eyvazpour R, Farrokhi B, Erfanian A. A general model based on Riemannian manifold for stable decoding movement trajectory from ECoG signals. *iScience*. 2026;29(2):114521.
[7] Yao L, Zhu B, Shoaran M. Fast and accurate decoding of finger movements from ECoG through Riemannian features and modern machine learning techniques. *Journal of Neural Engineering*. 2022;19(1):016037.
[8] Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*. 2004;88(2):365–411.
[9] Barachant A, Bonnet S, Congedo M, Jutten C. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*. 2013;112:172–178.
[10] Souriau R, Martel F, Aksenova T. Recursive sample weighted – N-way partial least squares for brain-computer interface decoder online learning with class imbalance. *Biomedical Signal Processing and Control*. 2026;113:109025.
[11] Huang Z, Gool LV. A Riemannian Network for SPD Matrix Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017;31(1).
[12] Brooks D, Schwander O, Barbaresco F, Schneider JY, Cord M. Riemannian batch normalization for SPD neural networks. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.
[13] Yamamoto MS, Lotte F, Yger F, Chevallier S. Class-distinctiveness-based frequency band selection on the Riemannian manifold for oscillatory activity-based BCIs: Preliminary results. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Jul. 2022, 3690–3693.
[14] Moly A et al. An adaptive closed-loop ECoG decoder for long-term and stable bimanual control of an exoskeleton by a tetraplegic. *Journal of Neural Engineering*. 2022;19(2):026021.

Why Simple Models Still Matter: Adaptive sLDA for Non-Stationary EEG Decoding

J. Egger¹ and G. R. Müller-Putz^{1,2}

¹ Institute of Neural Engineering, Graz University of Technology, Graz, Austria

² BioTechMed Graz, Austria

E-mail: gernot.mueller@tugraz.at

ABSTRACT: Electroencephalography (EEG)-based brain-computer interfaces (BCIs) are strongly affected by non-stationarities, causing covariate shifts between calibration and deployment and leading to performance degradation over time. While deep learning approaches address this issue, they typically require large training datasets that are often unavailable in practical BCI scenarios. We therefore investigate lightweight online adaptation strategies for shrinkage linear discriminant analysis (sLDA), a simple yet robust classifier for low-data EEG settings. We evaluate three unsupervised methods: exponential moving average feature normalization (*FNorm*), pooled mean adaptation of the classifier bias (*PMean*), and pooled mean with covariance adaptation (*PMean+Cov*). Across six sessions from 20 participants, without FP/min control, true-positive rates rose by 14.27% (*FNorm*), 29.13% (*PMean*) and 5.87% (*PMean+Cov*), while false-positives per minute changed to 3.71 (+2.19) for *FNorm*, 11.95 (+10.44) for *PMean*, and 1.05 (-0.47) for *PMean+Cov* (not statistically significant). *PMean+Cov* achieved the best balance between sensitivity and false positives, making it most suitable for real-time BCI speller applications.

INTRODUCTION

Locked-in syndrome is a condition in which individuals remain fully conscious but lose most voluntary motor control, severely limiting communication [1]. Within the INTRECOM project, we develop a fully implantable brain-computer interface (BCI) for real-time motor and speech decoding to enable home-based communication, particularly via a speller interface. As the implantable system is still under development, non-invasive electroencephalography (EEG) provides a practical platform to investigate decoding strategies, adaptation mechanisms, and real-time constraints. In this context, reliable control and low false-positive rates (FPR) are crucial, as unintended activations directly impair usability and increase user frustration.

A key challenge in EEG-based decoding is non-stationarity, which induces covariate shifts [2] between calibration (training) and deployment (test) data due to changes in e.g., cognition or feedback [3-5]. Although deep learning [6,7] and domain adaptation methods [8,9] address this problem, they often rely on larger

datasets and higher computational resources than linear models, limiting their suitability for short-calibration BCI settings. Adaptive extensions of linear discriminant analysis (LDA) have therefore been widely studied [10-14]. Here, we build on shrinkage LDA (sLDA) without modifying its shrinkage structure, preserving its robustness to low signal-to-noise ratios, inter-channel correlations, and limited training data in high-dimensional EEG [15,16]. The novelty of our approach lies in extending sLDA with lightweight online adaptation while retaining its regularization framework. We propose three online adaptation strategies: (1) exponential moving average (EMA) for feature normalization to track distributional shifts, a model-agnostic approach applicable beyond sLDA, (2) pooled mean adaptation to update the classifier bias while keeping weights fixed, building on prior LDA-based work [10,12], and (3) pooled mean and covariance adaptation to update, which updates both bias and classifier weights and represents the main methodological contribution, extending covariance adaptation to the shrinkage LDA while preserving its regularization structure under limited data conditions. To our knowledge, such an adaptation has not been previously investigated for sLDA in BCI settings. These methods enable efficient online adaptation with minimal complexity while explicitly controlling FPR for real-time speller application.

This work focuses on a controlled comparison within the sLDA framework to isolate the effect of lightweight adaptation under identical model complexity. A direct comparison to alternative adaptive models in terms of data efficiency and computational cost is beyond the scope of this study and remains subject to future work. We hypothesize that lightweight online adaptation of sLDA improves classification performance under non-stationary conditions while maintaining low FPRs, making it suitable for real-time speller applications.

MATERIALS AND METHODS

A. Dataset

The dataset [17] in use comprised hand gesture recordings from 22 healthy right-handed adults (20-40 years). EEG was recorded with 60 active gel-based

electrodes (actiCAP, Brain Products GmbH, Germany) placed according to the 10-10 system and four EOG electrodes for artifact correction (500 Hz sampling rate). A synchronized video motion capture system tracked a marker on the right index finger to determine movement onset. Participants completed six sessions between 2 p.m. and midnight, spaced two hours apart, including cognitive tasks, a standardized meal, and fatigue-inducing activities. Each session contained an eye movement paradigm for artifact correction [18] and a cue-based task with four right-hand gestures (fist, pistol, pincer, and the “Y” sign in American Sign Language). Each session consisted of eight runs, yielding 64 trials per gesture per participant. Two participants were excluded due to incomplete data.

B. Processing of recordings

Signal processing and decoding was performed in Python. Raw EEG data were visually inspected, and noisy channels were interpolated using spherical splines [19]. Powerline noise (50 Hz and harmonics) was removed with a 2nd-order IIR Butterworth comb filter, followed by a non-causal 0.3-70 Hz 4th-order Butterworth bandpass filter. Eye artifacts were removed using the SGEYESUB algorithm [18] trained on a separate eye-movement dataset. Signals were low-pass filtered at 3 Hz, down-sampled to 16 Hz (f_s), and re-referenced to a common average. Frontal and temporal channels were excluded due to poor signal quality.

C. Static Classification

Classification was performed using shrinkage LDA (sLDA) on sliding windows of low-frequency EEG features, employing the Ledoit-Wolf shrinkage for covariance regularization [20]. Binary labels for the continuous data were assigned following Crell et al. [21]. Windows of length w (from timepoint t_0 to t_0+w) were shifted in 0.0625 s steps ($1/f_s$). Using the motion capture data to extract the timepoint of movement onset t_{onset} , windows ending at $t_{onset}+d$ were labelled as *movement*, all others were labelled as *no movement*. The delay d incorporated post-onset information to reduce false positives (FPs). Hyperparameters ($w \in [0.5, 2.0]$ s; $d \in [0, w]$, step 0.1 s) were optimized per participant via 4-fold cross-validation on session 1 by maximizing the mean Matthews Correlation Coefficient (MCC). The final static model was retrained on session 1 with optimal parameters and applied to later sessions. The probability threshold p was determined on session 2 (0.1-0.9, step 0.1) to maximize MCC under the constraint of ≤ 1 false positive per minute (FP/min). Predictions within ± 0.25 s of the true movement onset were counted as true positives (TP).

D. Adaptive Classification

a. Adaptive Feature Normalization (FNorm)

Poor generalization arises from EEG non-stationarities

that shift feature distributions between training and test data [2]. A common strategy to address such covariate shifts is data normalization, which aligns these probability distributions [22]. We applied this using EMA filters to estimate the time-varying mean m_t and standard deviation s_t of the $(1 \times n)$ feature vectors using exponentially decaying weights, following Hehenberger et al. [3], further referred to as *FNorm* strategy. At each time point t , the incoming feature vector x_t is used to update the estimates of m_t and s_t . Based on the estimates from the previous time step ($t-1$), the update rule for each component i is defined as follows:

$$m_t^{(i)} = (1 - \eta)m_{t-1}^{(i)} + \eta x_t^{(i)} \quad (1)$$

$$s_t^{(i)} = \sqrt{(1 - \eta)s_{t-1}^{(i)2} + \eta(x_t^{(i)} - m_t^{(i)})^2} \quad (2)$$

The smoothing parameter η weighs the contribution of the current feature vector on normalization. Initial estimates (m_0 and s_0) were taken from session 1. The smoothing parameter η was defined as

$$\eta = 1 - (1 - p)^{\frac{1}{k}} \quad (3)$$

where k represents the number of observations whose cumulative weights account for a fraction p of the total EMA weights, i.e., the exponentially decaying coefficients that determine the contribution of past samples in the moving average. A participant-wise grid search on session 2 evaluated $p \in [10, 20]$ (step size 2) and $k \in [10, 70]$ (step size 5), selecting parameters that maximized $MCC - \lambda \cdot FP/min$, with $\lambda = 0.05$.

The normalized feature vector was then calculated as:

$$\tilde{x}_t = (x_t - m_t)\Sigma_t^{-\frac{1}{2}} \quad (4)$$

$$\text{with } \Sigma_t^{-\frac{1}{2}} = \text{diag}(s_t) = \begin{bmatrix} s_t^{(1)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_t^{(n)} \end{bmatrix} \quad (5)$$

b. Pooled Mean Adaptation (PMean)

Another unsupervised adaptation strategy, following Vidaurre et al. [10,12], updates the classifier parameters directly. Test data from session 2 to 6 were normalized using session 1 statistics. We adapted the pooled class mean μ using EMA filters (see Eq. (1)). Updating μ continuously adjusts the classifier bias term b

$$\mu = (\mu_1 + \mu_2)/2 \quad (6)$$

$$b = w^T \mu \quad (7)$$

while keeping the weights w fixed. In this paper, this approach is abbreviated by *PMean*.

c. Adaptive Pooled Mean and Sample Covariance (PMean+Cov)

Extending pooled mean adaptation approach, which

adjusts only the bias term b while keeping the weights w fixed, we incorporated a covariance update mechanism to also adapt the weights w [10,12]. Every minute, buffered unlabelled samples were used to estimate second-order statistics and update the inverse covariance matrix Σ_{t-1}^{-1} , yielding an unsupervised adaptation scheme. Instead of estimating class-specific covariances, a global second-order moment $E[xx^T]$ was computed. The inverse covariance matrix was updated recursively using a Woodbury identity-based formulation, allowing efficient incorporation of the new batch statistics without explicit recomputation or inversion of the full covariance matrix. To ensure numerical stability, Tikhonov regularization via diagonal loading was applied by adding a small-scaled identity matrix to the covariance estimate (or its inverse during recursive updates). The weights w were then recomputed as:

$$w_t = \Sigma_t^{-1}(\mu_1 - \mu_2) \quad (8)$$

This allowed continuous adaptation of both the decision boundary and bias b while retaining the session 1 model as initialization. In contrast to Eq. (1), separate update coefficients η_μ and η_Σ were introduced for the pooled mean and covariance adaptation, respectively, and were determined via a grid search on session 2 with η_μ and $\eta_\Sigma \in [0, 0.25]$. We refer to this adaptation strategy as *PMean+Cov*.

d. FP/min Control

To minimize FP, a probability threshold p adaptation mechanism was applied. The threshold p was used to convert classifier output probabilities into binary decisions (prediction = 1 if $P(\text{class}=1) \geq p$, otherwise 0). Predictions were generated run by run, and FP/min were computed after each run. The probability threshold p was adjusted via a feedback rule based on the deviation from a predefined target (FP/min = 1): $p \leftarrow p + \alpha \cdot (\text{FP/min} - \text{target})$, where $\alpha = 0.01$. Updates were only applied when the deviation exceeded a tolerance margin and were constrained to a fixed range ($p \in [0.1, 0.9]$). The threshold was increased when FP/min exceeded the target and decreased otherwise. This threshold controller was applied across all adaptive methods as a shared control layer. However, it represents a supervised post-hoc mechanism and is not intended for real-time application. Nonetheless, it was included to demonstrate the ability of the adaptive methods to reduce FP under controlled evaluation conditions.

E. Statistical Analysis

Statistical comparisons between the static sLDA classifier and adaptive methods were performed using linear mixed-effects models (LMMs) to account for repeated measurements across participants and sessions. For each performance metric (MCC, TPR, FP/min), the metric value was modelled as a function of method and session, with participants included as a random effect. Session-wise performance slopes were derived using session number as a continuous predictor, representing average performance change per session and capturing temporal trends. Analyses were conducted separately for adaptive configurations with and without FP/min control. Within each session, pairwise contrasts between the static sLDA baseline and each adaptive method were computed from fixed effects. One-sided p-values were calculated in the expected direction of improvement (adaptive > static for MCC and TPR; adaptive < static for FP/min), reflecting the prior, hypothesis that adaptive updates improve performance under non-stationary conditions. Effect sizes were obtained from estimated contrasts normalized by the residual standard deviation of the model. P-values were corrected for multiple comparisons across sessions using the Benjamini-Hochberg procedure. To compare adaptive strategies, group-level analyses were conducted using analogous LMMs that included all strategies simultaneously. These models enabled direct statistical comparisons between adaptive methods, the static sLDA baseline, and their interaction with session, while accounting for multiple testing via false discovery rate.

RESULTS

The performance of the static sLDA and adaptive classification methods across sessions for the different metrics, with and without FP/min control, is shown in Fig. 1. Tables 1 and 2 summarize the averaged performance across sessions of the adaptive methods relative to the static sLDA baseline. Tab. 1 corresponds to Fig. 1A (without FP/min control), and Tab. 2 to Fig. 1B (with FP/min control). For FP/min, negative differences to the baseline are desirable. Tab. 3 reports the best-performing adaptive method per metric and control condition. Tab. 4 presents the session-wise slopes of the best adaptive methods compared to static sLDA baseline to evaluate stability across sessions.

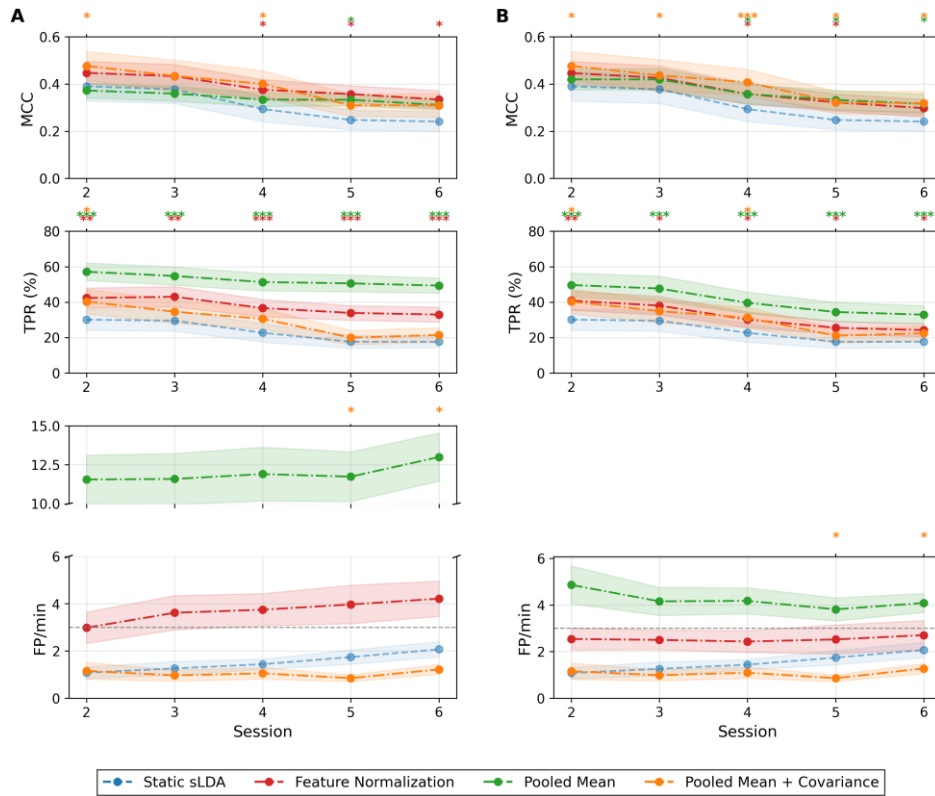


Figure 1: Performance metrics of the static sLDA classifier (blue) and the adaptive methods evaluated **A** without FP/min control and **B** with FP/min control. Color-coded asterisks denote statistically significant differences relative to the static baseline after Benjamini-Hochberg correction (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

Table 1: Performance metrics of each adaptive method without FP/min control. For both static and adaptive approaches, session-averaged values, minimum and maximum across sessions are reported along with the corresponding difference to the static baseline and the associated statistical significance corrected for multiple comparisons.

Metric	Method	Static			Adaptive			Delta			p-value
		Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	
MCC	<i>FNorm</i>				0.39	0.33	0.45	0.08	0.06	0.11	< 0.001
	<i>PMean</i>	0.31	0.24	0.39	0.34	0.31	0.37	0.03	-0.02	0.09	< 0.05
	<i>PMean +Cov</i>				0.39	0.31	0.48	0.08	0.06	0.11	< 0.001
TPR	<i>FNorm</i>				37.77	32.97	43.02	14.27	12.25	16.29	< 0.001
	<i>PMean</i>	23.5	17.6	30.09	52.62	49.31	57.16	29.13	25.23	32.99	< 0.001
	<i>PMean +Cov</i>				29.37	19.94	40.33	5.87	2.34	10.24	< 0.05
FP/min	<i>FNorm</i>				3.71	2.98	4.21	2.19	1.91	2.36	> 0.05
	<i>PMean</i>	1.52	1.08	2.07	11.95	11.55	12.99	10.44	9.99	10.92	> 0.05
	<i>PMean +Cov</i>				1.05	0.85	1.22	-0.47	-0.89	0.08	> 0.05

Table 3: Best adaptive method compared to static sLDA classification for each metric either with or without FP/min control including statistical significance corrected for multiple comparisons.

Metric	FP/min Control	Best Method	p-value
MCC	False	<i>FNorm</i>	< 0.001
	True	<i>PMean+Cov</i>	< 0.001
TPR	False	<i>PMean</i>	< 0.001
	True	<i>PMean</i>	< 0.001
FP/min	False	<i>PMean+Cov</i>	> 0.05
	True	<i>PMean+Cov</i>	> 0.05

Table 4: Comparison of the best adaptive method and the static sLDA baseline in terms of robustness across sessions, evaluated with and without FP/min control. Session-wise performance slopes are reported.

Metric	FP/min Control	Best Method	Adaptive Slope	Static Slope
MCC	False	<i>PMean</i>	-0.015	-0.043
	True	<i>PMean</i>	-0.029	
TPR	False	<i>PMean</i>	-1.984	-3.684
	True	<i>FNorm</i>	-4.585	
FP/min	False	<i>PMean+Cov</i>	0.001	0.246
	True	<i>PMean</i>	-0.19	

Table 2: Performance metrics of each adaptive method with FP/min control. Session-averaged values (min-max), differences relative to the static baseline and associated statistical significance corrected for multiple comparisons.

Metric	Method	Static			Adaptive			Delta			p-value
		Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	
MCC	<i>FNorm</i>				0.37	0.3	0.45	0.06	0.05	0.07	< 0.001
	<i>PMean</i>	0.31	0.24	0.39	0.37	0.32	0.42	0.06	0.03	0.09	< 0.001
	<i>PMean</i>				0.39	0.32	0.48	0.08	0.06	0.11	< 0.001
	+Cov										
TPR	<i>FNorm</i>				31.74	24.19	40.85	8.24	6.56	10.76	< 0.001
	<i>PMean</i>	23.5	17.6	30.09	40.85	32.89	49.64	17.35	15.27	19.55	< 0.001
	<i>PMean</i>				29.98	21.04	40.33	6.48	3.44	10.24	< 0.001
	+Cov										
FP/min	<i>FNorm</i>				2.54	2.43	2.71	1.02	0.64	1.46	> 0.05
	<i>PMean</i>	1.52	1.08	2.07	4.22	3.81	4.86	2.7	2.01	3.78	> 0.05
	<i>PMean</i>				1.07	0.86	1.27	-0.44	-0.88	0.08	> 0.05
	+Cov										

DISCUSSION

Group-level LMMs revealed significant performance differences between adaptive strategies and the static sLDA baseline. Without FP/min control, the largest TPR increase was observed for *PMean* (+29.13%, $p \approx 10^{-53}$, see Tab. 1 and Fig. 1A), followed by *PMean* with FP control evaluated in the unconstrained setting (+17.35%, $p \approx 10^{-20}$, see Tab. 2 and Fig. 1B). All adaptive methods achieved positive MCC contrasts relative to static baseline. However, these improvements were accompanied by substantially elevated FPRs, particularly for *PMean* without FP/min control (+10.44 FP/min, $p > 0.05$, see Tab. 1 and Fig. 1A), reflecting a sensitivity-specificity trade-off. This may be partly explained by class imbalance, where no-movement samples dominate. *PMean* assumes balanced class contributions; under imbalance, the running mean can become biased toward the majority class, shifting the decision boundary and thereby increasing the FPR. Session-wise analyses indicated general performance declines over time across methods. Nevertheless, *PMean* showed less steep negative slopes for MCC and TPR compared to static decoding (see Tab. 4), suggesting improved robustness to temporal variability, although none of the methods achieved consistently stable performance across sessions.

With FP/min control, adaptive methods continued to outperform the static baseline in MCC. Notably, *PMean+Cov* achieved the largest MCC increase (+0.082, $p < 10^{-6}$), followed by *PMean* (+0.059, $p < 10^{-3}$), as shown in Tab. 2 and Fig. 1B. In contrast to the uncontrolled setting, FP/min control substantially constrained FPRs. *PMean+Cov* yielded the lowest FP/min and slightly fewer FPs than static decoding (-0.44 and -0.47 FP/min, see Fig. 1 and Tab. 2 and Tab. 1 respectively), although these reductions were not statistically significant after correction.

Together, these findings highlight a trade-off between TPR and FPR. Without FP/min control, *PMean* maximized TPR but at the cost of markedly increased FPs, limiting usability in real-time BCI applications. For example, in a speller operating at six characters per

minute, FPRs above three per minute would result in roughly every second selection being incorrect, rendering the system impractical. Introducing FP/min control shifted adaptive methods toward a more practical operating point, preserving MCC gains while stabilizing FPRs. However, this controller relies on labelled feedback and is therefore not applicable in fully online scenarios.

Elevated FP/min in several adaptive strategies likely reflects increased sensitivity to non-stationary feature distributions. In addition, methods such as *PMean* may be sensitive to class imbalance, which is common in BCI settings, biasing the classifier toward more frequent classes. Among all methods, *PMean+Cov* achieved the best balance between sensitivity and specificity in both controlled and uncontrolled settings.

Overall, these findings demonstrate that simple adaptive extensions of linear models improve robustness to non-stationarity without requiring large datasets or computationally expensive deep learning architectures [6-9], making them suitable for low-trial EEG settings and real-time implantable BCIs.

CONCLUSION

This study investigated whether lightweight adaptation of sLDA improves decoding performance under non-stationary EEG conditions. The results support the hypothesis that adaptive extensions outperform static decoding. Among the evaluated strategies, pooled mean adaptation with covariance update (*PMean+Cov*) provided the best balance between sensitivity and FPRs and is therefore most suitable for real-time deployment. As an alternative, feature normalization (*FNorm*) demonstrated stable performance across sessions in both TPR and FP/min. Although FP/min remained elevated compared to *PMean+Cov*, it was substantially lower than for *PMean* while maintaining consistent detection performance. *FNorm* may therefore be preferable when cross-session stability is the primary objective.

However, methods relying on global statistics, such as *PMean*, may be sensitive to class imbalance, inherent to BCI paradigms. Future work should therefore address

class imbalance during adaptation to improve robustness and reduce FPs.

Overall, these results indicate that simple adaptive extensions of linear models provide an effective, practical approach for improving robustness to non-stationarity in real-time BCI systems.

ACKNOWLEDGMENTS

This research was funded by the European Union's HORIZON-EIC-2021-PATHFINDER CHALLENGES program under grant agreement No 101070939 and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00198.

REFERENCES

- [1] Birbaumer N, Ghanayim N, Hinterberger T, Iversen I, Kotchoubey B, Kübler A, et al. A spelling device for the paralysed. *Nature*. 1999 Mar;398(6725):297-8.
- [2] Raza H, Rathee D, Zhou SM, Cecotti H, Prasad G. Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related EEG-based brain-computer interface. *Neurocomputing (Amst)*. 2019 May 28;343:154-66.
- [3] Hehenberger L, Kobler RJ, Lopes-Dias C, Srisrisawang N, Tumfart P, Uroko JB, et al. Long-Term Mutual Training for the CYBATHLON BCI Race With a Tetraplegic Pilot: A Case Study on Inter-Session Transfer and Intra-Session Adaptation. *Front Hum Neurosci*. 2021 Feb 26;15:635777.
- [4] Buttfeld A, Ferrez PW, Millán J del R. Towards a robust BCI: error potentials and online learning. *IEEE Trans Neural Syst Rehabil Eng*. 2006 Jun;14(2):164-8.
- [5] Vidaurre C, Schlögl A, Cabeza R, Scherer R, Pfurtscheller G. Adaptive on-line classification for EEG-based brain computer interfaces with AAR parameters and band power estimates. *Biomed Tech (Berl)*. 2005 Nov;50(11):350-4.
- [6] Wu D, Xu Y, Lu BL. Transfer learning for EEG-based brain-computer interfaces: A review of progress made since 2016. *IEEE Trans Cogn Dev Syst*. 2022 Mar;14(1):4-19.
- [7] Wang X, Yang R, Huang M. An Unsupervised Deep-Transfer-Learning-Based Motor Imagery EEG Classification Scheme for Brain-Computer Interface. *Sensors (Basel)* [Internet]. 2022 Mar 14;22(6). Available from: <http://dx.doi.org/10.3390/s22062241>
- [8] Liang S, Li L, Zu W, Feng W, Hang W. Adaptive deep feature representation learning for cross-subject EEG decoding. *BMC Bioinformatics*. 2024 Dec 31;25(1):393. Elsevier, Amsterdam, Netherlands (1999)
- [9] Zhao H, Zheng Q, Ma K, Li H, Zheng Y. Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification. *IEEE Trans Neural Netw Learn Syst*. 2021 Feb;32(2):535-45.
- [10] Vidaurre C, Sannelli C, Müller KR, Blankertz B. Co-adaptive calibration to improve BCI efficiency. *J Neural Eng*. 2011 Apr;8(2):025009.
- [11] Vidaurre C, Schlögl A, Cabeza R, Scherer R, Pfurtscheller G. Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Trans Biomed Eng*. 2007 Mar;54(3):550-6.
- [12] Vidaurre C, Kawanabe M, von Büna P, Blankertz B, Müller KR. Toward unsupervised adaptation of LDA for brain-computer interfaces. *IEEE Trans Biomed Eng*. 2011 Mar;58(3):587-97.
- [13] Wu S, Bhadra K, Giraud AL, Marchesotti S. Adaptive LDA Classifier Enhances Real-Time Control of an EEG Brain-Computer Interface for Decoding Imagined Syllables. *Brain Sci* [Internet]. 2024 Feb 21;14(3).
- [14] Kapgate DD. Hybrid SSVEP + P300 brain-computer interface can deal with non-stationary cerebral responses with the use of adaptive classification. *J Neurorestoratology*. 2024 Jun;12(2):100109.
- [15] Peck R, Van Ness J. The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans Pattern Anal Mach Intell*. 1982 May;4(5):530-7.
- [16] Sosulski J, Tangermann M. Introducing block-Toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain-computer interfaces. *J Neural Eng* [Internet]. 2022 Nov 7;19(6). Available from: <http://dx.doi.org/10.1088/1741-2552/ac9c98>
- [17] Egger J, Kostoglou K, Müller-Putz GR. Chrono-EEG dynamics influencing hand gesture decoding: a 10-hour study. *Sci Rep*. 2024 Aug 30;14(1):20247.
- [18] Kobler, R. J., Sburlea, A. I., Lopes-Dias, C., Schwarz, A., Hirata, M., & Müller-Putz, G. R. (2020). Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals. *NeuroImage*. 2020 Sep 1;218:117000.
- [19] Perrin F, Pernier J, Bertrand O, Echallier JF. Spherical splines for scalp potential and current density mapping. *Electroencephalogr Clin Neurophysiol*. 1989 Feb;72(2):184-7.
- [20] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2), 365-411.
- [21] Crell MR, Kostoglou K, Suwandjieff P, Cruz JR da, Muller-Putz GR. A Non-Invasive, MRCP-Based BCI for Online Communication. *IEEE Trans Neural Syst Rehabil Eng*. 2026;34:1272-83.
- [22] Apicella A, Isgro F, Pollastro A, Prevete R. On the effects of data normalization for domain adaptation on EEG data. *Eng Appl Artif Intell*. 2023 Aug;123(106205):106205.

Wrapped One-Class Riemannian EEG Classifier for BCI-Based Detection of Anesthetic States

Valérie Marissens Cueva^{1,2†}, Thibault de Surrel^{3†}, Laurent Bougrain^{2,4},
Seyed Javad Bidgoli⁵, Guy Cheron⁶, Ana Maria Cebolla Alvarez⁶, Claude Meistelman⁷,
Fabien Lotte¹, Florian Yger⁸, Sébastien Rimbart¹
† These authors contributed equally to the paper.

¹Inria Center at Univ. Bordeaux / LaBRI, Talence, France; ²Université de Lorraine, CNRS, LORIA, Nancy, France; ³LAMSADE, CNRS, PSL Univ. Paris-Dauphine, France; ⁴Sorbonne Université, ICM, CNRS, Inria, Inserm, Paris, France; ⁵CHU Brugmann, Bruxelles, Belgium; ⁶Laboratory of Neurophysiology and Movement Biomechanics, Université Libre de Bruxelles, Bruxelles, Belgium; ⁷DevAH, University of Lorraine, Nancy, France; ⁸LITIS, INSA Rouen-Normandy, France;
E-mail: thibault.de-surrel@lamsade.dauphine.fr & valerie.marissens@loria.fr

ABSTRACT: Brain-computer interfaces based on ElectroEncephaloGraphy (EEG) often represent efficiently brain signals as covariance matrices and leverage Riemannian geometry for classification of mental states. However, current approaches typically require multiple classes for training. In many applications such as intraoperative monitoring of consciousness during anesthesia, only data from one class, such as the awake state, may be available, requiring one-class classification methods. We propose a novel Riemannian one-class classifier called the One-Class Wrapped Gaussian. Unlike existing methods that rely solely on the Riemannian mean, our approach incorporates second-order statistical information by using an anisotropic Gaussian-like distribution on the manifold of covariance matrices. We validated our method on EEG data from 19 patients undergoing general anesthesia. Results show that our classifier significantly outperforms state-of-the-art one-class methods for distinguishing between awake and anesthetized states, for clinically relevant electrode numbers. We further demonstrate the robustness of our approach by testing configurations with reduced electrode numbers, confirming its feasibility for real-world surgical settings where electrode placement is constrained.

INTRODUCTION

Current Brain-Computer Interfaces (BCIs) based on ElectroEncephaloGraphy (EEG) typically rely on a subject-specific calibration, due to the non-stationarity nature of EEG signals and the large inter-subject variability [1]. Supervised training is generally required to discriminate between mental states, which implies the availability of labeled data for each class. In many real world applications, however, this requirement may be difficult or even impossible to satisfy. In particular, clinical monitoring scenarios often only allow access to data from a single reference condition, while the alternative cannot be labeled in advance. For instance, to monitor the transition between two states of general anesthesia, awake

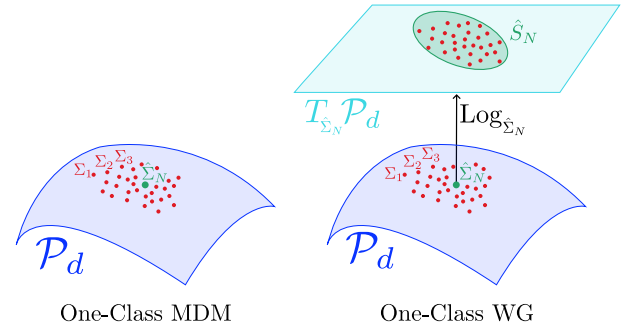


Figure 1: Illustration of the two compared algorithms: *One-Class MDM* (left) and *One-Class WG* (right). While *One-Class MDM* estimates only the Riemannian mean $\hat{\Sigma}_N$ of the training covariance matrices $\Sigma_1, \Sigma_2, \Sigma_3, \dots$, *One-Class WG* also estimates a dispersion matrix in the tangent space $T_{\hat{\Sigma}_N} \mathcal{P}_d$.

or under anesthesia, a classifier would require data from both states. However, before surgery, data can only be obtained from an awake patient. This constraint motivates the use of one-class classification strategies [2, 3], in which a model is trained exclusively on examples from a target class and detects deviations from this reference. While Riemannian geometry-based methods, relying on covariance matrices, have become a standard framework for EEG decoding in BCIs [4], existing approaches usually rely on two-class (or more) training.

Among Riemannian algorithms, the Minimum Distance to the Riemannian Mean (MDM) classifier [5] has been adapted to a one-class framework as a baseline method [6], where the target class is modeled by its geometric centroid. Although simple and efficient, this approach implicitly assumes that covariance matrices are equally distributed in all directions, i.e., that they follow an isotropic distribution [7]. As a result, it cannot explicitly model any anisotropic dispersion of data on the Riemannian manifold of symmetric positive definite (SPD) matrices.

In this work, we introduce a novel one-class classifier based on the Wrapped Gaussian distribution defined on the Riemannian manifold of SPD matrices. The proposed

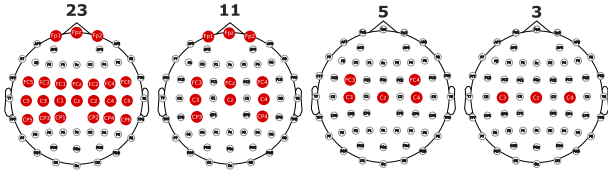


Figure 2: Electrode layouts tested compared to the full 64 electrodes montage.

method provides an explicit -and anisotropic- probabilistic description of the target class preserving the intrinsic geometry of covariance matrices. We evaluate the proposed approach in an EEG-based BCI designed to detect motor cortical activity [8]. Specifically, the motor pattern elicited by median nerve stimulation (MNS) is analyzed, as discriminating between MNS alone and MNS combined with motor intention (MI) has been shown to be more effective than distinguishing MI from rest [9]. The objective is to differentiate EEG responses recorded in awake subjects from those observed under altered brain dynamics induced by general anesthesia. Ultimately, this framework aims to contribute to the detection of intraoperative awareness by identifying potential motor intention from the anesthetized patient. This setting naturally defines a one-class learning problem, as only the reference condition can be collected before the surgery (here EEG during the awake state) for subject-specific calibration, while the alternative condition (here subject-specific EEG during anesthesia) is not accessible beforehand. Beyond this specific use case, the proposed method could be relevant for a broader range of BCI applications involving reference-only training, such as passive BCIs, vigilance monitoring, or brain-state change detection [3].

MATERIALS AND METHODS

Data description: Protocol: This study was conducted at the hospital of Brugmann in Belgium, approved by its ethical committee (CE 2021/225), and registered at EUDRACT (2021-006457-56) and ClinicalTrials.gov (NCT05272202) [8]. A total of 28 subjects scheduled for surgery under general anesthesia were initially enrolled. Four subjects (1, 2, 5 and 28) were excluded either due to technical or logistical reasons beyond our control. Five additional subjects (9, 10, 11, 15 and 25) were excluded to ensure dataset homogeneity in terms of stimulation side and anesthesia protocol. The final cohort consists of 19 subjects (10 females, 47.47 ± 12.51 years old). Loss of consciousness was induced using propofol, while the rest of the anesthesia protocol was determined by the attending anesthesiologist. Ketamine was not administered. The target propofol concentrations were recorded simultaneously with the EEG signals. Neuromuscular blocking agents (Rocuronium) were administered when clinically required (13 patients: 4, 6, 7, 8, 12, 13, 16, 17, 18, 22, 23, 24 and 26).

EEG acquisition and MNS: MNS consisted of 0.2 ms square electrical pulses, generated by the Micromed SD Ltm Stim Energy stimulator, and delivered to the left

wrist through a pair of grass gold cup electrodes. Stimulation intensity was individually adjusted to elicit a visible thumb twitch [10] and remained below 15 mA. EEG signals were recorded with a 64-channel ANT Neuro cap, and the eego mylab system at a sampling rate of 4096 Hz to preserve high-frequency content for future somatosensory evoked potential analyses. In this study, EEG signals were downsampled to 128 Hz and band-pass filtered (8-30 Hz). Artifacts were then visually rejected using EEGLAB[11] (28 ± 37 rejected epochs), and the data were epoched into 750 ms windows (250-1000 ms after MNS). Several electrode montages were selected (see Fig. 2), as current anesthesia monitors use frontal electrodes to detect changes in consciousness associated with general anesthesia [12], whereas, given that this is a motor task evoked by stimulation of the median nerve, the electrodes located in the motor cortex are also of interest.

Recording sessions and training setup: Each run had 150 MNS with a 3-4 s inter-stimulus interval across two sessions. In the preoperative session, at least one run was acquired with the awake patient, yielding 94-499 MNS trials per subject. During the intraoperative session, several EEG runs were recorded, the number of runs depending on the surgery duration, resulting in 415-2390 MNS trials in the anesthetized condition. One-class classifiers were trained on the first 65 preoperative trials, the awake class, and evaluated on the remaining preoperative trials (awake test class) and on the entire intraoperative set (anesthetized test class), resulting in 444-2472 test trials per subject. Classifications were performed using the MNE [13], Scikit-learn [14] and pyRiemann [15] packages in Python 3.10¹.

Riemannian tools used in BCI: In this paper, we represent an EEG trial X by its covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ where d is the number of EEG electrodes used. This covariance matrix contains information on the link between the different electrodes. Its underlying specific mathematical structure is a SPD matrix. Using the Riemannian geometry of the set $\mathcal{P}_d$ of SPD matrices has lead to state-of-the-art classifiers in EEG based BCIs [4]. Given an EEG trial X , one can compute its covariance matrix $\Sigma = \frac{1}{t-1}XX^T$, where t is the number of sampled time points in each trial. In this paper, we rely on the *Affine-Invariant Riemannian geometry* [16] on $\mathcal{P}_d$. This geometry induces a distance between two covariance matrices Σ_1 and Σ_2 , defined as follows:

$$\delta_R(\Sigma_1, \Sigma_2) = \|\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\|_F,$$

where $\log$ denotes the matrix logarithm and $\|\cdot\|_F$ the Frobenius norm. The tangent space $T_{\Sigma}\mathcal{P}_d$ is a linear approximation of the curved space $\mathcal{P}_d$ at Σ and can be identified with the flat space of $d \times d$ symmetric matrices $\mathcal{S}_d$, and thus, is of dimension $d(d+1)/2$. To move between $\mathcal{P}_d$ and $T_{\Sigma}\mathcal{P}_d$, we use the Riemannian exponential Exp_{Σ} and its inverse, the logarithm map, defined

¹Find the code at https://github.com/valeriemarissens/one_class_wrapped_gaussian

for $\Sigma_1, \Sigma_2 \in \mathcal{P}_d$ and $U \in T_{\Sigma_1} \mathcal{P}_d$ as:

$$\begin{aligned} \text{Exp}_{\Sigma_1}(U) &= \Sigma_1^{1/2} \exp(\Sigma_1^{-1/2} U \Sigma_1^{-1/2}) \Sigma_1^{1/2}, \\ \text{Log}_{\Sigma_1}(\Sigma_2) &= \Sigma_1^{1/2} \log(\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}) \Sigma_1^{1/2}. \end{aligned} \quad (1)$$

Finally, the Riemannian mean of N covariance matrices $\Sigma_1, \dots, \Sigma_N$ is the unique minimizer $\hat{\Sigma}_N$ of

$$\hat{\Sigma}_N = \arg \min_{\Sigma \in \mathcal{P}_d} \sum_{i=1}^N \delta_R^2(\Sigma, \Sigma_i). \quad (2)$$

It can be computed using a Riemannian gradient descent.

Wrapped Gaussian on $\mathcal{P}_d$: Let us now define a probability distribution on the manifold $\mathcal{P}_d$ of SPD matrices. This distribution is a version of a multivariate Gaussian defined directly on $\mathcal{P}_d$. For this, we need to define the vectorization $\text{Vect}_{\Sigma}(U)$ at $\Sigma \in \mathcal{P}_d$ of a symmetric matrix U , that is, taking the upper triangular portion of $\Sigma^{-1/2} U \Sigma^{-1/2}$, with off-diagonal elements scaled by $\sqrt{2}$, forming a vector in $\mathbb{R}^{d(d+1)/2}$. One can also consider the inverse operation, $\text{Unvect}_{\Sigma}(t)$, that takes a vector $t \in \mathbb{R}^{d(d+1)/2}$, and creates a symmetric matrix. Let $\Sigma \in \mathcal{P}_d$ and S be a SPD matrix of size $d(d+1)/2 \times d(d+1)/2$, then we say that the random variable X on $\mathcal{P}_d$ follows a *wrapped Gaussian*, denoted $\text{WG}(\Sigma, S)$ if

$$X = \text{Exp}_{\Sigma}(\text{Unvect}_{\Sigma}(t)), \quad t \sim \mathcal{N}(0, S).$$

The idea is to linearize the curved space $\mathcal{P}_d$ at the point Σ using the tangent space $T_{\Sigma} \mathcal{P}_d$ and to draw the Gaussian distribution $\mathcal{N}(0, S)$ on this flat space. Then, to respect the natural curvature of $\mathcal{P}_d$, the Exp_{Σ} wraps this flat distribution back onto the manifold. The parameter Σ is the center of the distribution and S is the dispersion matrix in the tangent space $T_{\Sigma} \mathcal{P}_d$. This wrapped Gaussian distribution has been studied in depth in [17] and several learning algorithms were derived in [18]. One can compute the density of this distribution in a closed form, which we omit here (see Theorem 4.2 of [17]). Given a sample $(\Sigma_1, \dots, \Sigma_N)$ of SPD matrices following a wrapped Gaussian $\text{WG}(\Sigma, S)$, one can estimate Σ using the Riemannian mean $\hat{\Sigma}_N$ (see Equation 2) and estimate S as follows:

$$\hat{S}_N = \frac{1}{N} \sum_{i=1}^N \text{VLog}_{\hat{\Sigma}_N}(\Sigma_i) \text{VLog}_{\hat{\Sigma}_N}(\Sigma_i)^{\top},$$

where $\text{VLog}_{\hat{\Sigma}_N}(\Sigma_i) = \text{Vect}_{\hat{\Sigma}_N}(\text{Log}_{\hat{\Sigma}_N}(\Sigma_i))$. These estimators are the maximum likelihood estimators.

One-Class Riemannian classifiers: We consider two types of one-class Riemannian classifiers. They are illustrated at Figure 1.

One-Class MDM: The first one-class classifier that we consider in this study is based on the *Minimum Distance to the Mean* (MDM) classifier introduced in [5]. We call this one-class classifier the *One-Class MDM* [6]. Given a training dataset $\Sigma_1, \dots, \Sigma_N$ of N covariance matrices of the class awake, the *One-Class MDM* computes their Riemannian mean $\hat{\Sigma}_N$ using equation 2. Then, given a new

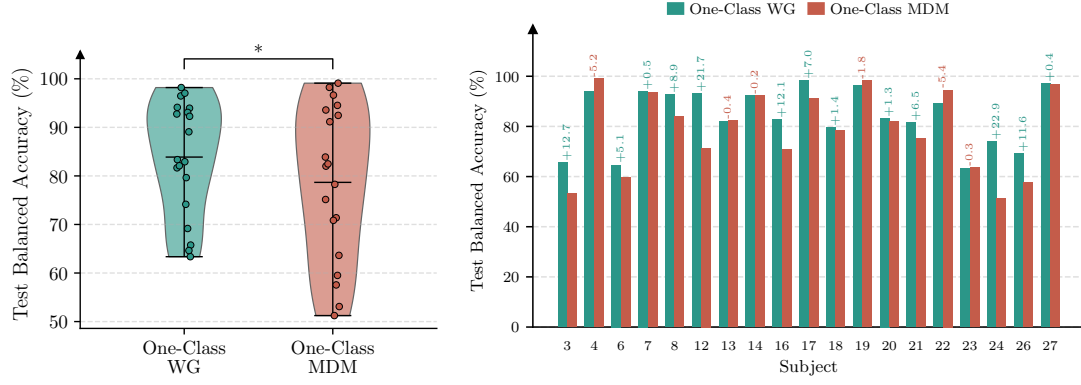
covariance matrix $\tilde{\Sigma}$, it computes the Riemannian distance $\delta_R(\tilde{\Sigma}, \hat{\Sigma}_N)$ between $\tilde{\Sigma}$ and $\hat{\Sigma}_N$ and classifies $\tilde{\Sigma}$ as awake if $\delta_R(\tilde{\Sigma}, \hat{\Sigma}_N) < \varepsilon$ and anesthetized otherwise, where ε is a threshold. In the experiments, the threshold ε was set to the median of the distances of the training data to $\hat{\Sigma}_N$ plus three times the standard deviation (same threshold as the Riemannian potato outlier detection method [19]).

One-Class WG (proposed): Now we introduce a novel one-class classifier, called *One-Class WG*, based on the Wrapped Gaussian distribution introduced in section *Wrapped Gaussian on $\mathcal{P}_d$* , which does not rely solely on the Riemannian mean of the training data. Given training covariance matrices $\Sigma_1, \dots, \Sigma_N$ of N from the awake class, our classifier estimates the parameters Σ and S of the wrapped Gaussian $\text{WG}(\Sigma, S)$ using the estimators $\hat{\Sigma}_N$ and $\hat{S}_N$. For a new covariance matrix $\tilde{\Sigma}$, the log-likelihood of $\tilde{\Sigma}$ under $\text{WG}(\hat{\Sigma}_N, \hat{S}_N)$ is computed and the trial is classified as awake if it exceeds a given threshold ε and anesthetized otherwise. In the experiments, the threshold ε was set to 95% of the training log-likelihoods. Notably, if the dispersion matrix is constrained to $S = \alpha I$, the classification becomes equivalent to a distance-based classifier. Thus, *One-Class MDM* is a particular isotropic case of the proposed *One-Class WG* model [18].

RESULTS

Main results: Let us start by comparing the mean balanced accuracy across all subjects of both classifiers *One-Class MDM* and *One-Class WG* on the data described in section *Materials and Methods*. The choice of this metric was driven by the highly asymmetric class imbalance. *One-Class MDM* achieved $78.67\% \pm 15.93$, whereas *One-Class WG* reached $83.88\% \pm 11.69$ (see Fig. 3a). A two-sided paired t-test confirmed that *One-Class WG* significantly outperformed *One-Class MDM* ($p = 0.0118$). Fig. 3b shows the accuracy of both classifiers for each subject, and for 13 subjects out of the 19, *One-Class WG* substantially improves the performances. These results were obtained when considering only five electrodes (Fig. 2) and for the *One-Class WG*, the estimated dispersion matrix $\hat{S}_N$ is only diagonal. We conducted an ablation study on the number of electrodes, on whether $\hat{S}_N$ should be diagonal, and on the choice of the log-likelihood threshold. Although performance varied only moderately across $\varepsilon \in \{0.85, 0.90, 0.95, 0.99\}$, significant differences between *One-Class WG* and *One-Class MDM* were obtained only at $\varepsilon = 0.95$.

Influence of the number of electrodes: We studied five different setups with an increasing number of electrodes (see Fig. 2). The results for the different setups, as well as the significance level of the two-sided paired t-test, with a Benjamini-Hochberg correction, between *One-Class WG* and *One-Class MDM* are given in Figure 4. One can see that, for the proposed *One-Class WG*, the best number of electrodes is 5, while for the *One-Class MDM*, the best number of electrodes is 11. With more electrodes, *One-*



(a) Distribution of the balanced accuracy across all subjects using 5 electrodes. A two-sided paired t-test shows that *One-Class WG* significantly outperforms *One-Class MDM* (* $p < 0.05$).

Figure 3: Balanced test accuracy of the two one-class classifiers (*One-Class WG* vs *One-Class MDM*) on the dataset. This is done on five electrodes (C3, C4, Cz, FC4 and FC3) and for the *One-Class WG*, the estimated Σ is only diagonal.

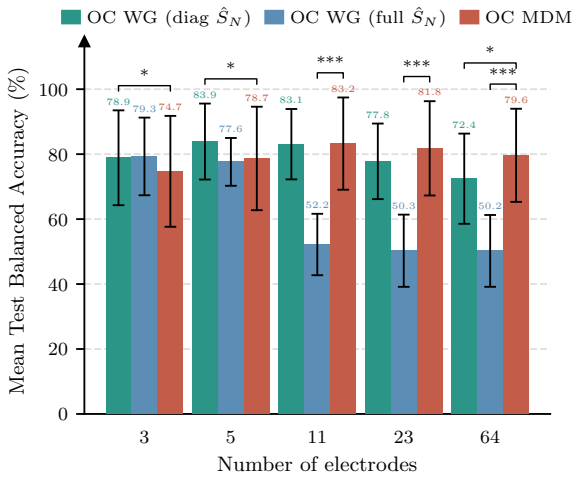


Figure 4: Mean test balanced accuracy of One-Class (OC) classifiers across different electrode setups. Statistical significance was assessed using a two-sided paired t-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Class MDM outperforms our proposed method.

Influence of the type of the estimated $\hat{S}_N$: We also studied whether or not we should constrain the estimated dispersion matrix $\hat{S}_N$ to be diagonal: $\hat{S}_N = \text{diag}(\hat{s}_{N,1}^2, \dots, \hat{s}_{N,d(d+1)/2}^2)$. This parameterization reduces the number of parameters to estimate from $d(d+1)/2 \times d(d+1)/2$ to simply $d(d+1)/2$. The coefficients $\hat{s}_{N,j}^2$ are then simply the variances of the $(\text{Vect}_{\hat{S}_N}(\text{Log}_{\hat{S}_N}(\Sigma_i)))_i$. This leads to two variants of the proposed *One-Class WG*: with a full dispersion matrix $\hat{S}_N$ and with a diagonal one. The results of these two classifiers using the different montages are also in figure 4. One can see that, as soon as the number of electrodes is at least 5, the performances of the *One-Class WG* with a diagonal $\hat{S}_N$ are much better than the one with a full $\hat{S}_N$.

DISCUSSION

Overall, the proposed *One-Class WG* improved discrimination between awake and anesthetized conditions compared to the *One-Class MDM*, with reduced electrode montages, despite the variability of a real clinical setting. The following discussion examines these results in terms of continuous behavior of model scores across anesthetic depth, robustness to clinical variability, influence of montage, and the role of anisotropic dispersion modeling.

Continuous validation of one-class model scores through different levels of anesthetic depth:

To further understand model behaviors, we analyze the temporal evolution of their respective decision scores in relation to propofol concentration (see Fig. 5). Rather than focusing only on binary classification accuracy, this analysis allows us to evaluate how each model responds continuously to changes in anesthetic depth. *One-Class WG* score corresponds to the log-likelihood of each median nerve stimulation-evoked covariance matrix under the Wrapped Gaussian distribution. *One-Class MDM* score is the Riemannian distance to the awake centroid. Propofol concentration is displayed in the background to assess consistency with anesthetic depth. Both training (first 65 awake trials) and test data are displayed.

For subject 4 (94% *One-Class WG*, 99% *One-Class MDM*), both models clearly separate awake and anesthetized trials, with scores progressively diverging from the awake reference as propofol increases. The slightly lower *One-Class WG* accuracy may reflect a conservative decision threshold, rather than a failure of the underlying distributional modeling.

In contrast, subject 24 exhibits the largest discrepancy (74% *One-Class WG* vs. 51% *One-Class MDM*). While both models deviate from the awake reference beyond 4-5 $\mu\text{g/mL}$, *One-Class WG* yields a more consistent separation. One possible explanation lies in the EEG dynamics at high propofol concentrations, often associated with burst suppression patterns which are composed of periods of isoelectric activity and high-amplitude bursts [20]. Such dynamics induce strong variability in covari-

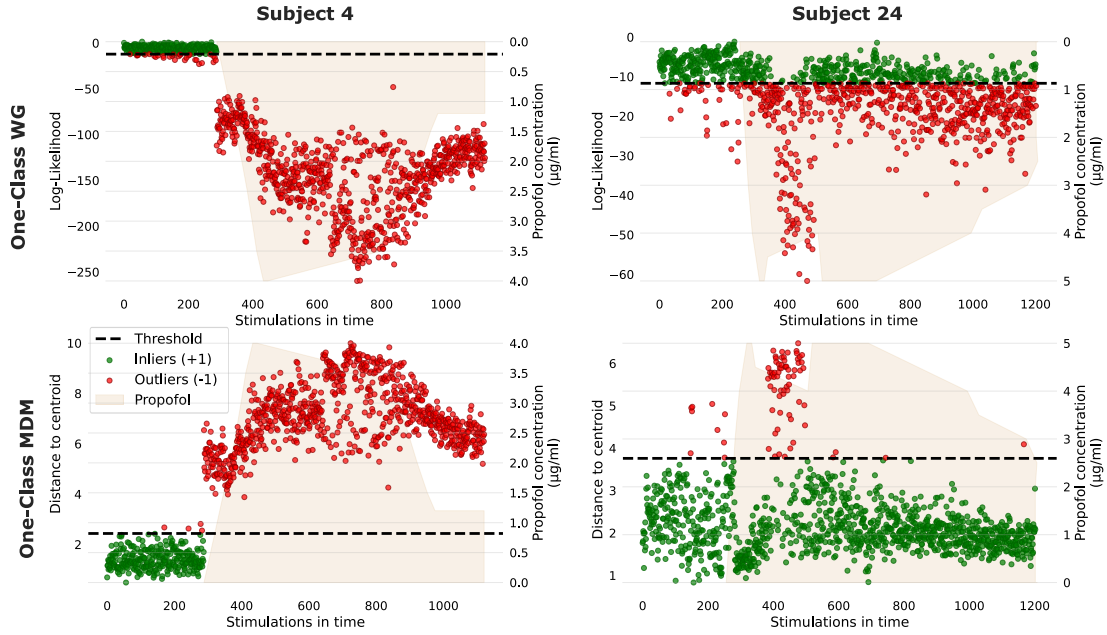


Figure 5: Evolution of one-class model scores over time, for two representative subjects: subject 4 (*One-Class WG*: 94%, *One-Class MDM*: 99%) and subject 24 (*One-Class WG*: 74%, *One-Class MDM*: 51%). These examples illustrate the inter-subject variability observed across the dataset.

ance structures. Because *One-Class MDM* relies solely on distance to a centroid, it does not explicitly model dispersion, whereas *One-Class WG* models not only the centroid but also the variability of the awake distribution through the dispersion matrix, yet covariance matrices corresponding to burst activity may not be uniformly distant from the awake centroid. Future analyses could quantify burst suppression via burst suppression ratio [12] and relate it to one-class model scores, or use Riemannian visualization techniques [21] to enable a geometric interpretation of how deep anesthesia shifts the data distribution relative to the awake reference. Further work will improve robustness in subjects exhibiting lower classification performance.

Robustness in the clinical anesthesia setting: An important aspect of the proposed approach is its robustness to variability. Unlike controlled laboratory BCI datasets, these data were acquired in a real surgical environment, introducing multiple sources of variability, including operating room noise, patient positioning affecting EEG stability, heterogeneous anesthesia protocols, and inter-subject differences. Despite this variability, *One-Class WG* outperformed *One-Class MDM*, suggesting that explicitly modeling dispersion of the reference class may improve robustness compared to centroid-based methods.

Effect of the montage: EEG was recorded with 64 channels, which leads to 64×64 covariance matrices. With *One-Class WG*, the estimated dispersion matrix $\hat{S}_N$ will be of dimension $2,080 \times 2,080$ (indeed, $64 * (64 + 1)/2 = 2,080$). The small number of training data (65 trials) and their short duration of 750 ms is not sufficient to correctly estimate $\hat{S}_N$ (see Section 5 of [17]). This is why we tried different montages with a lower number of electrodes. Moreover, in a real-life surgical setting, simpler setups are preferred. As shown in our results, the

proposed method works best with 3-5 electrodes, improving mean accuracy across all subjects by up to 5% over *One-Class MDM*. The best results are achieved with *One-Class WG* using only 5 electrodes and a diagonal $\hat{S}_N$.

Isotropy vs anisotropy: Let us now make an hypothesis to explain why the proposed *One-Class WG* performs better than *One-Class MDM*. *One-Class MDM* estimates only the mean of the training covariance matrices and therefore relies only on first order information. *One-Class WG*, estimates also a dispersion matrix, gathering second order information. The estimated dispersion matrix $\hat{S}_N$ captures the spread of the point cloud and may exhibit preferred directions, reflecting the anisotropic structure of the data rather than an isotropic one. This anisotropic behavior of the point cloud cannot be captured by the *One-Class MDM*, as it focuses only on first order information. To quantify this effect, we looked at the variance of the diagonal elements of $\hat{S}_N$ in the 5-electrodes setup and when $\hat{S}_N$ is only diagonal (where *One-Class WG* performs the best). Large variance across diagonal coefficients suggest an anisotropic point cloud, whereas similar diagonal coefficients indicate isotropy. Let us write the matrix $\hat{S}_N = \text{diag}(\hat{s}_{N,1}^2, \dots, \hat{s}_{N,d(d+1)/2}^2)$, then, we compute the variance of the diagonal coefficients ($\hat{S}_N$ has $d(d+1)/2$ diagonal coefficients), yielding one variance per subject. Subjects are then partitioned according to whether *One-Class WG* or *One-Class MDM* achieved higher accuracy, and we compute the mean variance within each group. Across all subjects, the group in which *One-Class WG* performs best has a mean variance of 3.021×10^{-2} , about 16 times larger than that of the group in which *One-Class MDM* performs best (1.891×10^{-3}). This suggests that *One-Class WG* tends to perform better when the point cloud is more anisotropic. Although this difference is partly driven by

a single outlier with unusually high variance, removing it still leaves the mean variance of the *One-Class WG* group 2.8 times larger. Statistical tests were not conducted due to the small number of subjects (6) in the latter group.

Perspectives: Several directions may extend this work. First, the clear separation between awake and anesthetized scores observed in Fig. 5 suggests that Riemannian change-point detection algorithms [22] could exploit the temporal structure of the data instead of treating trials independently. Second, the interpretability of Riemannian models could be assessed [23] as it remains a key aspect for clinical adoption. Future work will also include validation on additional EEG datasets, and extension to multimodal one-class models. Although no comparison with standard depth-of-anesthesia monitors was performed in the present study, such analyses are currently underway using a newly acquired dataset.

CONCLUSION

A new probabilistic one-class Riemannian classifier based on Wrapped Gaussians, *One-Class WG*, was introduced and evaluated in a BCI to detect anesthetic states. The proposed approach significantly outperformed the baseline *One-Class MDM* with a clinically-relevant number of EEG sensors, with a lower inter-subject variability, highlighting the importance of modeling the dispersion on the Riemannian manifold rather than relying solely on distance to a centroid. Notably, an optimal performance was achieved with fewer electrodes, making the approach particularly suitable for a practical BCI deployment in surgical environments. In summary, the proposed model offers a solution for one-class Riemannian BCI applications beyond the present anesthetic setting.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency (ANR), with projects PROTEUS (grant ANR-22-CE33-0015-01) and BCI4IA (grant ANR-22-CE19-0016-03).

REFERENCES

- [1] Lotte F et al. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *J Neur Eng*. 2018.
- [2] Khan SS, Madden MG. One-class classification: Taxonomy of study and review of techniques. *CoRR*. 2013;abs/1312.0049.
- [3] Ruff L et al. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE*. 2021;109(5):756–795.
- [4] Yger F, Berar M, Lotte F. Riemannian Approaches in Brain-Computer Interfaces: A Review. *IEEE Trans Neur Syst Rehab*. 2017;25(10).
- [5] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass Brain-Computer Interface Classification by Riemannian Geometry. *IEEE Trans Biomed Eng*. 2012;59(4):920–928.
- [6] Marissens Cueva V et al. One-Class Riemannian EEG Classifier to Detect Anesthesia. In: *Proc. NEC*. 2024.
- [7] Said S, Bombrun L, Berthoumieu Y, Manton JH. Riemannian gaussian distributions on the space of symmetric positive definite matrices. *IEEE Trans. Inf. Theory*. 2017;63(4):2153–2170.
- [8] Rimbert S et al. Detection of motor cerebral activity after median nerve stimulation during general anesthesia (stim-motana): Protocol for a prospective interventional study. *JMIR Res. Protoc*. 2023;12:e43870.
- [9] Rimbert S, Riff P, Gayraud N, Schmartz D, Bougrain L. Median nerve stimulation based bci: A new approach to detect intraoperative awareness during general anesthesia. *Front. Neurosci*. 2019;13.
- [10] Schnitzler A, Salenius S, Salmelin R, Jousmäki V, Hari R. Involvement of primary motor cortex in motor imagery: A neuromagnetic study. *NeuroImage*. 1997;6(3):201–208.
- [11] Delorme A, Makeig S. Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *J Neurosc Meth*. 2004;134(1):9–21.
- [12] Purdon PL, Sampson A, Pavone KJ, Brown EN. Clinical Electroencephalography for Anesthesiologists: Part I: Background and Basic Signatures. *Anesth*. 2015;123(4):937–960.
- [13] Gramfort A et al. Meg and eeg data analysis with mne-python. *Front Neuroinf*. 2013;7:267.
- [14] Pedregosa F et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
- [15] Barachant A et al. Pyriemann. *Zenodo*. 2023.
- [16] Pennec X. Manifold-valued image processing with SPD matrices. In: *Riemannian Geometric Statistics in Medical Image Analysis*. Elsevier, 2020.
- [17] de Surré T, Lotte F, Chevallier S, Yger F. Wrapped Gaussian on the manifold of symmetric positive definite matrices. In: *Proc. ICML*. 2025.
- [18] de Surré T, Yger F, Lotte F, Chevallier S. A probabilistic view on riemannian machine learning models for SPD matrices. In: *Proc. GSI*. 2025.
- [19] Barthélemy Q, Mayaud L, Ojeda D, Congedo M. The Riemannian Potato Field: A Tool for Online Signal Quality Index of EEG. *IEEE Trans Neural Syst Rehab*. 2019;27(2):244–255.
- [20] Pawar N, Barreto Chang OL. Burst suppression during general anesthesia and postoperative outcomes: Mini review. *Front. in Syst. Neurosci.*. 2022;15.
- [21] de Surré T, Chevallier S, Lotte F, Yger F. Geometry-aware visualization of high dimensional symmetric positive definite matrices. *TMLR*. 2025.
- [22] Wang X, Borsoi R, Breloy A, Richard C. Riemannian change point detection on manifolds with robust centroid estimation. 2025. eprint: 2508.18045 (cs.LG).
- [23] de Surré T, Venot T, Corsi MC, Yger F. Interpretability of Riemannian tools used in brain computer interfaces. In: *IEEE MLSP*. 2025, 1–6.

Filling the gap: sex-specific differences in post-task interval dynamics

Michael Leitner¹, Selina C. Wriessnegger¹

¹Institute of Neural Engineering, Graz University of Technology, Graz, Austria

E-mail: s.wriessnegger@tugraz.at

ABSTRACT: Electroencephalography (EEG) is widely used to investigate neural correlates of cognitive workload, fatigue, and stress through changes in oscillatory band power. While many studies have focused on neural activity during task execution, the post-task resting period following cognitive effort remains comparatively underexplored, despite its potential relevance for understanding how cognitive resources are restored. Moreover, little is known about whether neural dynamics differ between male and female participants. In this study, we analyzed EEG band power during the post-task resting interval following a cognitively demanding working memory paradigm. Relative band power (rBP) was examined across multiple frequency bands, scalp regions, and rest epochs to characterize temporal and spatial patterns. Significant sex-related differences were observed in the theta, alpha, and beta bands. Females exhibited higher theta activity, particularly in central regions, whereas males showed elevated alpha band power. Across resting epochs, alpha activity displayed the most pronounced increase, suggesting progressive reallocation of attentional resources. Topographic analyses further revealed region-specific differences in central, parietal, and occipital areas. These findings suggest that post-task interval dynamics exhibit distinct oscillatory characteristics that vary between sexes. Future studies might consider such baseline phases leading to improved EEG-based detection of cognitive states and inform the design of neuroadaptive systems.

INTRODUCTION

Cognitive state changes are reflected in measurable neurophysiological responses that vary across individuals. Factors such as age, biological sex and hormonal cycles (in females) represent important sources of variability, alongside numerous individual differences. Among these factors, sex differences have received increasing attention in recent neuroscience research. Evidence from electroencephalography (EEG) studies indicates that male and female participants may exhibit distinct neurophysiological patterns [1–3] during cognitive processing and mental state changes.

Changes in cognitive state are reflected in specific neurophysiological signatures. In EEG research, spectral characteristics, particularly changes in bandpower (BP) [4, 5] are frequently used to describe alterations associated

with mental workload, fatigue or stress. In particular, increases in frontal theta activity and changes in alpha power have frequently been associated with increased cognitive demand and attentional control. Such patterns are highly relevant for applications such as passive brain-computer interfaces (BCIs) [6–8] and for advancing the understanding of mechanisms underlying EEG dynamics. While BP changes during task execution have been widely described and resting-state EEG has been extensively investigated [3, 9–11], comparatively little attention has been given to the dynamics that occur after task completion, in the between-task phases. However, spatial and temporal characteristics of inter-block interval dynamics remain poorly understood, especially considering sex-specific patterns. Most studies focus on task periods, whereas the post-task resting intervals remain left behind.

Oscillatory patterns observed in electroencephalography (EEG) provide important insights into the brain's activity during cognitive processes. Different frequency bands have been associated with distinct aspects of cognitive control and information processing. In particular, theta-band activity (4–7 Hz), especially over frontal regions, has been widely linked to cognitive control, working memory demands and increased mental workload, reflecting the recruitment of executive control mechanisms during demanding tasks [2, 12, 13]. In contrast, alpha-band oscillations (8–12 Hz) are commonly associated with attentional regulation and inhibitory control, where increases in alpha power are thought to reflect the suppression of task-irrelevant information or neural inhibition of competing processes [2, 3, 14]. Beta-band activity (13–30 Hz) has frequently been related to task engagement, sensorimotor processing and active maintenance of cognitive states, reflecting ongoing information processing during sustained task performance [1, 15]. Together, these oscillatory dynamics provide a useful framework for interpreting EEG signatures of cognitive workload and mental state transitions.

Working memory tasks represent a common paradigm for investigating cognitive workload and associated neural oscillations. Cognitive load and stress are frequently studied constructs in this context. Increasing evidence suggests the presence of sex-specific patterns in neurophysiological responses to cognitive stress [5, 16, 17], pointing to evidence of spatial and functional differences in brain oscillations between sexes. However, potential sex-specific variations in neurophysiological patterns

raise the question of whether fatigue or mental workload affect males and females differently during resting intervals. Such differences could have implications for adaptive technologies, fatigue monitoring systems, and therapeutic interventions that aim to support cognitive resting and regeneration. Previous studies suggest that working memory processing can show sex-related differences in neural dynamics. EEG investigations have reported differences in oscillatory responses and network variability between males and females during visuo-spatial working memory tasks [18], while event-related oscillatory activity appears to vary with both age and sex [19]. Behavioral studies further indicate that environmental factors may interact with sex-related differences in working memory performance [20].

In particular, the resting interval remains underexplored. This period reflects the reorganization and restoration of cognitive and attentional resources following sustained mental effort. Understanding post-task resting dynamics is important because adaptive cognitive systems and neurotechnology applications increasingly rely on accurate detection of mental states and their transitions. Moreover, resting processes may influence subsequent task performance and overall cognitive resilience. Several findings suggest that moderate changes in BP dynamics may occur during resting interval, yet systematic investigations remain scarce. Existing studies often rely on small sample sizes or focus primarily on task-related effects rather than resting dynamics. Consequently, the functional significance of resting-related EEG patterns and their potential sex-specific characteristics remains unclear.

The present work aims to provide an initial exploration of sex-specific post-task resting characteristics in EEG following cognitively demanding tasks. By analyzing BP dynamics during inter-task rest phases, this study investigates whether systematic differences in resting-related brain activity emerge between male and female participants.

MATERIALS AND METHODS

Signal processing was performed in Python 3.10 using the libraries MNE-Python [21], NumPy, SciPy, and scikit-learn.

Participants: The dataset consisted of 30 healthy participants with an equal sex distribution (15 male, 15 female). Male participants had a mean age of 26.1 ± 3.0 years, while female participants had a mean age of 22.9 ± 3.2 years. The study was approved by the local ethics committee and all participants provided informed consent.

Experimental setup and paradigm: The electroencephalogram (EEG) was recorded using a 32-electrode setup with a LiveAmp amplifier (Brain Products GmbH, Gilching, Germany) at a sampling rate of 500 Hz. Participants performed a modified Sternberg working memory task [22] designed to induce varying levels of cognitive load and stress. In each trial, a sequence of randomly 3–5

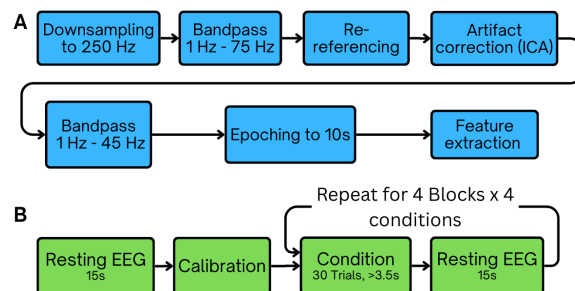


Figure 1: **A:** Pre-processing pipeline of EEG signals during post-task resting phase, **B:** Experimental procedure of paradigm

digits was presented for 2 s. Participants were required to respond under time pressure, with the response deadline set to $0.9 \times$ the individual reaction time obtained prior to the experiment. The response was a single digit letter, which was present for the calibration time. Each block consisted of 30 trials followed by a resting period during which EEG activity was recorded. To maintain participant engagement, feedback on task performance was displayed after each block, intentionally underestimating the actual performance by 15%. In a short try-out (calibration), 10 equal trials without any constrictions (time pressure or environmental sounds) were presented. The protocol is visualized in figure 1. The response time was then used as calibration reaction time. Four experimental conditions were implemented: baseline (simple sternberg without manipulation), timing pressure, noise distraction and timing pressure + noise distraction. The experiment consisted of four blocks per condition (16 blocks in total). Each block consisted of 30 trials. The order of conditions was pseudo-randomized inside a single block, the order of the blocks was also pseudo-randomized to mitigate potential learning effects. Prior to the task, participants completed a baseline recording of 15 s. Additionally, resting phase followed each block to capture neural dynamics. EEG segments used for analysis were extracted from the beginning of each rest phase. For completeness, we did explain the full paradigm but did not analyze the active paradigm phases as the focus was laid on the inter-block phases.

Data processing: EEG pre-processing was performed using a multi-step pipeline. The signals were first downsampled to 250 Hz to improve computational efficiency. A band-pass filter (4th-order Butterworth) between 1–75 Hz was then applied, and the data were re-referenced to the common average. Independent component analysis (ICA) using the extended Infomax ICA was subsequently performed with 15 components. Artifact components were identified and removed using a combination of automated and manual procedures. Muscle artifacts were detected automatically based on spatially isolated patterns and power spectral density (PSD) slope characteristics. We used mne-python implemented detection algorithms with a threshold of 0.85 to remove muscle artifact. Eye artifacts were identified manually based on frontal topography and characteristic temporal patterns in the component time courses which resulted in the re-

removal of 2-3 components per participant. Following artifact removal, the signals were filtered again between 1–40 Hz. For the analysis of the post-task intervals, EEG recordings were segmented into 10 s epochs starting 2 s after the onset of each resting phase (2–12 s window). The pre-processing pipeline as well as the experimental timeline can be seen in figure 1.

Feature extraction: Power spectral density (PSD) was estimated using Welch's method with segment lengths of 500 samples, 50% overlap, and an FFT length of 1024. To account for inter-subject variability in overall power, spectral estimates were normalized by the average power across the full spectrum, resulting in rBP measures. rBP was computed for five canonical EEG frequency bands: delta (0.5–3.5 Hz), theta (4–7.5 Hz), alpha (8–12.5 Hz), beta (13–29.5 Hz), and gamma (30–40 Hz). To assess global band power changes independent two-sample *t*-tests were performed to evaluate sex-specific differences in electrode-average rBP. To account for multiple comparisons, false discovery rate (FDR) correction was applied. Topographical maps (topomaps) were used to visualize the spatial distribution of rBP differences between sexes across electrodes.

In addition, several regions of interest (ROIs) were defined to display the temporal evolution of the post-task period for each frequency band. Electrodes were grouped into four regions: frontal (F3, Fz, F4, FCz), central (C3, Cz, C4, CPz), parietal (P3, Pz, P4), and occipital (O1, Oz, O2). For each region, the average rBP per interval was calculated separately for female and male participants.

RESULTS

Row one of figure 2 displays the spatial distribution of rBP differences across the scalp, shown as the difference (female–male). In the delta band, males show higher activity in central, parietal and occipital regions, whereas females exhibit stronger activation in right dorso-lateral as well as left frontal and temporal areas. In the theta band, overall activity is higher in females, particularly in central regions with a pronounced increase around electrode C4. Additional increases are observed in right parieto-temporal and left temporal areas. In the alpha band, higher rBP is observed in frontal regions for females. Most non-frontal regions show higher values in males, with local maxima in central areas extending toward the left temporal region, the right dorso-lateral region, and the right parietal area. In the beta band, higher activity is visible in central regions with a slight lateralization toward the left hemisphere, as well as in the right parietal and occipital areas. For the gamma band, no notable spatial differences were observed between sexes. Figure 2 shows the sex-specific differences in band power across all frequency bands in row two. Statistical testing revealed significant differences in the theta ($p = 0.0123$), alpha ($p = 0.024$), and beta ($p < 0.001$) bands. No significant differences were observed in the delta or gamma bands. Alpha rBP is higher

in male participants. Rows 3–6 of Figure 2 displays the regional rBP values across epochs. The largest differences between sexes were observed in the central theta, alpha, and beta bands. Delta rBP shows an overall decreasing trend across epochs, accompanied by relatively high variability across regions. Theta activity remains mostly stable over time. In central theta, females initially show higher rBP values, while the difference decreases toward later epochs. Additionally, theta shows an initial increase in females within the frontal, parietal, and occipital ROIs during the first epochs. Alpha rBP shows a gradual increase across epochs in all regions, with a slightly stronger increase in the parietal ROI and a consistent offset between sexes in the central region. Beta activity remains largely stable across epochs, but notable differences between sexes are visible in the central ROI. In the gamma band, an initial increase is observed in males in the frontal, parietal, and occipital regions, followed by a slight decrease in the central and parietal ROIs.

DISCUSSION

In this work, we investigated resting interval dynamics in EEG rBP. Significant sex-specific differences were observed in the theta, alpha, and beta bands. Male participants exhibited higher alpha rBP, whereas females showed elevated theta activity. Topographic differences were most pronounced in central regions across the theta, alpha, and beta bands, with additional effects observed in parietal theta and occipital beta activity. Across the resting epochs, alpha rBP showed the most pronounced increase over time. No significant differences were observed in the gamma band, while delta activity displayed a gradual decrease over time with relatively indistinct spatial patterns.

Theta activity is typically elevated in frontal regions and is associated with cognitive control, working memory demands, and mental workload [12, 13]. In the present study, no pronounced increase in frontal theta activity was observed across the resting time course. Instead, theta activity was predominantly centered in central regions. Within this region, females exhibited consistently higher theta rBP compared to males, and this difference remained relatively stable throughout the experiment. In parietal regions, the difference between sexes was smaller and became more pronounced during later experimental blocks (blocks 7–8). Additionally, frontal, parietal, and occipital regions showed an initial increase in theta rBP during the first experimental block. This early shift in theta activity may reflect the restoration of executive or memory-related resources following the task. Elevated central theta activity may also indicate prolonged cognitive engagement or a slower resting of cognitive control processes.

Alpha oscillations are commonly associated with attentional regulation, inhibitory control, and resting processes, but have also been linked to mental fatigue and workload [2, 3, 8]. In the present study, alpha rBP exhib-

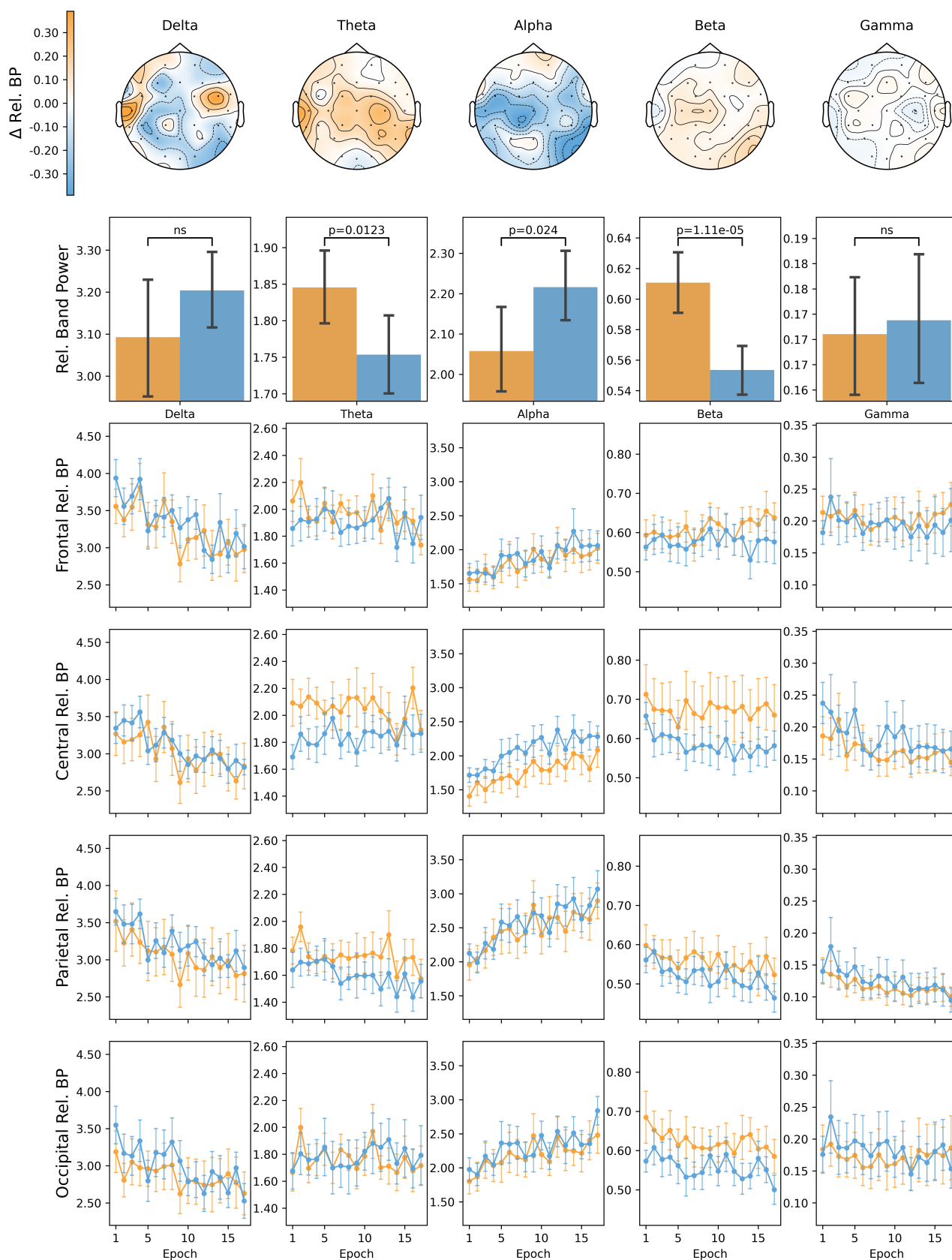


Figure 2: Overall results displaying observed sex-specific differences. Blue represents data or tendencies towards males, red female data or tendencies towards females. All bars indicating derivative measure are representing the standard error. Row 1: Topographical difference (female (orange) - male (blue)), Row 2: Global rBP power comparison with indicated significance level (ns = non-significant, p-value if significant), Rows 3-6: Regional rBP across epochs for female and male participants. Rows represent anatomical regions of interest (frontal, central, parietal, occipital), and columns correspond to EEG frequency bands. Error bars indicate the standard error of the mean (SEM).

ited a gradual increase across the post-task phases, with parietal and occipital regions generally showing higher values than frontal or central regions. This pattern may reflect ongoing visual processing in parietal and occipital areas, while increases in frontal and central regions could indicate progressive mental fatigue and a gradual shift in attentional allocation. In the frontal regions, there also seems to be a slight shift towards females. Towards central left areas, males showed consistently higher alpha rBP compared to females, although both groups followed a similar temporal trend. These dynamics may indicate an initially strong engagement of visual and attentional systems during the task, followed by a gradual disengagement with the proceeding task.

Beta oscillations have frequently been associated with task engagement and the maintenance of cognitive states [1, 15]. In the present study, beta activity showed a slightly increasing trend across the experimental timeline which is parallel to maintenance of task-related cognitive states. Central regions exhibited higher beta rBP in females throughout the task, which might indicate a stronger maintenance of task-related engagement. In contrast, parietal and occipital regions showed patterns that could reflect a gradual disengagement from visual processing. In occipital regions in particular, females tended to show slightly stronger beta activity compared to males, which is linked to a potentially higher level of sustained task engagement.

Gamma band activity did not show significant differences between sexes. However, during the first competitive block involving feedback, a transient increase in gamma activity was observed. This initial spike may reflect increased cognitive resource recruitment during the first exposure to the competitive task condition. Taken together, these observations suggest that resting interval exhibit distinct oscillatory dynamics in male and female participants. Such differences may have important implications for experimental paradigms and applications that rely on the assessment of mental workload or cognitive stress, such as confrontation therapy or adaptive cognitive systems. Ignoring these dynamics could lead to misinterpretations of mental workload or fatigue, particularly when sex-specific neural patterns are not taken into account. Adaptive systems and neurotechnology applications may therefore benefit from incorporating sex-specific resting characteristics into models of cognitive state monitoring. Beside our interesting findings some limitations should be considered for future work. Although the dataset was balanced with respect to sex, a larger sample size would further improve statistical robustness. The inclusion of behavioural analysis would also provide clearer understanding of the detailed rBP changes. Additionally, the study employed a modified experimental paradigm within a controlled laboratory environment. Future research could investigate resting dynamics in more ecologically valid settings in order to strengthen the link to real-world therapeutic or neuroadaptive applications. Further work may also explore longer resting periods

and incorporate additional physiological measures, such as hormonal monitoring (e.g., cortisol levels) or multimodal neurophysiological recordings combining EEG with fNIRS or ECG.

CONCLUSION

This study investigated sex-specific differences in EEG rBP during the post-task intervals following a cognitively demanding working memory task. Significant differences were observed in theta, alpha, and beta oscillations, with females exhibiting stronger theta activity and males showing higher alpha rBP. Spatial analyses further revealed distinct topographic patterns, particularly in central, parietal, and occipital regions. These findings suggest that resting dynamics following cognitive effort exhibit measurable oscillatory characteristics that differ between male and female participants. Importantly, the results highlight the relevance of considering resting interval, which have received comparatively little attention in EEG research. Future work should investigate longer periods, larger cohorts, and multimodal physiological measurements to further characterize the mechanisms underlying cognitive rest.

REFERENCES

- [1] Engel AK, Fries P. Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology*. 2010;20(2):156–165.
- [2] Klimesch W. Eeg alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain research reviews*. 1999;29(2-3):169–195.
- [3] Klimesch W. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*. 2012;16(12):606–617.
- [4] Wriessnegger S, Leitner M, Kostoglou K. The brain under pressure: Exploring neurophysiological responses to cognitive stress. *Brain and Cognition*. 2024;182:106239.
- [5] Wriessnegger S, Leitner M, Kostoglou K. Neurophysiological features of mental stress and their sex-specific differences. 2025.
- [6] Käthner I, Wriessnegger SC, Müller-Putz GR, Kübler A, Halder S. Effects of mental workload and fatigue on the p300, alpha and theta band power during operation of an erp (p300) brain–computer interface. *Biological Psychology*. 2014;102:118–129.
- [7] Gregory SE, Wang H, Kessler K. Eeg alpha and theta signatures of socially and non-socially cued working memory in virtual reality. *Social Cognitive and Affective Neuroscience*. 2022;17(6):531–540.
- [8] Herbst SK et al. Alpha power indexes working memory load for durations. *iScience*. 2026;29(1).
- [9] Raichle ME. The brain’s default mode network. *Annual review of neuroscience*. 2015;38(1):433–447.
- [10] Barry RJ, Clarke AR, Johnstone SJ, Magee CA, Rushby JA. Eeg differences between eyes-closed and

eyes-open resting conditions. *Clinical Neurophysiology*. 2007;118(12):2765–2773.

[11] Kang JH, Bae JH, Jeon YJ. Age-related characteristics of resting-state electroencephalographic signals and the corresponding analytic approaches: A review. *Bio-engineering*. 2024;11(5).

[12] Cavanagh JF, Frank MJ. Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*. 2014;18(8):414–421.

[13] Brouwer AM, Hogervorst MA, Van Erp JB, Heffelaar T, Zimmerman PH, Oostenveld R. Estimating workload using eeg spectral power and erps in the n-back task. *Journal of neural engineering*. 2012;9(4):045008.

[14] Magosso E, Ricci G, Ursino M, et al. Alpha and theta mechanisms operating in internal-external attention competition. *Journal of Integrative Neuroscience*. 2021;20(1):1–19.

[15] Feltman KA, Bernhardt KA, Kelley AM. Measuring the domain specificity of workload using eeg: Auditory and visual domains in rotary-wing simulated flight. *Human Factors*. 2021;63(7):1271–1283.

[16] Wriessnegger SC, Lorenzer L, Kostoglou K. Neurocardiac signatures of acute mental stress: A sex-comparative study. *Frontiers in Neuroscience*. 2025;19:1633295.

[17] Leitner M, Wriessnegger SC. Detecting mental stress using eeg and biosignals: Towards a neuro-adaptive system. In: 2025 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine). 2025, 1125–1130.

[18] Jouzizadeh M. Eeg-assessed network and signal variability features in males and females during a visuospatial task. Ph.D. dissertation. Université d'Ottawa/University of Ottawa. 2024.

[19] Yener G, Kıyı İ, Düzenli-Öztürk S, Yerlikaya D. Age-related aspects of sex differences in event-related brain oscillatory responses: A turkish study. *Brain Sciences*. 2024;14(6):567.

[20] Rhee JH, Schermer B, Lee KH. Effects of indoor nature density and sex differences on working memory. *Current Psychology*. 2024;43(39):30677–30690.

[21] Gramfort A et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013;7(267):1–13.

[22] Sternberg S. High-speed scanning in human memory. *Science*. 1966;153(3736):652–654.

Is User-Independent MI-BCI Performance Influenced by Gender Proportion in Training Users?

Simon Kojima¹, Fabien Lotte¹

¹Inria Centre at the University of Bordeaux, Talence, France

E-mail: simon.kojima@inria.fr

ABSTRACT: In motor-imagery brain-computer interfaces (MI-BCI) research, how user gender influences BCI classification performance and ERD/S responses is not yet fully understood. Thus, investigating the impact of gender on BCI systems is an important step toward realizing BCIs that are accessible to all users. In this study, we examined how varying the proportion of female users in the training data in cross-user MI-BCI affects BCI classification performance. The results (on 139 users from 2 data bases) revealed positive correlations (0.33–0.71) between the proportion of female users in the training data and BCI performance, and showed that including more female users in the training data was associated with performance improvements of up to 5.5%. These findings indicate that machine-learning models in MI-BCI may be influenced by gender-related biases and highlight the importance of considering gender distribution in the design of future BCI research.

INTRODUCTION

Motor Imagery (MI)-based brain-computer interfaces (BCIs) offer intuitive control and a wide range of potential applications; however, it is well known that approximately 15–30% of users fail to achieve sufficient control accuracy, a phenomenon referred to as BCI inefficiency [1]. Although various factors have been investigated, the fundamental causes of this issue remain unclear, and no definitive solution has been established. MI-BCIs operate by detecting, using machine learning, the power decreases and increases in the μ (8–13 Hz) and β (15–30 Hz) bands, known as event-related desynchronization/synchronization (ERD/S), elicited by MI of body-part movements.

Multiple factors affecting BCI performance and ERD/S strength have been reported, particularly biological sex or gender-related factors, which have been shown to influence ERD amplitude and BCI performance. Randolph et al. reported that ERD amplitudes were larger in females [2]. Similar tendencies have been reported in [3, 4], where higher MI-BCI performance in females was also observed. The higher MI-BCI performance in females was observed in other studies [5, 6]. Furthermore, an fMRI study on post-stroke patients reported greater activation in the sensorimotor cortex in females compared with males [7]. Additionally, it has been suggested that

the interaction between the genders of the experimenter and the user may influence MI-BCI performance [8]. Another study reported slightly better, though not statistically significant, MI-BCI performance in female users, and also showed that features learned by deep learning models can be used to classify user gender [9]. In contrast, [10–12] reported no significant gender differences. Thus, there is currently no clear consensus on how gender affects MI-BCI performance. Nevertheless, beyond BCI research, it is well known that various machine-learning algorithms may embed gender-related biases [13]. Therefore, a careful investigation of gender effects is essential for realizing BCIs that are accessible to all users.

Previous studies mainly compared male vs. female performance but did not examine how the composition of the training data affects performance. In this study, we investigated an aspect that has not been sufficiently examined so far, namely how imbalances in the number of male and female samples in the training data - for cross-user MI-BCIs - affect the BCI performance of male and female users. In addition, by quantifying the variability of ERD/S responses under these conditions, we also examined potential gender differences in the ERD/S variability to interpret the performance difference.

MATERIALS AND METHODS

Dataset Description: We used two open datasets, (A) Dreyer2023 [14] and (B) Lee2019 [15], which were accessed via the Python package MOABB [16]. For both datasets, only EEG channels were used, and participant sex (male/female) was obtained from the dataset meta-data and is referred to as gender in the remainder of the paper.

(A) Dreyer2023: It contains EEG recordings from 87 users performing left- and right-hand motor imagery (MI) tasks [14]. For each user, a single-session EEG recording consisting of six runs of MI tasks was provided. The first and second runs constituted the acquisition phase, during which data were collected for classifier training, whereas runs 3 to 6 comprised the online phase, during which users received real-time feedback from the BCI. In each run, 20 trials were performed per class, i.e., left- and right-hand MI. Among the 87 users, two (4 and 59) were excluded from further analysis due to excessive noise in the recordings (4) and missing runs (59), leaving 85 users

(41 females and 44 males) for analysis. All 27 available EEG channels were used (cf. [14]). The original sampling frequency was 512 Hz.

(B) Lee2019: It contains EEG recordings from 54 users (23 females and 31 males) performing left- and right-hand motor imagery (MI) tasks [15]. For each user, two-session EEG recordings consisting of two runs (a training run and a test run) of MI tasks were provided. As the test run does not include true class labels (see the original paper for details [15]), only the training run was used in this study. Each run consisted of 100 trials (50 trials per class). All 62 available EEG channels were used (cf. [15]). The original sampling frequency was 1000 Hz.

Classification: In this study, the cross-user classification task was evaluated for each test user across varying proportions of female users in the training data.

For each dataset, one user was selected as the test user U_t , and the training data were constructed by selecting N_t users from the remaining users such that the proportion of female users in the training data was $R_f \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$. Here, N_t was set to 40 for Dreyer2023 and 20 for Lee2019. The training and test data were generated as follows.

For each condition defined by the test user U_t and the female ratio R_f , the following steps were performed:

1. From the users excluding U_t , randomly select N_t users as the proportion of female users equals R_f .
2. Apply preprocessing (described later) to the data of each selected user and compute the covariance matrix for each trial. After applying Ledoit-Wolf regularization [17] to the covariance matrices, Riemannian Procrustes Analysis (RPA) [18] was applied to reduce between-user variability. In this study, only the recentering and rescaling steps of RPA were used, normalizing the data such that the geometric mean became the identity matrix and the dispersion around it was equal to 1. The rotation step of RPA was not applied.
3. Concatenate the data from all selected users to form the training dataset.
4. Apply the same preprocessing to the data of the test user U_t , compute the covariance matrices, and apply RPA. To simulate a realistic BCI usage scenario, RPA was applied sequentially; i.e., for the n -th trial, RPA was estimated and applied using only the data from the first n trials.
5. Train the classifier using the training data and classify the data of the test user. As the classifier, Riemannian Tangent Space + Logistic Regression (TSLR) [19] was used. For Logistic Regression, the 1bfgs solver with an L2 penalty ($C = 1.0$) and $max_iter = 100$ was employed.
6. The above procedure was repeated three times per test user, and the results were averaged.

Each user in the dataset was selected once as the test user, and the above procedure was performed for all conditions of R_f (the proportion of female users in the training data) for each test user.

The data preprocessing was performed as follows. First, a fourth-order Butterworth filter was applied bidirectionally in the 8–30 Hz band, and the MI task interval was extracted. The extraction interval was 0.5–4.5 s for Dreyer2023 and 0.5–3.5 s for Lee2019, post MI instruction cue. The data were then resampled to 128 Hz. Since Lee2019 contains data from two sessions, preprocessing and RPA were applied independently to each session, after which the two sessions were concatenated to form the data for each user. Classification performance was evaluated using accuracy. No artifact removal was applied to either the classification or the variability calculation.

Variability: For each test user, we examined (i) how variable the training data were with respect to that user (test-referenced variability) and (ii) how much variability existed within each gender group (gender-wise variability). Variability was evaluated for both the temporal waveform and the spatial distribution. To assess variability, we used the class-wise mean ERD/S pattern of each user. In this study, variability was uniformly defined as the average distance between each user's ERD/S pattern and a reference pattern M . The definition of the reference pattern M was adapted depending on the analysis setting. Specifically, for test-referenced variability, the ERD/S pattern of the test user served as the reference pattern, whereas for gender-wise variability, the representative pattern for each gender group was used.

First, preprocessing was applied to each user's data. A fourth-order Butterworth filter in the 1–45 Hz band was applied bidirectionally, and time–frequency representations were estimated from 1 to 45 Hz with a resolution of 1 Hz using the Discrete Prolate Spheroidal Sequences (DPSS) multitaper method [20], followed by decimation by a factor of 2. ERD/S was then computed using the interval -2.0 – 0.0 s as the baseline. Subsequently, ERD/S was averaged across trials for each class and user, and the resulting class-wise mean ERD/S pattern was used as the representative data for each user.

Test-referenced variability: First, we evaluated how much the ERD/S patterns of the N_t training users selected for a given test user U_t varied around U_t . This metric quantifies how varied the training users' ERD/S patterns are from the test user's pattern.

For temporal variability, the corresponding channel for each class (C4 for left-hand imagery and C3 for right-hand imagery) was extracted and their ERD/S averaged over the 8–30 Hz band. This resulted in a temporal vector $x_t \in \mathbb{R}^T$ for each user, where T denotes the number of time samples.

The temporal variability v_t was quantified using DTW (Dynamic Time Warping), a distance measure based on nonlinear alignment between two time series [21]. It was defined using the reference pattern M , given by the temporal waveform of the test user ($M = x_t^{c, U_t}$), as follows:

$$v_t = \frac{1}{N_c} \frac{1}{N_t} \sum_{c=1}^{N_c} \sum_{i=1}^{N_t} \text{DTW}(x_t^{c,i}, M), \quad M = x_t^{c, U_t} \quad (1)$$

Here, N_c denotes the number of classes, c denotes the class (left-hand and right-hand imagery), and i indexes the training users.

Next, spatial variability was evaluated. All EEG channels were used, and ERD/S data were averaged over the 8–30 Hz band and subsequently over time. This resulted in a spatial vector $x_s \in \mathbb{R}^C$ for each user, where C is the number of EEG channels.

The spatial variability v_s was defined using the reference pattern M as the spatial pattern of the test user, $M = x_s^{c, U_t}$, as follows:

$$v_s = \frac{1}{N_c} \frac{1}{N_t} \sum_{c=1}^{N_c} \sum_{i=1}^{N_t} \arccos \left(\frac{x_s^{c,i} \cdot M}{\|x_s^{c,i}\| \|M\|} \right), \quad M = x_s^{c, U_t} \quad (2)$$

Here, $d = \arccos((w_1 \cdot w_2) / (\|w_1\| \|w_2\|))$ compute the angular distance between two vectors (w_1 and w_2).

These two metrics represent the variability around the test user U_t in terms of the temporal waveform and spatial distribution, respectively.

Gender-wise variability: Next, for each dataset, we evaluated the variability within the male group, within the female group, and within the entire user population. Here, the reference pattern M was defined as the representative ERD/S pattern of each group.

For the temporal waveform, the reference pattern M in Eq. (1) was defined as the Euclidean mean over the set of users belonging to the group $U \in \{U_{\text{male}}, U_{\text{female}}, U_{\text{all}}\}$:

$$M = \frac{1}{N_U} \sum_{i \in U} x_t^{c, U} \quad (3)$$

where N_U is the number of users in the given group U . For the spatial distribution, the reference pattern M in Eq. (2) was defined as the Fréchet mean on the hypersphere for the set of spatial vectors:

$$M = \operatorname{argmin}_{\|m\|=1} \sum_{i \in U} \arccos \left(\frac{x_s^{c,i} \cdot m}{\|x_s^{c,i}\| \|m\|} \right) \quad (4)$$

Statistical Analysis: In the correlation analysis, we evaluated the relationships between the proportion of female users in the training data, R_f , and each metric: machine-learning classification accuracy and test-referenced variability metrics. Specifically, for each test user, repeated-measures data obtained under different R_f conditions were used to analyze the correlations between R_f and (i) classification accuracy, (ii) temporal waveform variability, and (iii) spatial distribution variability, using repeated measures correlation. In this study, a total of 12 correlation analyses were conducted (2 datasets $\times$ 2 genders $\times$ 3 metrics: accuracy, temporal variability, and spatial variability). Therefore, multiple comparison correction using the Benjamini–Hochberg (BH) procedure was applied to the resulting p -values, with a significance level of 0.05.

In addition, to examine whether each metric changed significantly with variations in the female ratio R_f , the

Table 1: The results of the post hoc comparisons using the Wilcoxon signed-rank test are presented. A and B denote the levels of the female ratio R_f that were compared. p denotes the p -values corrected using the Benjamini–Hochberg procedure. With a significance level of 0.05, results showing significant differences are indicated by bold p -values. G., TempVar and SpatVar denote gender, temporal variability and spatial variability, respectively.

Dataset	Modality	G.	A	B	p
Dreyer 2023	Accuracy	F	0.0	0.5	7.7×10^{-2}
			0.5	1.0	8.7×10^{-2}
			0.0	1.0	5.5×10^{-2}
	TempVar	F	0.0	0.5	2.2×10^{-4}
			0.5	1.0	6.9×10^{-3}
			0.0	1.0	4.5×10^{-11}
		M	0.0	0.5	7.8×10^{-6}
			0.5	1.0	6.7×10^{-3}
			0.0	1.0	2.3×10^{-12}
	SpatVar	F	0.0	0.5	1.9×10^{-6}
			0.5	1.0	8.4×10^{-8}
			0.0	1.0	1.1×10^{-7}
		M	0.0	0.5	1.3×10^{-4}
			0.5	1.0	4.9×10^{-5}
			0.0	1.0	4.1×10^{-5}
Lee 2019	Accuracy	F	0.0	0.5	8.3×10^{-3}
			0.5	1.0	4.1×10^{-4}
			0.0	1.0	6.0×10^{-5}
		M	0.0	0.5	2.7×10^{-3}
			0.5	1.0	7.7×10^{-3}
			0.0	1.0	7.8×10^{-4}

following analysis was conducted. First, for the total of 12 conditions (2 datasets $\times$ 2 genders $\times$ 3 metrics), repeated-measures data for each test user were analyzed using a repeated-measures ANOVA with the female ratio R_f as a within-subject factor. This allowed us to test whether there was a significant main effect of R_f on each metric. The resulting 12 p -values were corrected for multiple comparisons using the Benjamini–Hochberg (BH) procedure, and statistical significance was determined with a significance level of $\alpha = 0.05$.

For the conditions in which a significant main effect was observed in the ANOVA, post hoc comparisons were performed for all pairwise combinations of the five levels of R_f (10 pairs in total) using Wilcoxon signed-rank tests. The p -values obtained from these comparisons (10 values in total) were also corrected for multiple comparisons using the BH procedure.

RESULTS

Fig. 1 shows bar plots of the results for classification accuracy and each test-referenced variability metric across different proportions of female users in the training data R_f , separately for male and female test users. The R values in the figure denote the correlation coefficients obtained by repeated measures correlation.

For Dreyer2023, a positive correlation between R_f

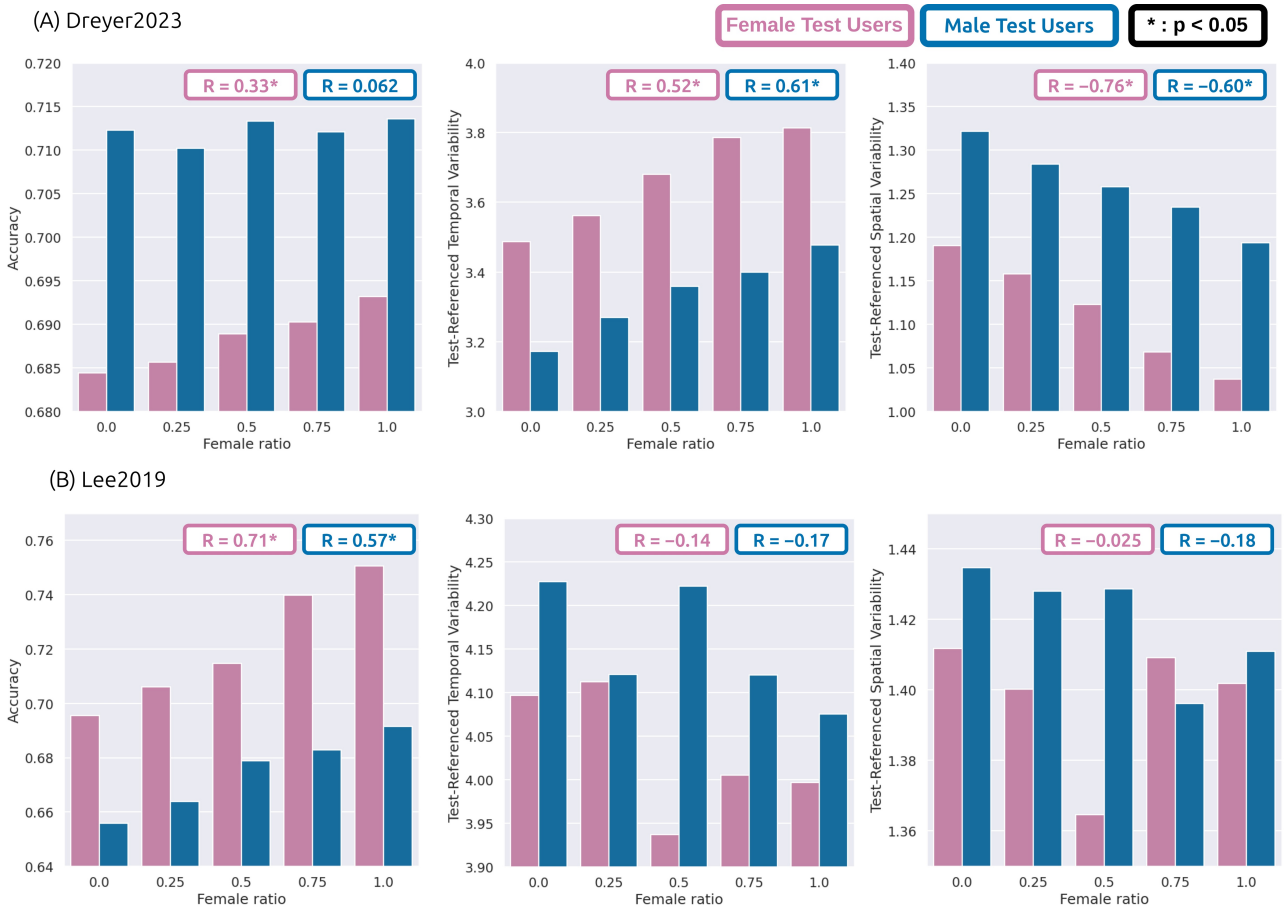


Figure 1: Classification accuracy as well as test-referenced temporal and spatial variability are shown for different values of the female ratio R_f in the training data. The correlation coefficients R obtained by repeated measures correlation are also indicated in the figure. An asterisk (*) attached to a correlation coefficient denotes that the correlation was significant at $\alpha = 0.05$ after applying the Benjamini-Hochberg correction to the obtained p -values. Note that the scale of the y-axis differs across the plots.

and classification accuracy was observed for female test users. For temporal variability, moderate positive correlations between R_f and variability were observed for both male and female test users. In contrast, for spatial variability, negative correlations between R_f and variability were observed for both genders.

For Lee2019, regarding classification accuracy, positive correlations between R_f and accuracy were observed for both male and female test users. On the other hand, for both temporal variability and spatial variability, no significant correlations were found for either gender.

A repeated-measures ANOVA was conducted to examine the main effects of the female ratio R_f on each metric. For Dreyer2023, significant main effects were observed for classification accuracy (female test users; $p = 1.5 \times 10^{-2}$), temporal variability (female; $p = 6.0 \times 10^{-11}$, male; $p = 1.2 \times 10^{-7}$), and spatial variability (female; $p = 1.2 \times 10^{-11}$, male; $p = 2.7 \times 10^{-6}$). In contrast, for Lee2019, significant main effects were observed only for classification accuracy (female; $p = 2.4 \times 10^{-8}$, male; $p = 3.1 \times 10^{-6}$).

In the post hoc comparisons, among the five levels of R_f , the results for the three pairs 0.0 vs. 0.5, 0.5 vs. 1.0, and 0.0 vs. 1.0, out of ten pairs, are reported in Table 1.

Significant differences were found in all conditions except for the classification accuracy in Dreyer2023 when female users were used as the test users. Additionally, significant differences were found in 0.75 vs. 1.0 in Lee2019 regardless the gender of the test user (female; 4.8×10^{-2} and male; 4.3×10^{-2}).

In Lee2019, the highest classification accuracy was obtained when the model was trained using only female data ($R_f = 1.0$), regardless of the gender of the test user. Even when male users were used as the test users, the highest accuracy was achieved when training was performed using only female data. When the female user was the test user, the accuracy was improved by 5.5% when $R_f = 1.0$ against $R_f = 0.0$.

Fig. 2 shows the gender-wise variability for each dataset, and the distances between the centroids of the female and male groups. In Dreyer2023, males exhibited lower variability in temporal variability, whereas females exhibited lower variability in spatial variability. In contrast, in Lee2019, females tended to show lower variability for both temporal and spatial metrics.

DISCUSSION

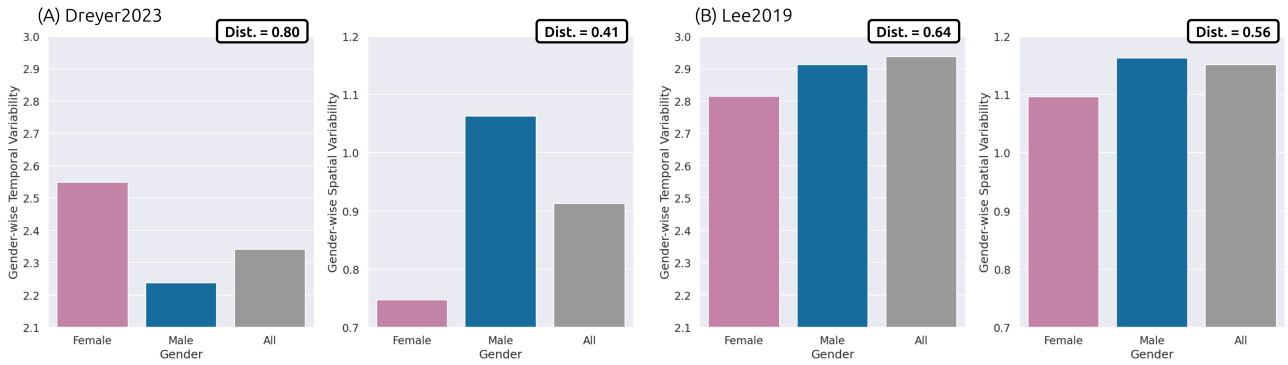


Figure 2: Results of the gender-wise variability analysis. Temporal and spatial variability are shown for the male group, the female group, and all users. The brown box in the upper-right corner of each subfigure shows the distance between the centroids of the female and male groups (DTW for temporal variability and angular distance for spatial variability). The y-axis scales are identical across datasets for each metric.

First, this study revealed that, in cross-user MI-BCI classification, bias may arise depending on the gender ratio of users included in the training data. Both the correlation analysis between the female ratio R_f and BCI performance (Fig. 1) and the differences in BCI performance across R_f conditions within each setting (Table 1) showed significant correlations (3 out of 4 conditions) and significant differences (2 out of 4 conditions). These findings provide important evidence that the gender distribution can influence performance in the design and experimentation of cross-user MI-BCI systems.

The results suggested a tendency toward improved performance when a higher proportion of female user data was included in the training set. Significant positive correlations between R_f and BCI performance were observed for female users in Dreyer2023 and for both male and female users in Lee2019. In particular, for female users in Lee2019, a strong correlation coefficient of 0.71 was obtained. In addition, the results in Table 1 indicate that, in Lee2019, training the classifier using only female users ($R_f = 1.0$) yielded significantly higher performance than the other R_f conditions. Moreover, for both male and female users in Dreyer2023, although no significant differences were observed, the highest average accuracy among all R_f conditions was also obtained when $R_f = 1.0$. As discussed in the Introduction, several studies [3–6] have reported that females tend to exhibit higher MI-BCI performance and larger ERD amplitudes; our findings further suggest the potential advantage of using more female data for training.

The results suggest that differences in ERD/S responses may exhibit a hierarchical scale structure. In Fig. 2, the within-group variability for both males and females was consistently larger than the centroid distance between the two gender groups ($v_f > d$ and $v_m > d$, where d denotes the centroid distance and v_f and v_m denote the variability of females and males, respectively). This implies that inter-individual differences exceed gender differences. Furthermore, the centroid distance between gender groups was larger than the difference in within-group variability ($d > |v_f - v_m|$) in all conditions. This

indicates that the direct gender difference in mean responses may be larger than the difference in variability within each gender group. Taken together, these findings suggest a hierarchical scale structure of inter-individual differences $>$ gender differences $>$ differences in within-gender variability. In other words, although inter-individual differences are likely to have the largest impact on BCI performance, direct gender differences may exert a greater influence than within-gender variability differences. Therefore, when designing cross-users BCI systems that account for gender effects, ensuring robustness to differences in mean responses between genders may be more important than addressing differences in variability within each gender group. However, in this study, each variability metric was estimated as a single group-level value using all available data, and thus, statistical tests of gender differences in variability were not performed. Future work should further investigate gender-related differences in variability using resampling approaches such as bootstrap methods.

The performance changes with varying R_f in Dreyer2023 may be partly explained by variability. For example, spatial variability showed a strong negative correlation with R_f , suggesting that lower spatial variability was associated with higher accuracy. In contrast, temporal variability exhibited a moderate positive correlation with R_f , implying that higher temporal variability may have had a relatively smaller impact on accuracy. These results suggest that the Riemannian TS+LR classifier used in this study may be sensitive to spatial variability while being relatively robust to temporal variability. However, a similar tendency was not observed in Lee2019, and further validation using a larger number of datasets will be necessary.

In this study, we used two datasets, Dreyer2023 and Lee2019; however, these data were collected at Inria Bordeaux (Talence, France) and Korea University (Seoul, Korea), respectively, and may therefore reflect populations with different geographical backgrounds. Such geographical population differences could be one of the factors leading to the differing results observed between the

datasets. Future work should thus include validation using a larger number of datasets to enable evaluation based on geographically diverse populations.

In addition, the present analysis was conducted using only TS+LR, and different trends may emerge when other classifiers are employed. Therefore, validation using alternative classifiers is an important direction for future research.

CONCLUSION

In this study, we investigated how varying the gender ratio in the training data affects BCI performance in cross-user MI-BCI. The results showed that classification accuracy was significantly higher in three out of four conditions when the machine-learning model was trained using only female users. In addition, across the same three conditions, positive correlations (up to 0.71) were observed between the proportion of female users in the training data and BCI performance, suggesting that training the classifier with more female user data may be associated with improved BCI performance, even for male users. Several previous studies have also suggested a female advantage in MI-BCI performance, and the present findings can be considered to be consistent with these observations. Moreover, since our results indicate that machine-learning models for MI-BCI may be subject to gender-related biases in the training data, they suggest that incorporating a gender-aware perspective will be important in future BCI research.

REFERENCES

- [1] Vidaurre C, Blankertz B. Towards a Cure for BCI Illiteracy. *Brain Topography*. 2010;23(2):194–198.
- [2] Randolph AB. Not All Created Equal: Individual-Technology Fit of Brain-Computer Interfaces. In: 2012 45th Hawaii International Conference on System Sciences. Jan. 2012, 572–578.
- [3] Cantillo-Negrete J, Gutierrez-Martinez J, Carino-Escobar RI, Carrillo-Mora P, Elias-Vinas D. An approach to improve the performance of subject-independent BCIs-based on motor imagery allocating subjects by gender. *BioMedical Engineering OnLine*. 2014;13(1):158.
- [4] Alimardani M, Gherman DE. Individual Differences in Motor Imagery BCIs: A Study of Gender, Mental States and Mu Suppression. In: International Winter Conference on Brain-Computer Interface. 2022, 1–7.
- [5] Cantillo-Negrete J, Gutiérrez-Martínez J, Cariño-Escobar RI, Flores-Rodríguez TB, Elías-Viñas D. Time-frequency analysis of EEG signals from healthy subjects allocated by gender for a subject-independent BCI-based on motor imagery. In: International IEEE/EMBS Conference on Neural Engineering. 2013, 37–40.
- [6] Wang P, Huang S, Jamil Z, Cheung VCK, Chan RH. Exploring Gender Differences in Motor Imagery EEG for Brain-Computer Interface Applications*. In: IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology. 2023, 15–16.
- [7] Klick SM et al. Females versus males exhibit greater brain activation in sensorimotor regions during motor imagery after stroke. *Neuroscience*. 2025;582:104–113.
- [8] Roc, Aline, Pillette, Léa, N’Kaoua, Bernard, Lotte, Fabien. Would Motor-Imagery based BCI user training benefit from more women experimenters? 2019.
- [9] Zorzet B, Echeveste R, Milone D, Peterson V. Can sex be decoded from MI features in deep learning based BCIs?: An exploratory analysis. In: Int. BCI Meeting. Verlag der Technischen Universität Graz, 2025.
- [10] Hagedorn LJ, Leeuwis N, Alimardani M. Prediction of Inefficient BCI Users Based on Cognitive Skills and Personality Traits. In: Neural Information Processing. 2021, 81–89.
- [11] Jeunet C, N’Kaoua B, Subramanian S, Hachet M, Lotte F. Predicting Mental Imagery-Based BCI Performance from Personality, Cognitive Profile and Neurophysiological Patterns. *PLOS ONE*. 2015;10(12):e0143962.
- [12] von Groll VG et al. Large scale investigation of the effect of gender on mu rhythm suppression in motor imagery brain-computer interfaces. *Brain-Computer Interfaces*. 2024;11(3):87–97.
- [13] Nadeem A, Abedin B, Marjanovic O. Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies. *ACIS 2020 Proceedings*. 2020.
- [14] Dreyer P, Roc A, Pillette L, Rimbert S, Lotte F. A large EEG database with users’ profile information for motor imagery brain-computer interface research. *Scientific Data*. 2023;10(1):580.
- [15] Lee MH et al. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience*. 2019;8(5):giz002.
- [16] Chevallier S et al. The largest EEG-based BCI reproducibility study for open science: The MOABB benchmark. Apr. 2024. arXiv: 2404.15319.
- [17] Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*. 2004;88(2):365–411.
- [18] Rodrigues PLC, Jutten C, Congedo M. Riemannian Procrustes Analysis: Transfer Learning for Brain-Computer Interfaces. *IEEE Transactions on Biomedical Engineering*. 2019;66(8):2390–2401.
- [19] Yger F, Berar M, Lotte F. Riemannian Approaches in Brain-Computer Interfaces: A Review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2017;25(10):1753–1762.
- [20] Slepian D. Prolate spheroidal wave functions, fourier analysis, and uncertainty — V: The discrete case. *The Bell System Technical Journal*. 1978;57(5):1371–1430.
- [21] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Acoust Speech Sig Proc*. 1978;26(1):43–49.

Non-invasive measurement of somatosensory evoked spinal cord potentials

Bernhard Golser¹, Markus Oberndorfer¹, Kyriaki Kostoglou¹, Gernot R. Müller-Putz^{1,2}

¹Institute of Neural Engineering, Graz Brain-Spine Lab, Graz University of Technology, Graz, Austria

²BioTechMed Graz, Graz, Austria

E-mail: gernot.mueller@tugraz.at

ABSTRACT: The spinal cord is more than a mere relay station between the brain and the peripheral nervous system. Recent research indicates that complex processing already occurs at the spinal level. The non-invasive recording of spinal cord activity acquired with high-density surface electrode arrays may advance our understanding of spinal network dynamics and the integration of spinal data could enhance existing brain-computer interfaces. In the present study, electroencephalography and electrospinography were used to measure cortical and spinal somatosensory evoked potentials elicited by median and tibial nerve stimulation in 10 healthy individuals. In contrast to previous studies, participants were seated upright with no external support of the head and stimulus intensities were kept strictly below the movement threshold. Under these more physiologically realistic settings, two of the three well-characterized spinal cord components (N13 and P22) were successfully detected, indicating a broader variety of potentially successful experimental setups.

INTRODUCTION

Being the most caudal part of the central nervous system (CNS), the spinal cord links the peripheral nervous system (PNS) to the brain. Several studies have shown that complex processing already occurs at the spinal level, contributing to both motor and somatosensory functions [1] [2]. Capturing this spinal activity non-invasively is possible using surface electrodes and is referred to as the electrospinogram (ESG). However, compared to the electroencephalogram (EEG) the ESG has received considerably less attention in scientific research. This may be, in part, due to the technical challenges associated with recording the ESG. Small signal amplitudes and large artifactual sources such as electric activity from the heart and surrounding muscles complicate its measurement. Existing research mainly focuses on the somatosensory evoked spinal cord potential (SCP), since its phase-locked nature facilitates signal enhancement by averaging across multiple trials [3] [4] [5] [6]. These potentials reflect activity along the somatosensory pathway. Afferent potentials traveling along this pathway pass through at least three neurons: primary afferents of the PNS are pseudounipolar neurons located in the dorsal

root ganglion. Their dendrites pick up signals generated by various sensory receptors, which are consequently relayed to the CNS via the neurons' axons. They enter the spinal cord through the dorsal roots and project to neurons in the CNS. These secondary afferent neurons are located in the dorsal horn, the medulla oblongata or the pons and in turn project to the thalamus. Subsequently, the signals are relayed to the somatosensory cortex by tertiary afferent neurons [7]. The so-called segmental SCP is evoked by stimulation of a peripheral nerve or a dorsal root and is recorded at the corresponding spinal segment [8]. It is composed of three main peaks characterized by their latency in ms. The first, positive peak (P9) represents afferent volleys traveling towards the dorsal root and entering the spinal cord [8]. Others have argued that P9 could reflect far field potentials stemming from afferent volleys in the brachial plexus [3]. The second, negative peak (N13) represents activity of secondary afferents and interneurons within the spinal segments targeted by stimulation [8]. The final well defined characteristic of the SCP is a slow, positive wave termed P22. Just as the two preceding components it is also detectable in the spinal segment corresponding to the stimulation site. It has been hypothesized that it reflects presynaptic inhibition by primary afferent depolarization [5] [8]. Although measurements conducted using epidural electrodes or needle electrodes yield a higher signal quality, the aforementioned SCP components have also been detected in surface electrode recordings, demonstrating the feasibility of non-invasive ESG recordings [6] [9] [10]. However, previous studies employed mixed nerve stimulation above the motor threshold, which yields higher SCP amplitudes, but is not representative of physiological somatosensory stimuli such as touch or pain. Additionally, previous studies minimized electromyogram (EMG) contamination by recording participants in supine or reclined positions that promote relaxation of neck muscles. While such conditions enhance the signal-to-noise ratio, they restrict the range of possible experimental approaches. In the present study, we aim to address these limitations by recording cervical SCPs following median nerve stimulation (MNS) at a stimulation intensity strictly below the motor threshold to avoid the recording of movement-related components. Recordings are obtained with participants seated in an upright position,

forcing them to engage the stabilizing muscles in their neck. Comparative measurements in response to tibial nerve stimulation (TNS) are conducted to validate the neurogenic nature of SCPs following MNS. We hypothesize that MNS will elicit the previously described waveform, whereas TNS will not produce cervical responses at the corresponding latencies, since the spinal segments associated with the tibial nerve are located further caudally. Concurrent cortical somatosensory evoked potentials (SEPs) are recorded to confirm successful stimulation. Importantly, we introduce a novel and reproducible electrode placement system for cervical ESG recordings based on easily identifiable anatomical landmarks, providing a standardized framework for future non-invasive spinal measurements.

MATERIALS AND METHODS

Participants: Ten able-bodied students participated in the study (three female). The mean age was 26.6 years (range: 25–30 years). Participants received financial compensation for their participation.

Paradigm: All recordings were conducted in a quiet, electromagnetically shielded room at a pleasant temperature. The participants were seated in a comfortable chair and were asked to adopt a relaxed position. During recordings they were asked to sit still and keep their gaze focused on a fixed point, but were specifically instructed to adjust their seating position any time they felt uncomfortable and to breath and blink normally. These instructions were given to prevent excessive EMG contamination due to a tense posture.

Two stimulation conditions were defined according to the major nerves targeted: MNS of the upper and TNS of the lower extremity. A total of 4000 stimulations were delivered per participant and condition, amounting to 8000 stimulations per session. The inter-stimulus-interval was sampled from a uniform distribution ranging from 700 to 900 ms. To minimize discomfort, the recordings were split into ten runs (five for each condition) of 800 stimulations, each run lasting approximately 11 minutes. MNS and TNS runs were executed in random order to avoid biases due to fatigue or habituation. Prior to the stimulation runs, 14 minutes of data for the resting state condition were recorded. Including breaks, the total recording time amounted to approximately three hours.

Stimulation setup: Electrical stimulation was delivered using the *Digitimer DS7R Constant Current Stimulator* (Digitimer, Letchworth Garden City, UK) and a pair of *Flextrode Plus 32mm* surface electrodes (Krauth+Timmermann, Hamburg, Germany). The device was set to monophasic stimulation with a pulse duration of 200 μ s. In all participants, stimulations were delivered to the right side of the body for both MNS and TNS with stimulation currents strictly below the motor threshold to avoid motor components in the recorded signal. To minimize artifacts caused by the stimulation current, a ground-clamp was attached proximal to the respective

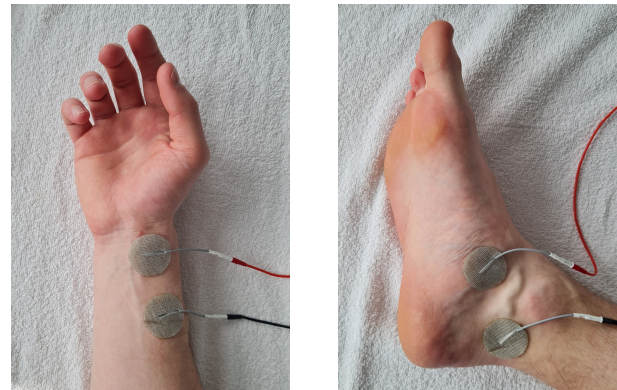


Figure 1: **Stimulation setup.** Approximate location of the stimulation electrodes for MNS (left) and TNS (right).

stimulation site. For MNS, the stimulation electrodes were placed on the radial side of the forearm, with the anode just proximal to the wrist. The cathode was placed proximal to the anode. The exact electrode positions were determined by trial and error: after an initial positioning, a stimulus above the motor threshold was delivered and the motor response was observed. The electrode positions were re-adjusted until the stimulus led to a twitch of the thumb, indicating a correct stimulation of the median nerve [11]. Once the electrode positions were established, the motor threshold was determined by gradually reducing the stimulation current until no motor response was visible. One participant did not tolerate stimulation currents high enough to trigger a motor response. In this case, the highest bearable stimulation current was applied and the electrode positions were determined based on the experience gained with previous participants. The same procedure was followed for TNS. The electrodes were placed on the medial side of the participant's ankle. The exact positions were once again determined by applying a stimulus above the motor threshold and observing the motor response. A plantar flexion of the first toe indicated the correct position [11]. Figure 1 shows the approximate electrode sites for MNS and TNS.

EEG recording: Both EEG and ESG were recorded using active electrodes and the *BrainAmp Standard* amplifier (Brain Products, Gilching, Germany). The amplifier's internal highpass filter and lowpass filter were set to 1 Hz and 1000 Hz, respectively. Signals were acquired with a sampling rate of 5000 Hz. The EEG was recorded using 28 electrodes. The electrode montage followed the standard 10-10 system, using the following electrode positions: FC5, FC3, FC1, FCZ, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPZ, CP2, CP4, CP6, P5, P3, P1, Pz, P2, P4, P6. Additionally, electrodes were placed on the left and right mastoids for later re-referencing. The ground electrode was placed on the forehead while the reference electrode was placed on the tip of the nose to provide a symmetric reference.

ESG electrode system: In the course of this study, a reproducible electrode system for high-density ESG recordings was developed. Derived from the well-established 10-10 system for EEG measurements, it uti-

lizes visible or palpable landmarks to define anatomically sensible positions for 54 electrodes. The vertical inter-electrode distance is defined as one fifth of the distance between the spinous process of the second cervical vertebra and the spinous process of the seventh cervical vertebra. Similarly, the horizontal inter-electrode distance is defined as one fifth of the distance between the left and right spinous process of the first cervical vertebra. Figure 2 displays the mentioned reference points and distances and the full ESG electrode layout. The ESG nomenclature was defined according to the following system: the first letter, *S* for *spine*, clearly distinguishes the ESG from the EEG nomenclature. It is followed by a number to determine the electrode's horizontal position. As in EEG nomenclature, uneven numbers are used for the left and even numbers for the right side, with higher numbers indicating a more lateral position. The number is followed by a *C* for *cervical* or a *T* for *thoracic*. The final number indicates the approximate vertical level of the electrode with respect to the closest spinous process. The active electrodes were attached to the participant's neck with adhesive rings. To ensure precise placement, the necessary distances were measured and markings were made on the participant's skin. The skin of the neck was thoroughly cleansed with alcohol prior to attaching the electrodes. Electrode gel was applied and allowed to settle in order to reduce impedance. As with the EEG electrodes, the ESG electrodes were also referenced to the nose-electrode. EEG and ESG recordings as well as triggers were synchronized via Lab Streaming Layer [12].

Processing: Signal processing was performed in *Matlab R2023b* [13]. Prior to any further processing, all recording sessions were visually inspected. Channels that were corrupted throughout an entire run were excluded from the respective session. Subsequently, short sections contaminated by large EMG artifacts (e.g., due to participants adjusting their sitting posture) were removed manually. The resulting dataset was re-referenced to the left and right mastoids' average potential. Power-line interference at 50 Hz and its harmonics were removed using a comb filter. Additionally, a 4th order butterworth highpass filter at 0.5 Hz was applied. To remove artifactual signal components such as the electrocardiogram (ECG) from the data, independent component analysis (ICA) was employed using the function *runica* implemented in *EEGLAB* [14]. To improve ICA decomposition while preserving high-frequency components, the ICA weights were computed on data band-pass filtered between 1 and 60 Hz. These weights were subsequently applied to the unfiltered dataset, and the resulting independent components (ICs) were visually inspected for rejection. Topographical maps were generated for each IC to assist classification. Subsequently, MNS and TNS data were epoched relative to the stimulus onset, whereas resting state data were segmented into non-overlapping epochs. Approximately 5-10% of epochs of each participant were rejected based on amplitude thresholding, their joint probability and their kurtosis [14]. All epochs

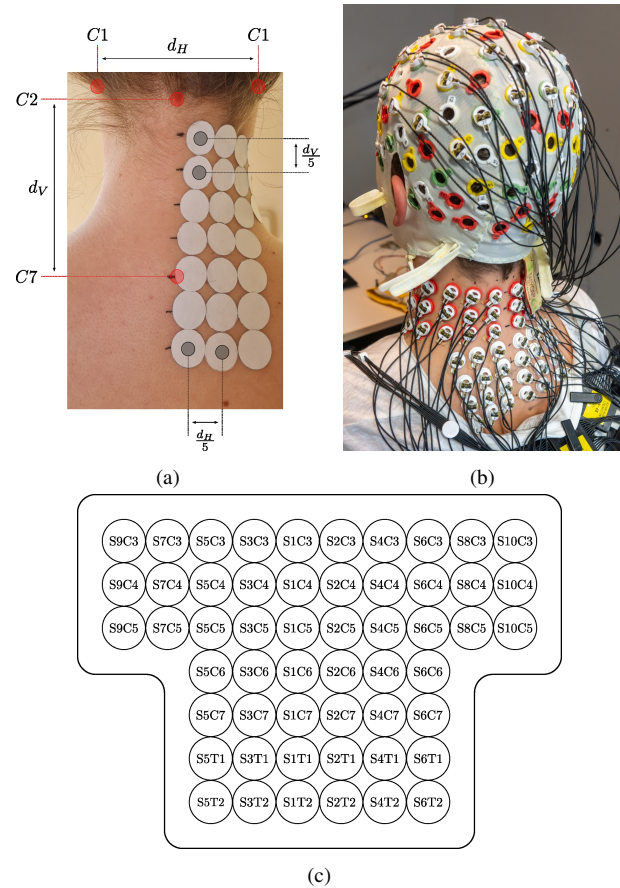


Figure 2: **ESG electrode system and measurement setup.**

a) Explanation of the ESG electrode system. C2 and C7 mark the locations of the respective spinous processes as determined by palpation. The labels C1 indicate the transverse processes of the first cervical vertebra. d_V is the vertical distance between C2 and C7 while d_H is defined as the horizontal distance between the transverse processes of C1. The vertical and horizontal inter-electrode distances are defined by $\frac{d_V}{5}$ and $\frac{d_H}{5}$, respectively. **b)** Full EEG and ESG electrode setup. **c)** ESG electrode layout and nomenclature. The letter S stands for spine, distinguishing ESG from EEG electrodes. It is followed by a number to determine the electrode's horizontal position. Uneven numbers are used for the left and even numbers for the right side, with higher numbers indicating a more lateral position. This is followed by a combination of the letter C or T (for *cervical* or *thoracic*) and a number, indicating the approximate vertical level of the electrode with respect to the closest spinous process.

were baseline corrected with respect to the interval -55 to -5 ms. Subsequently, grand averages were calculated. For visualization purposes, 95% confidence intervals were calculated for each condition and displayed as shaded areas around the average time courses. Local extrema around the expected latencies of N13 ($t = 12.4$ ms) and P22 ($t = 22.2$ ms) could be identified. The MNS and TNS conditions were tested against the rest condition at these latencies using linear mixed-effect models with the condition as a fixed effect and the participant as a random intercept to account for repeated measurements within participants. Benjamini-Hochberg correction was applied to account for multiple comparisons [15]. ESG

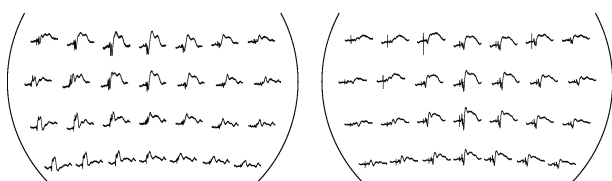


Figure 3: **Cortical somatosensory evoked potentials.** SEPs recorded in response to MNS (left) and TNS (right) displayed at their respective locations on the scalp. The traces are plotted in the time range -120 to 430 ms with respect to the stimulation onset for both conditions.

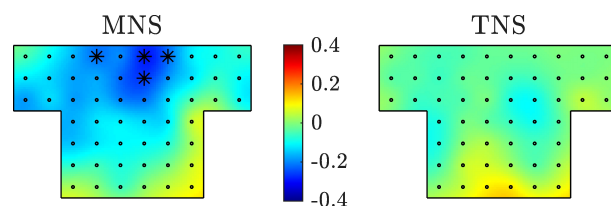
topoplots at the tested latencies were created by spline interpolation between the individual electrode sites.

RESULTS

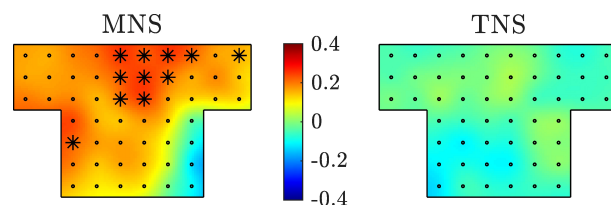
Cortical somatosensory evoked potentials: EEG measurements were performed, in part, to validate the effective stimulation of the median and tibial nerve. Figure 3 illustrates the SEPs recorded across the scalp. The contralateral nature of the evoked potentials following MNS was clearly visible, while the response to TNS was more central with a slight tendency towards the ipsilateral (right) side in the later components.

Somatosensory evoked spinal cord potentials: At the group level, two of the three major peaks described in literature could be identified in our SCP recordings. Figure 4 displays evoked waveforms and their corresponding topographical distributions following MNS and TNS. At a latency corresponding to N13, four electrodes, most of them located rather rostral and medial, showed significant amplitudes after MNS compared to rest (see Subfigure 4a). No clear lateralisation was visible and the wave had a duration of approximately 5 ms. The comparison to TNS served as a control condition. No significant amplitudes were visible at this latency in response to TNS, which had been expected. The peak at a latency associated with P22 covered a much larger area of the neck, with its maximum intensity in the rostral area on the ipsilateral (right) side (Subfigure 4b). A region of reduced amplitudes was distinguishable in the ipsilateral, caudal part of the electrode grid. With a duration of approximately 20 ms, this wave was much broader than the previously described N13. Again, TNS did not lead to any significant activity at the latency of interest.

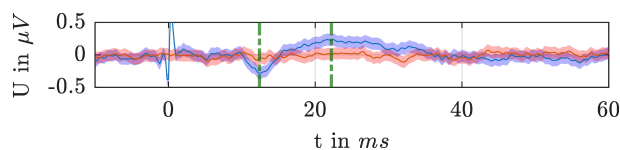
Somatosensory pathway: Figure 5 depicts the potentials evoked by MNS recorded along the somatosensory pathway in this study, all of which have been previously described in literature [8] [16]. The figure clearly shows that the earliest component registered in this study was a negative peak that appeared over the cervical spinal cord (N13), which is attributed to postsynaptic activity in the dorsal horn [8]. N13 was followed by far field potentials measured on the scalp representing activity in the area of the cuneate nuclei and the medial lemniscus (P14) [17]. Subsequently, a more localized potential registered above the contralateral somatosensory areas indicated the afferent signal's arrival at the cortex (N20). Finally, P22 represented subsequent activity at the spinal level, ascribed to primary afferent depolarisation and top-down feedback from supraspinal structures [8] [18].



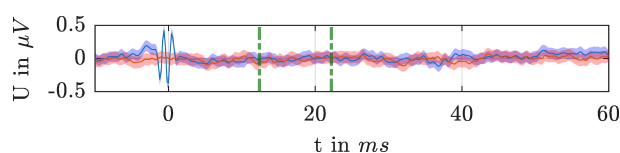
(a) Topographic maps at a latency of 12.4 ms after MNS and TNS



(b) Topographic maps at a latency of 22.2 ms after MNS and TNS



(c) Grand average time course of S2C3 after MNS



(d) Grand average time course of S2C3 after TNS

Figure 4: **SCPs after MNS and TNS.** The topographic maps on the left in **a)** and **b)** display the signal distribution after MNS at the maximum of N13 and P22, respectively. Electrodes marked with an asterisk showed significant activity compared to rest at the corresponding latency. Note the central and rostral location of N13 and the broader area affected by P22. The topographic plots on the right in **a)** and **b)** display the signal distribution at the same latencies after TNS. Note that no significant activity compared to rest was captured at these latencies as a result of TNS. **c)** Grand average SCP elicited by MNS in one representative electrode. **d)** Signal captured in the same electrode in response to TNS. The blue traces represent the signal recorded after MNS (**c)** and TNS (**d**), the red trace represents the rest condition. The dashed vertical lines mark the latencies of N13 and P22. Shaded areas indicate the 95% confidence interval.

resented subsequent activity at the spinal level, ascribed to primary afferent depolarisation and top-down feedback from supraspinal structures [8] [18].

DISCUSSION

Cortical somatosensory evoked potentials: In general, the grand average SEPs recorded after both MNS and TNS were consistent with existing literature. As expected, the topographic plots showed a lateralization towards the contralateral side with respect to the stimulated side in case of MNS, which reflects the location of the respective areas within the somatosensory cortex [16]. The evoked potentials following TNS were localized above the center of the somatosensory cortex and

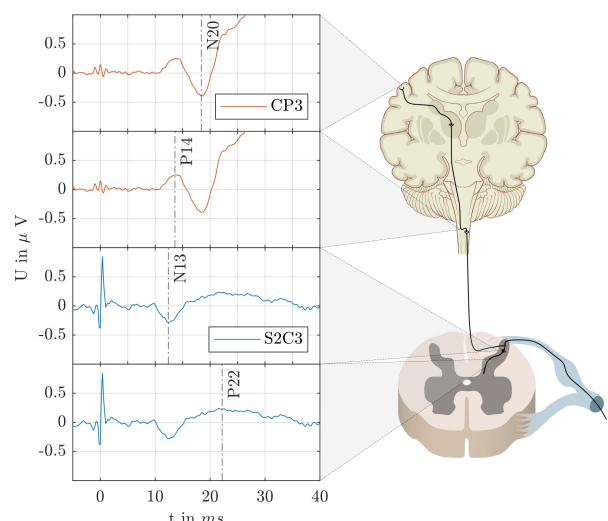


Figure 5: Responses to MNS along the somatosensory pathway recorded in this study. The peaks marked in the traces of SEP and SCP on the left hand side of the figure are allocated to their respective regions of origin in the sketch on the right. The blue traces in the two bottommost plots are SCPs recorded in the ESG electrode S2C3, showing the peaks N13 and P22. The upper two orange traces are SEPs recorded in CP3, reflecting P14 and N20. Sketches adapted from [19] [20].

showed a deviation to the ipsilateral side. This deviation is unexpected and could not be explained. One reason for recording the EEG response was to validate MNS and TNS. Considering the grand average traces and the fact, that SEPs were evoked in all participants at the individual level, it can be concluded that both median and tibial nerve stimulation were successful.

Somatosensory evoked spinal cord potentials: At the group level, the evoked SCP was clearly visible. Although the precise anatomical origins of the individual components cannot be determined from surface electrode recordings, the observed waveforms were consistent with descriptions reported in literature [6] [8] [18] [21] [22] [23]. As previously described, N13 and P22 were both observable in our data. Interestingly, both potentials were recorded at quite rostral levels, with N13 being maximal at the level of the spinous processes C3 and C4. Considering that the median nerve projects to the spinal segments Th1 to C6 this is unexpected. Similarly rostral locations of N13 have been reported previously [4] [22]. Part of this upward shift might be explained by the spine's anatomy: the processes used to estimate the location of the corresponding spinal segments protrude in caudal direction. Additionally, the respective spinal segment is situated above the corresponding vertebra. As a result, the electrode row ascribed to a certain spinal segment was probably placed too far caudally with respect to this segment. This at least partly explains why N13 was recorded so far rostrally. However, since shape and timing of the recorded peaks are in line with literature, it can be concluded that we successfully recorded N13 (re-

flecting postsynaptic activity in the dorsal horn) and P22 (reflecting activity resulting from top-down feedback or primary afferent depolarisation). However, P9 was not apparent in our data. The observation that N13 and P22 were detectable whereas P9 was not may be attributed to the postsynaptic origin of N13 and P22. P9 reflects afferent volleys and the strength of these volleys is related to the amount of nerve fibres recruited by stimulation. Small stimuli result in a low number of recruited fibres [24], which in turn results in small signal amplitudes when recording summed potentials of said volleys using surface electrodes. However, N13 is caused by the activity of a network of neurons that receives input from these primary afferents. Coombs et al. [25] proposed that repetitive firing within this network might cause the prolonged duration of the N13 peak. It is plausible that such sustained postsynaptic activity facilitates the detectability of N13 in our setup. A similar argument may explain the detectability of P22. Even though the afferent volleys elicited by sub-threshold stimulation may have been too weak to produce a measurable P9 component, they were sufficient to evoke robust cortical responses. The resulting top-down feedback hypothesized to underlie P22 likely generated sufficient activity within the corresponding spinal segment to permit non-invasive recording.

Taken together, our findings demonstrate that cervical somatosensory evoked SCPs can be recorded non-invasively using a high-density surface electrode array. To our knowledge, this is the first report of cervical SCPs obtained with sub-motor-threshold MNS in participants seated upright without head support. Despite these less optimized conditions, the components N13 and P22 were reliably detectable at the group level, whereas the component P9 was not observed. Achieving statistically significant effects under these conditions required a substantially larger number of stimuli compared to studies employing higher stimulation intensities or supine recordings. Nevertheless, these results extend previous work and indicate that non-invasive ESG recordings remain feasible in more naturalistic experimental settings. Recent advances in spinal electrophysiology, including single-trial SCP detection using spatial filtering techniques [23], reports of attentional modulation of P22 [18] and the proposal of an alternative ESG electrode layout [22], highlight a rekindled interest in non-invasive assessment of spinal processing. In this context, the present dataset provides a valuable basis for future research, such as the analysis of spino-cortical connectivity.

Limitations: Despite these promising findings, several limitations should be considered. The proposed electrode layout is reproducible and anatomically sensible. However, the anatomical relationship between spinal segments and their corresponding vertebra does not hold for more caudal regions of the spinal column. Hence, adjustments would be necessary for thoracic and lumbar measurements. Regarding signal processing, the manual removal of ECG via ICA was successful but time-consuming. Other promising methods for ECG removal

have been proposed [22] [23].

CONCLUSION

Our findings demonstrate the feasibility of non-invasively recording somatosensory evoked spinal cord potentials under more naturalistic conditions, suggesting a broader range of viable paradigms. In light of the renewed interest in non-invasive spinal recordings across several groups, such approaches may play an important role in advancing our understanding of spinal-level processing. Furthermore, incorporating spinal signals into existing BCI approaches may provide a missing link with the potential to enhance system performance in the future.

REFERENCES

- [1] Chirila AM et al. Mechanoreceptor signal convergence and transformation in the dorsal horn flexibly shape a diversity of outputs to the brain. *Cell*. 2022;185(24):4541–4559.e23.
- [2] Lorach H et al. Walking naturally after spinal cord injury using a brain-spine interface. *Nature*. 2023;618(7963):126–133.
- [3] Desmedt JE, Nguyen TH. Bit-mapped colour imaging of the potential fields of propagated and segmental subcortical components of somatosensory evoked potentials in man. *Electroencephalography and clinical neurophysiology*. 1984;58(6):481–497.
- [4] Hallström YT, Lindblom U, Meyerson BA, Prevec TS. Epidurally recorded cervical spinal activity evoked by electrical and mechanical stimulation in pain patients. *Electroencephalography and clinical neurophysiology*. 1989;74(3):175–185.
- [5] Jeanmonod D, Sindou M, Mauguière F. The human cervical and lumbo-sacral evoked electrospinogram. data from intra-operative spinal cord surface recordings. *Electroencephalography and clinical neurophysiology*. 1991;80(6):477–489.
- [6] Shimoji K, Shimizu H, Maruyama Y. Origin of somatosensory evoked responses recorded from the cervical skin surface. *Journal of neurosurgery*. 1978;48(6):980–984.
- [7] Waldeyer A, Anderhuber F, Pera F, Streicher J. Waldeyer - anatomie des menschen. DE GRUYTER (2012).
- [8] Kano T, Katayama Y, Fukuda S, Shimoji K, Willis WD. Evoked spinal cord potentials. Springer Japan: Tokyo (2006).
- [9] Di Lionardo A et al. Modulation of the n13 component of the somatosensory evoked potentials in an experimental model of central sensitization in humans. *Scientific reports*. 2021;11(1):20838.
- [10] Di Pietro G et al. The n13 spinal component of somatosensory evoked potentials is modulated by heterotopic noxious conditioning stimulation suggesting an involvement of spinal wide dynamic range neurons. *Neurophysiologie clinique = Clinical neurophysiology*. 2021;51(6):517–523.
- [11] Mauguière F et al. Somatosensory evoked potentials. the international federation of clinical neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*. 1999;52:79–90.
- [12] Kothe C et al. The lab streaming layer for synchronized multimodal recording. *Imaging neuroscience (Cambridge, Mass.)*. 2025;3.
- [13] Inc. TM. *Matlab version: 9.13.0 (r2023b)*. Natick, Massachusetts, United States, 2023. [Online]. Available: <https://www.mathworks.com>
- [14] Delorme A, Makeig S. Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*. 2004;134(1):9–21.
- [15] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1995;57(1):289–300.
- [16] Stöhr M. *Evozierte potenziale: Sep - vep - aep - ekp - mep*. Springer Medizin Verlag Heidelberg: Berlin, Heidelberg (2005).
- [17] Eisen A. The somatosensory evoked potential. *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques*. 1982;9(2):65–77.
- [18] Gabrieli G, Somervail R, Mouraux A, Leandri M, Haggard P, Iannetti GD. Electrical spinal imaging: A noninvasive, high-resolution approach that enables electrophysiological mapping of the human spinal cord. *PLoS biology*. 2025;23(11):e3003116.
- [19] Aguayo M. Imgnotraçat arc refelx. [Online; accessed April 8, 2025]. 2014. [Online]. Available: https://commons.wikimedia.org/wiki/File:Imgnotra%C3%A7at_arc_refelx_esp.svg
- [20] Lynch PJ, Jaffe CC. Human brain coronal section. [Online; accessed April 8, 2025]. 2006. [Online]. Available: https://commons.wikimedia.org/wiki/File:Brain_human_coronal_section.svg
- [21] Cracco RQ. Spinal evoked response: Peripheral nerve stimulation in man. *Electroencephalography and clinical neurophysiology*. 1973;35(4):379–386.
- [22] Mehra P et al. Sc10x/u: A high-density electrode system for standardised surface recording of neural activity of the cervical spinal cord. *Clinical Neurophysiology*. 2025;177:2110824.
- [23] Nierula B et al. A multichannel electrophysiological approach to noninvasively and precisely record human spinal cord activity. *PLOS Biology*. 2024;22(10):1–41.
- [24] Müller-Putz G, Rupp R. *Neuroprosthetics and brain-computer interfaces in spinal cord injury*. Springer International Publishing: Cham (2021).
- [25] Coombs JS, Curtis DR, Landgren S. Spinal cord potentials generated by impulses in muscle and cutaneous afferent fibres. 1956;19(5):452–467.

Towards thermal imagery-based brain-computer interfaces

T. Lefeuve¹, E. Savalle¹, J. Petit¹, M. J-M Macé¹, A. Lécuyer², L. Pillette¹

¹Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

²Inria, Univ. Rennes, IRISA, CNRS, Rennes, France

E-mail: theo.lefeuvre@inria.fr

ABSTRACT: Mental task-based brain-computer interfaces (BCIs) enable users to control a system by voluntarily modulating their brain activity without overt movement. Research in the field has focused mainly on motor imagery, whereas tactile imagery, defined as the imagination of tactile stimuli on the skin, remains comparatively underexplored. Although pressure and vibratory imagery have received growing attention, other modalities such as thermal imagery remain largely neglected, despite their potential clinical relevance. In this EEG-based BCI study, healthy participants learned to control a BCI by imagining a warm sensation in the right hand, representing, to our knowledge, the first thermal imagery-based BCI user training. To facilitate mental representation, participants received visual, thermal, or combined visuothermal instructions. Thermal imagery elicited event-related desynchronization, primarily over left sensorimotor regions, and online performance improved significantly across runs. Comparisons between instruction modalities suggested trends favoring visual and thermal guidance. Together, these findings support the feasibility of thermal imagery as a promising alternative control paradigm for BCI applications.

INTRODUCTION

Brain-computer interfaces (BCI) are technologies that enable the control of external devices by translating central nervous system activity into control commands [1]. BCIs have been investigated for both clinical and non-clinical applications [2]. One of the most widely used paradigms to control a BCI is kinesthetic motor imagery (KMI), which involves imagining moving a limb without performing the action [3] combined with imagining sensations associated with its execution (e.g., muscle tension and proprioceptive sensations). However, current BCI paradigms remain insufficiently reliable for widespread deployment beyond research settings. One of the major challenges remains the phenomenon of BCI inefficiency, where BCI control based on classical paradigms such as KMI fails in approximately 15-30% of users. Previous work suggests that this problem may, at least in part, stem from limitations in the design of current BCI systems [4, 5].

Alternative BCI paradigms have been proposed to address the difficulties participants may encounter when

training with conventional approaches. Unlike KMI, which relies on imagined movement, tactile imagery (TI) relies on the imagination of tactile stimulation on the skin, most commonly pressure and vibratory stimulation, and several studies have already demonstrated its feasibility in BCI paradigms [6–9]. KMI and TI both rely on oscillatory brain dynamics, commonly quantified via event-related desynchronization/synchronization (ERD/ERS) in sensorimotor rhythms [6–8], reflecting decreases (ERD) or increases (ERS) in oscillatory power relative to a baseline period. Both KMI and TI elicit ERD in the alpha (8-13Hz) and beta (13-30Hz) bands, with TI showing particularly pronounced alpha-range modulations [7, 8] and statistical comparisons between the two revealing significant differences mainly in that frequency band. TI has been shown to exhibit neural activation patterns closely resembling those observed during real tactile stimulation [8, 10], whereas KMI recruits largely overlapping neural networks with those involved in voluntary movement.

Within the context of TI, the existing literature mostly focuses on the imagery of pressure or vibratory tactile stimulation. Other modalities, such as cold or heat imagination, to the best of our knowledge, have received no attention despite their potential relevance, notably for clinical applications. Indeed, following a stroke, approximately 50% of patients exhibit impairment affecting not only motor but also sensory functions such as pressure, touch, pain, and temperature perception. It is not yet clear which somatosensory abilities are the most frequently impacted [11, 12]. Given that TI has been shown to recruit sensory-related regions [6], it may represent a relevant approach for rehabilitation protocols, particularly when combined with congruent sensory feedback to facilitate neuroplasticity mechanisms [13]. In this context, thermal imagery (ThI) appears to be a potentially relevant approach for the rehabilitation of sensory deficits. Moreover, similarly to pressure or vibratory TI, which shares neural activation patterns with actual tactile stimulation [8], it can be hypothesized that ThI activates neural networks comparable to those recruited during real thermal stimulation. This is of particular interest given that thermal stimulation has been shown to facilitate recovery in post-stroke patients [14]. Furthermore, exploring alternative forms of mental imagery may enable individuals for whom classical paradigms, such as KMI, are ineffective

to identify better-suited strategies. Such diversification of paradigms could contribute to mitigating the issue of BCI inefficiency by broadening the range of viable control strategies across users. However, BCI user training, during which BCI users learn to control their brain activity based on ThI, has not yet been explored, despite also being an imagery-based approach relying on the modulation of neural rhythms.

To facilitate the reconstruction of sensory representations, some studies have introduced a brief stimulation phase to facilitate participants' tasks, particularly in pressure and vibratory tactile imagery [9, 15]. These stimuli, delivered either at the beginning of each trial [15] or before the imagery condition [9], had the purpose of helping participants better reproduce the target sensations during the imagery task. However, the effect of different instruction modalities on brain activity or user experience has not been reported. We hypothesise that for ThI, the modality of instruction will significantly impact the resulting change in brain activity and participants' performance.

In this context, the present experiment aimed to evaluate a novel BCI paradigm based on ThI by addressing two main questions: (i) Can ThI be used to control a BCI, and can its efficiency be enhanced through user training? (ii) Does the type of instruction significantly influence BCI performance?

MATERIALS AND METHODS

To examine the effect of the instruction modality presented at the beginning of each imagery trial, participants completed three different instruction conditions: visual, thermal and visuothermal, presented in a counterbalanced order across participants. To assess training effects within each condition, participants first performed one ThI run without feedback, followed by two runs with feedback. For each condition and at the end of the session, questionnaires were used to collect participants' subjective evaluations.

Participants:

Twenty healthy right-handed participants (13 women, 7 men, age 28.2 ± 5.7 years) took part in the experiment. All participants were fluent in either French or English, had normal or corrected-to-normal vision and hearing, and reported no history of neurological or psychiatric disorders. Participants were excluded if they reported traumatic brain injury, epilepsy, or chronic cardiovascular/metabolic diseases. All participants provided written informed consent before the experiment. The study protocol was reviewed and approved by a national ethical committee (Coerle n°2025-62). Each participant received a compensation of 20 euros for their participation.

Experimental protocol:

The experiment lasted approximately two hours, including setup, calibration, and cleaning procedures. The actual recording time was about one hour. The experimental plan is displayed in Figure 1. Written instructions

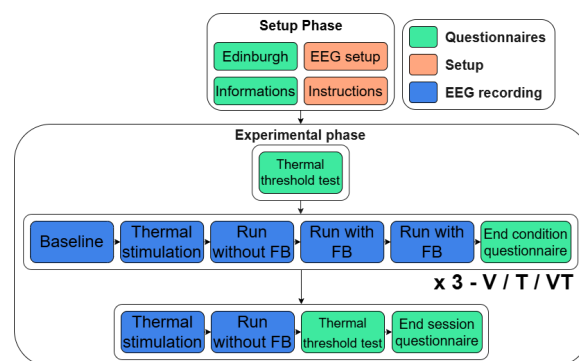


Figure 1: Experimental protocol. The three conditions i.e., visual (V), thermal (T), visuothermal (VT) were presented in a counterbalanced order across participants.

were provided to each participant to ensure consistency and reproducibility of the experiment. Upon arrival, participants completed demographic and Edinburgh Handedness Inventory handedness questionnaires (see section Questionnaires). To immerse participants in a realistic scenario and provide embodied feedback, we used a virtual environment displayed in the VR headset. The scene was a neutral room with a table aligned with the real table in front of which the participants sat. On the table was an animated bonfire used for instruction and feedback. The VR headset was rigidly mounted to a custom headrest system. This setup both reduced head movement-related artifacts and limited small electrode shifts that can occur when the headset is removed between conditions. Height and depth of the VR headset were adjusted for comfort and a stable fit.

At the beginning of the experiment, a thermal threshold test based on the method of levels (LV) paradigm from [16] was conducted to determine the minimal temperature participants were able to detect. For each condition, participants received a series of twenty non-painful thermal stimuli. They were instructed to remain still and focus on the perceived sensations. Then, participants completed one run of ThI without feedback, followed by two runs with feedback. Each run consisted of 20 trials and followed the same structure, 2 s of rest, during which participants were instructed to relax and remain still, 2 s of the condition specific instruction, 5 s of ThI performed with or without feedback including 1 s of cue indicating the onset of the imagery period and a final rest period randomly varying between 1 and 2 s. The first run without feedback of each condition served as a calibration phase. ERDS features extracted during the imagery periods of this run were used to compute the feedback delivered in the following runs with feedback. After completing the two runs with feedback, participants filled in a questionnaire. Then after completing the three conditions, participants performed an additional thermal stimulation run and an additional run of ThI without feedback, followed by a final thermal threshold test and an end-of-session questionnaire.

Different instruction modalities were used across both runs, with and without feedback, depending on the condi-

tion. The order of the conditions was pseudo-randomized and counterbalanced across participants. In each condition, participants were asked to imagine a sensation of warmth according to the instructions provided as follows:

- Visual (V) – visual animation in VR of a campfire burning displayed between two virtual arms. Participants were asked to imagine a feeling of heat on the back of their hand as if the campfire were real.
- Thermal (T) – thermal stimulation delivered on the back of the hand (see Section Thermal stimulation). Participants were asked to imagine the same sensation of heat that they felt during the stimulus.
- Visuothermal (VT) – a combination of visual and thermal instructions. Participants were asked to imagine the same sensation of heat on the back of their hand that they felt during the stimulus, and as if the campfire were real.

Thermal stimulation:

To deliver non-painful thermal stimulation on the back of the participants' hands, a device was designed based on the system developed in [17], with a HS-85MG servo motor. The device integrates a Force-Sensing Resistor to ensure a controlled and consistent pressure across all participants. Temperature regulation is achieved using a thermistor. The temperature applied to each participant was individualized and based on their thermal threshold test results (see Section Experimental protocol). Specifically, the absolute value of the threshold measured at the beginning of the experiment was used, and 10 °C were added to determine the stimulation temperature to ensure that the thermal stimuli were detectable but not painful. Participants were systematically asked to confirm whether the stimulus was not painful, both verbally and through post-task questionnaires.

EEG Recordings:

The EEG data was recorded with 30 active electrodes, using a g.USBamp EEG amplifier (g.tec, Austria). The electrodes were placed over the sensorimotor cortex (at locations Fp1, Fp2, F3, F4, FC5, FC6, FC3, FC4, FC1, FC2, FCz, T7, T8, C5, C6, C3, C4, C1, C2, Cz, CP5, CP6, CP3, CP4, CP1, CP2, CPz, P3, P4 and Oz in the 10-20 system), referenced to the left earlobe and grounded to AFz. Data was sampled at 512 Hz and processed using OpenViBE 3.6.0 [18]. Two participants were excluded from further analysis due to excessive noise in the recorded EEG data.

Online Feedback:

During runs with and without feedback, the online data used to compute ERDS was processed as follows. We first selected the signal from electrodes CP3, C1, C5, and P3, which was then filtered between 8 and 30 Hz. We then applied a Laplacian filter centered on CP3, which was chosen due to its location over the somatosensory cortex. Based on previous literature indicating that TI elicit activation in the somatosensory cortex [6–9], this approach was chosen to enhance spatial resolution [19].

During the resting phases of each trial (2 s), the output signal was epoched using a 1 s window, sliding every 0.1 s. For each epoch, the signal was squared and then time-averaged to get the band power. Outlier epochs outside 3 standard deviations were removed, and the remaining epochs were averaged and used as the Rest value to compute the ERDS of the following ThI task. During the ThI periods of each trial, the output signal from the Laplacian filter was epoched using a 250 ms window every 250 ms. The signal of each epoch was squared and then time-averaged to estimate the band power. For every epoch, this average was used as the Task value to compute the online ERDS as follows:

$$ERDS_{online} = \frac{Task - Rest}{Rest} * 100 \quad (1)$$

During the run without feedback of each condition, a calibration score $ERDS_{calib}$ was computed. At the end of the run, all the $ERDS_{online}$ values calculated during the run were sorted in ascending order, and the value at the 30th percentile was selected as $ERDS_{calib}$. During the runs with feedback, the $ERDS_{calib}$ obtained from the run without feedback was used to compute an ERDS score (S_{ERDS}) based on the $ERDS_{online}$ measured during the task as follows:

$$S_{ERDS} = \frac{ERDS_{online}}{ERDS_{calib}} * 100 \quad (2)$$

The S_{ERDS} was then discretized to define the participant's performance score and determine the intensity of the feedback provided to the participant (see Figure 2), with level 0 displaying no intensity and 3 displaying maximum intensity.

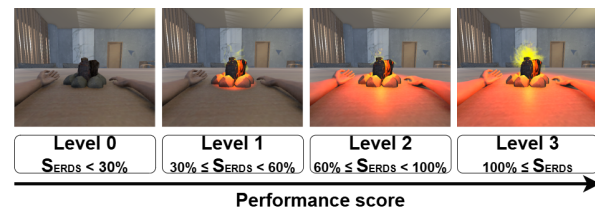


Figure 2: Visual feedback provided during the runs based on the user performance S_{ERDS}

Questionnaires:

First, we assessed participant characteristics that could potentially influence the results using two questionnaires. These include the evaluation of lateralization through the Edinburgh questionnaire [20], as well as the collection of demographic variables. We also examined the user experience associated with ThI for BCI control at the end of each condition and the session.

Analysis:

First, we investigated whether participants' performance improved over time across the two runs with feedback of each condition, and whether differences were observed between the condition-specific instructions. We used a two-way repeated-measures ANOVA on the averaged

S_{ERDS} measured, with *Condition* (levels: Visual, Thermal, Visuo-thermal) and *Run* (levels: Run 1 and Run 2) as independent variables. Post-hoc pairwise comparisons were performed using Holm correction to account for multiple comparisons. Outliers can strongly influence statistical tests and may artificially affect the significance of the results. To ensure more reliable statistical analysis, participants presenting outlier values according to the predefined criteria were excluded from the analysis. For each run with feedback of each condition, participants' averaged performance was used to identify outliers, defined as the values outside two standard deviations from the mean of the run. Participants who exhibited ERDS outliers in three or more of the six total runs were excluded from statistical analysis. In total, one participant was excluded from this analysis. Then, to investigate whether participants' performance improved over the course of the experiment during runs without feedback, we conducted a one-way (1×4) repeated-measures ANOVA on the averaged S_{ERDS} measured during the four runs without feedback, with *Run* (levels: Run 1, Run 2, Run 3, and Run 4) as a within-subject factor. Post-hoc pairwise comparisons were performed using Holm correction to account for multiple comparisons. To examine the systematic evolution of the dependent variable across the four runs, polynomial contrasts were applied to the *Run* factor. This approach allowed us to specifically test for linear, quadratic, and cubic trends in the repeated measures. Inspection of the coefficient allowed us to verify if this trend was negative, as expected if participants improved their performance over the course of the experiment. For each run without feedback, participants' averaged performance was used to identify outliers, defined as the values outside two standard deviations from the mean of the run. Participants who exhibited ERDS outliers in two or more of the four runs were excluded from statistical analysis. In total, three participants were excluded from this analysis.

RESULTS

Two-way repeated-measures ANOVA on participants' performance during the two runs with feedback showed a significant effect of *Condition* (see Figure 3). Since Mauchly's test of sphericity was significant for *Condition*, the Greenhouse-Geisser correction was used ($F(1.296, 20.742) = 4.124$, $p_{corrected} = 0.046$, $\eta^2 = 0.106$). Among the 17 participants, one participant showed ERS on average in both runs of the visuo-thermal condition as well as in one run of the visual and thermal conditions. Two other participants showed ERS in one run of the visuo-thermal condition. The remaining participants showed ERD on average during all the runs. In addition, values are in average more negative in the visual ($M = -27.087$, $SD = 16.665$) and thermal ($M = -24.718$, $SD = 19.355$) conditions compared to visuo-thermal ($M = -19.571$, $SD = 21.308$). Post-hoc comparisons (Holm-corrected) revealed a tendency be-

tween thermal and visuo-thermal conditions ($p = 0.070$, $d = -0.247$) and between visual and visuo-thermal conditions ($p = 0.098$, $d = -0.360$). A marginal effect was found for *Run* ($F(1, 16) = 3.234$, $p = 0.091$, $\eta^2 = 0.025$), with Run 2 ($M = -25.329$, $SD = 18.805$) showing more negative values than Run 1 ($M = -22.256$, $SD = 18.294$). However, no significant effect was found for the interaction *Condition* $\times$ *Run* ($F(2, 32) = 1.117$, $p = 0.336$, $\eta^2 = 0.022$). Participants were divided based on ERDS magnitude of the two runs averaged. Strong responders (ERDS $< -20\%$) represented 44% of the participants ($n = 8$), moderate responders ($-20\% < \text{ERDS} < -10\%$) accounted for 33% ($n = 6$) and low responders ($-10\% < \text{ERDS} < 0\%$) accounted for 22% ($n = 4$).

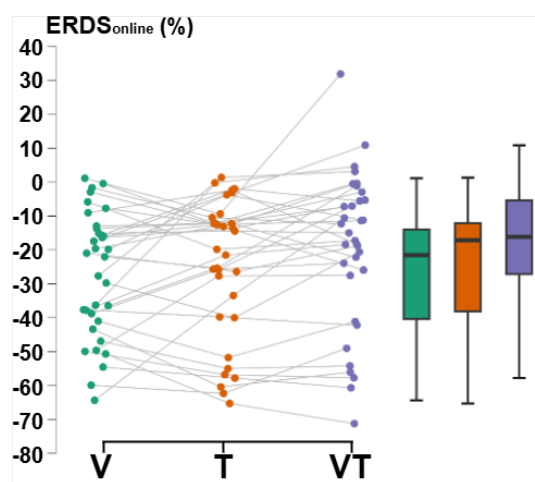


Figure 3: Averaged $ERDS_{online}$ during runs with feedback averaged over all participants for visual (V), thermal (T) and visuo-thermal (VT) conditions.

The analysis on the averaged $ERDS_{online}$ measured during the runs without feedback (see Figure 4) indicated that *Run* had a significant effect ($F(3, 42) = 3.315$, $p = 0.034$, $\eta^2 = 0.211$). ERD became progressively more negative from Run 1 ($M = -6.518$, $SD = 16.701$) to Run 3 ($M = -16.470$, $SD = 16.214$), then partially rebounded in Run 4 ($M = -13.693$, $SD = 15.342$). From the first run, desynchronization is apparent, with most values presenting negative synchronization changes. Post-hoc comparisons (Holm-corrected) revealed no further significant effect between runs.

In order to examine whether the effect of *Run* followed a specific trend, we conducted a polynomial contrast after the ANOVA. The results revealed a moderate linear trend ($t(17) = -2.169$, $d = -0.334$, $p = 0.051$), indicating that ERD decreased linearly across runs ($Estimate = -5.505$, $SE = 2.538$). The quadratic trend showed a similar trend ($t(17) = 1.834$, $p = 0.092$, $d = 0.293$), suggesting weak quadratic shape ($Estimate = 4.818$, $SE = 2.627$). The cubic trend was not significant ($t(17) = 0.212$, $p = 0.835$, $d = 0.020$), indicating no higher-order trend. Participants were grouped according to the ERDS magnitude observed during the run. Strong responders (ERDS $< -20\%$) represented 33% of the participants ($n = 6$), mod-

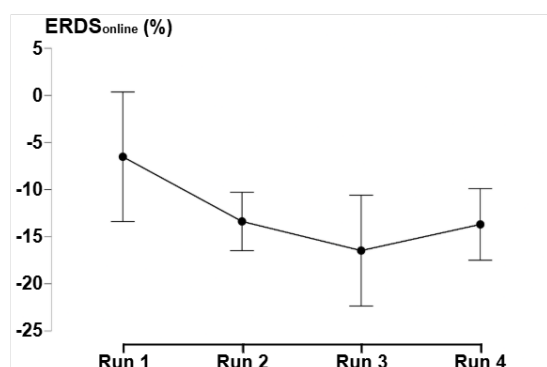


Figure 4: Evolution of the average $ERDS_{online}$ measured during runs without feedback, averaged over all participants, with a 95% confidence interval.

erate responders ($-20\% < ERDS < -10\%$) accounted for 22% ($n = 4$), and low responders ($-10\% < ERDS < 0\%$) accounted for 22% ($n = 4$). The remaining 22% ($n = 4$) were classified as non-responders ($ERDS > 0\%$), exhibiting ERS.

DISCUSSION

The present study examined the feasibility of using thermal imagery and whether participants could learn to regulate their own brain activity through this approach. To this end, participants were immersed in a virtual environment with a small bonfire placed in front of them and instructed to imagine non-harmful heat sensations on their right hand under three different instruction conditions: visual (animated bonfire), thermal (Peltier-based stimulation), or visuothermal (combining both). Then, they received visual feedback on their imagery performance through the bonfire animation. The following sections discuss the study's research questions presented in the Introduction.

Can somatosensory imagery be used to control a BCI and its efficiency be enhanced through user training?:

Our first objective was to determine whether ThI can elicit reliable neurophysiological responses suitable for BCI control. Using ERDS computed online during the imagery phase, we observed ERD over the left sensorimotor cortex at electrode CP3, during both runs with and without feedback. This pattern is consistent with findings in TI studies, where imagining pressure or vibratory tactile sensations induced lateralized ERD in the α and β bands [6–8] in the sensorimotor cortex. Beyond previous TI studies, our observation of an online ERD over the left sensorimotor cortex during right-hand thermal imagery is consistent with EEG evidence showing contralateral alpha desynchronization during tactile imagery, especially over central electrodes, and with source-imaging and fMRI studies [10] indicating that tactile imagery recruits contralateral sensorimotor and somatosensory regions with partial overlap with real tactile stimulation. These results support the feasibility of using ThI as a control strategy for BCI applications. Additionally, a compa-

table proportion of low or non-responders was observed, which is in line with the findings in the literature on motor imagery [4]. We then examined whether participants improved over time. When looking at the runs without feedback, we observed a significant improvement in participants' performance over the course of the experiment, even in the absence of feedback. However, the absence of a control group prevents us from concluding whether performance improvements are due to participant learning or to other factors, such as repeated exposure to thermal stimuli, which could also activate the somatosensory cortex in a similar way to ThI, as it has been shown with TI and tactile stimuli [6, 8]. No significant improvement but a tendency toward improvement was observed across the runs with feedback within each condition. One possible explanation is that the runs were conducted in close temporal succession, limiting the measurable impact of the training between two runs. These findings may indicate feedback-independent learning processes unfolding during BCI user training. It is also possible that performance improves naturally with repeated trials. We will test this hypothesis in a future double-blind study including a control group receiving sham feedback.

Does the type of instruction significantly affect BCI performance?:

During the experiment, we included an instruction phase preceding the imagery period aimed at helping and preparing participants to imagine a warm sensation, similar to Yao et al., 2018 [15]. We investigated whether significant differences exist between visual, haptic, and visuothermal modalities, with the goal of better guiding participants in performing the task. Analysis of the online performance revealed a significant difference between conditions, with a trend between the visuothermal instruction and the other two. A visual inspection of the results suggests a larger ERD for visual and thermal conditions compared to the visuothermal condition. This pattern may be consistent with previous findings suggesting that multimodal feedback may not be optimal during the early stages of learning, as novices may have difficulty integrating information from multiple modalities [21]. A similar effect could therefore be expected with multimodal instructions, especially for a combination of visual and thermal instructions, which differed in terms of continuity and discreteness. The spatial placement also differed, as the visual instruction consisted of a campfire presented in front of the participant, whereas the thermal instruction involved a warm stimulus applied to the participant's hand.

Limitations:

Several participants reported experiencing discomfort during the instruction phase of the visuothermal condition. In particular, the lack of coherence between the continuous visual part of the instruction, consisting of a continuous campfire animation, and the thermal part, delivered as a discrete thermal tactile stimulus applied to the back of the hand, was reported by multiple participants. This qualitative feedback will help us design fu-

ture protocols, in which more coherent visual stimuli will be proposed to better match the sensations perceived and imagined by the participants.

CONCLUSION

In BCI research, TI paradigms remain largely limited to the imagination of pressure or vibratory stimulation, despite the potential relevance of other sensory modalities. To the best of our knowledge, ThI has not previously been investigated as a BCI control strategy, including in user-training paradigms. In the present study, we examined whether imagined warm sensations could be detected from EEG signals, whether performance improved with training over time, and whether it could be enhanced through visual, thermal, or combined visuothermal instructions. Our results demonstrate that a modulation in brain activity is observable based on changes in sensorimotor rhythms in the α - β bands, and could be used to detect whether a participant is performing ThI, potentially serving as a control strategy for BCI. Participants also exhibited significant improvements in online performance across runs without feedback. Although instructions significantly influenced performance at a global level, further analysis did not reveal significant differences between the instruction modalities. However, there was a tendency toward stronger effects with visual and thermal instructions. Overall, this study provides evidence that a novel imagery paradigm using the imagination of warm sensations can be implemented. This may represent a promising direction for clinical applications and individuals who experience difficulties with classical imagery paradigms. Further analysis will investigate whether neural response to a thermal stimulus could predict ThI performance and its potential ability to influence thermal perception in healthy participants.

REFERENCES

- [1] McFarland DJ, Wolpaw JR. EEG-based brain-computer interfaces. current opinion in Biomedical Engineering. 2017;4:194–200.
- [2] Pfurtscheller G, Guger C, Müller G, Krausz G, Neuper C. Brain oscillations control hand orthosis in a tetraplegic. Neuroscience letters. 2000;292(3):211–214.
- [3] Roc A *et al.* A review of user training methods in brain computer interfaces based on mental tasks. Journal of Neural Engineering. 2021;18(1):011002.
- [4] Vidaurre C, Blankertz B. Towards a cure for BCI illiteracy. Brain topography. 2010;23(2):194–198.
- [5] Thompson MC. Critiquing the concept of BCI illiteracy. Science and engineering ethics. 2019;25(4):1217–1233.
- [6] Yao L, Jiang N, Mrachacz-Kersting N, Zhu X, Farina D, Wang Y. Performance variation of a somatosensory BCI based on imagined sensation: A large population study. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022;30:2486–2493.
- [7] Yakovlev L, Syrov N, Miroshnikov A, Lebedev M, Kaplan A. Event-related desynchronization induced by tactile imagery: An EEG study. eneuro. 2023;10(6).
- [8] Wen H, Zhong Y, Yao L, Wang Y. Neural correlates of motor/tactile imagery and tactile sensation in a BCI paradigm: A high-density EEG source imaging study. Cyborg and Bionic Systems. 2024;5:0118.
- [9] Morozova M, Yakovlev L, Syrov N, Lebedev M, Kaplan A. Cortical responses to tactile imagery: A high-density EEG study of the μ -rhythm event-related desynchronization and somatosensory evoked potentials. NeuroImage. 2025:121440.
- [10] Yoo SS, Freeman DK, McCarthy III JJ, Jolesz FA. Neural substrates of tactile imagery: A functional MRI study. Neuroreport. 2003;14(4):581–585.
- [11] Pillette L, Lotte F, N’Kaoua B, Joseph PA, Jeunet C, Glize B. Why we should systematically assess, control and report somatosensory impairments in BCI-based motor rehabilitation after stroke studies. NeuroImage: Clinical. 2020;28:102417.
- [12] Kessner SS, Bingel U, Thomalla G. Somatosensory deficits after stroke: A scoping review. Topics in stroke rehabilitation. 2016;23(2):136–146.
- [13] Mane R, Chouhan T, Guan C. BCI for stroke rehabilitation: Motor and beyond. Journal of neural engineering. 2020;17(4):041001.
- [14] Chen JC, Liang CC, Shaw FZ. Facilitation of sensory and motor recovery by thermal intervention for the hemiplegic upper limb in acute stroke patients: A single-blind randomized clinical trial. Stroke. 2005;36(12):2665–2669.
- [15] Yao L, Mrachacz-Kersting N, Sheng X, Zhu X, Farina D, Jiang N. A multi-class BCI based on somatosensory imagery. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2018;26(8):1508–1515.
- [16] Jakovljević M, Mekjavić IB. Reliability of the method of levels for determining cutaneous temperature sensitivity. International journal of biometeorology. 2012;56(5):811–821.
- [17] Troisi D *et al.* Studying the influence of contact force on thermal perception at the fingertip. In: International Conference on Human Haptic Sensing and Touch Enabled Computer Applications. 2024, 154–161.
- [18] Renard Y *et al.* OpenViBE: An Open-Source Software Platform to Design, Test and Use Brain-Computer Interfaces in Real and Virtual Environments. Presence Teleoperators & Virtual Environments. 2010;19.
- [19] McFarland DJ. The advantages of the surface Laplacian in brain-computer interface research. International Journal of Psychophysiology. 2015;97(3):271–276.
- [20] Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia. 1971;9(1):97–113.
- [21] Pillette L, N’Kaoua B, Sabau R, Glize B, Lotte F. Multi-session influence of two modalities of feedback and their order of presentation on MI-BCI user training. Multimodal Technologies and Interaction. 2021;5(3):12.

Movement-Onset-Aligned EEG for Generalizable Arm and Hand Movement Decoding: A Large-Scale Multi-Experiment Dataset

K. Sterk¹, K. Kostoglou¹, G.R. Müller-Putz^{1,2}

¹Institute of Neural Engineering, Graz University of Technology, Graz, Austria

²BioTechMed Graz, Graz, Austria

E-mail: gernot.mueller@tugraz.at

ABSTRACT: This work presents a curated, movement-onset-aligned electroencephalography (EEG) dataset comprising nine experiments with 188 able-bodied participants performing hand and arm movement execution tasks as well as rest conditions. Movement onsets were aligned using auxiliary sensors. All recordings were uniformly processed using causal filtering to ensure compatibility with online brain-computer interface (BCI) applications. Further processing steps include resampling, independent component analysis (ICA) for removal of ocular and noise components, and strict epoch rejection. Across experiments, the datasets include recordings with different channel configurations, covering 94 distinct EEG channels. In total, the dataset comprises over 95,000 annotated trials with detailed movement labels and is provided in a standardized HDF5 structure to facilitate flexible data selection in machine learning workflows. Signal analyses confirm the presence of physiologically plausible movement-related patterns over sensorimotor regions. This resource aims to support the development of cross-participant and cross-experiment generalizing movement decoders and transfer learning in EEG-based BCIs. The data will be released in the near future.

INTRODUCTION

Electroencephalography (EEG)-based brain-computer interfaces (BCIs) provide a direct communication pathway between the brain and external devices by translating neural activity to control signals. In motor-related BCIs, control signals are commonly derived from neural processes associated with voluntary motor activity, including movement execution [1], attempted movement [2], and motor imagery [3]. Such systems have attracted significant attention as assistive technologies for individuals with severe motor impairments. For example, EEG-based BCIs for persons with impairments of the upper limbs can be used to control (neuro)prosthetic devices [4].

In recent years, deep learning approaches have increasingly been explored for EEG decoding tasks due to their ability to learn hierarchical representations from training data. They have demonstrated promising results in a variety of BCI applications [5] by modeling nonlinear relationships and spatial-temporal dependencies within the data. However, deep learning models typically require

large amounts of training data to achieve robust performance and to generalize well. In practical BCI applications, training data are usually collected during calibration sessions with the intended user. These sessions are often limited in duration, resulting in relatively small subject-specific datasets, especially for motor-impaired users where prolonged calibration sessions can be physically and mentally demanding. Consequently, insufficient training data remain a major challenge for developing reliable decoding models for EEG-based BCIs.

One promising strategy to address this challenge is the usage of generic deep learning classifiers pre-trained on a larger population [6]. Such models may capture generalizable neural patterns associated with motor processes and can subsequently be adapted to individual users through transfer learning and fine-tuning with a relatively small amount of subject-specific data. Despite the growing interest in data-driven methods for EEG decoding, the availability of large, well-curated datasets suitable for developing and evaluating such approaches remains limited. In particular, datasets that integrate recordings across multiple experimental paradigms and participants while maintaining consistent preprocessing and annotation are still relatively scarce. To support research in this direction, the present work introduces a curated large-scale EEG dataset intended to facilitate the development of machine learning and specifically deep learning methods for motor-related EEG decoding in BCI applications.

MATERIALS AND METHODS

Raw Data: Data from nine experiments were used for the creation of the dataset. The experiments were conducted at the Institute of Neural Engineering, Graz University of Technology, between 2017 and 2025. The individual raw datasets were assigned identifiers A to I. All experiments comprise EEG recordings acquired during right-hand and/or arm movement execution tasks performed by able-bodied participants. The signals were acquired using gel-based EEG systems. While some experiments focused exclusively on either arm or hand movements, others used combined motions. Movements were executed in response to visual cues. Most experiments additionally included resting state recordings, either active rest or passive rest. Active rest denotes trial-based

rest periods explicitly cued to remain at rest, whereas passive rest refers to continuous resting periods without trial-specific cues. The original recordings differ in the number of EEG electrodes and the sampling frequency. An overview of all experiments is provided in Table 1.

Table 1: Overview of the datasets used with information on their original publication, number of participants, number of used EEG channels, and types of movements included

ID	Source	# Parti- cipants	# EEG Channels	Movement Types
A	[1]	15	61	Arm, Hand, Rest
B	[7]	15	58	Arm&Hand, Rest
C	[8]	20	60	Arm, Rest
D	[9]	20	60	Arm, Rest
E	[10]	15	61	Arm&Hand, Rest
F	[11]	34	61	Arm&Hand
G	[12]	25	60	Hand, Rest
H	[13]	22	60	Hand, Rest
I	[14]	22	60	Hand, Rest

Movement Onset Detection: Due to differences in the experimental paradigms, the time interval between cue onset and actual movement execution varied substantially across experiments. To ensure consistent alignment of the trials to the true movement onset, precise onset timings were extracted from auxiliary movement tracking systems. These systems include exoskeletons [1], data gloves [1][10], camera-based motion capture systems [8][9][12][13][14], pressure buttons [10], and Myo armbands [11]. Movement onsets were determined using threshold-based detection applied to task-relevant metrics, such as hand velocity or acceleration. Trials in which no movement could be reliably detected were excluded from further analysis. For the active rest condition, a virtual movement onset was defined to account for the temporal influence of the visual cue. To this end, the average delay d_{MO} between cue onset (CO) and movement onset (MO) was calculated for each participant during movement trials. The virtual onset for rest trials was then defined as $t_{MO,rest} = t_{CO,rest} + d_{MO} + r \cdot std(d_{MO})$, where $r \in [-1, 1]$ is a uniformly sampled random factor and $std(d_{MO})$ denotes the standard deviation of movement onset delays. The random delay was introduced to mimic the variability in participants' reaction times observed during movement trials.

EEG Preprocessing: The EEG data preprocessing pipeline was designed with two main objectives: (i) to transform data from heterogeneous experiments into a uniform structure, and (ii) to remove strong artifacts and noise while preserving task-relevant EEG information. Processing was performed using MNE-Python [15]. All filters were chosen as causal filters to allow the dataset to be used for online BCI applications. First, line noise was attenuated using notch filters at 50 Hz and 100 Hz. Subsequently, a 0.5 Hz to 90 Hz bandpass filter was applied. The filters were implemented as 4th-order causal Butterworth filters. EEG channels identified as floating (i.e., exhibiting strongly reduced band power across frequencies) were interpolated using spherical spline inter-

polation. The data were then resampled to 256 Hz using a causal anti-aliasing finite impulse response lowpass filter with Kaiser window [16]. Afterwards, independent component analysis (ICA) with Picard algorithm [17] was computed to identify and remove ocular artifacts. Further noise components, such as muscle artifacts and channel noise, were addressed using the MNE_ICALabel extension [18] for automated component classification. Components with a predicted artifact probability greater than 95% were removed. Additionally, components with class probability to represent noise or artifacts greater than 85% were removed if confirmed by visual inspection. Next, the data were segmented into 8-s epochs around the movement onsets, comprising 3 s pre-onset and 5 s post-onset. An additional epoch rejection step was performed to exclude residual noisy epochs. For each participant, the standard deviation was computed per channel across the entire recording. An epoch was rejected if any channel exhibited (i) a standard deviation exceeding five times or (ii) an absolute amplitude exceeding eight times the channel's global standard deviation. Finally, epochs were mapped to a uniform EEG channel layout comprising all electrodes used across experiments. Channels not present in the respective experiment were inserted and filled with NaN values to preserve structural consistency.

Dataset Structure: The final curated dataset is structured as hierarchical data format (HDF), version 5, files. To limit file size, each experiment is stored in a separate HDF5 file. Each file contains the processed and epoched EEG data, information on the timings of the movement onsets, and additional metadata stored as group or file attributes. Each experiment file *experiment_X.hdf5* is organized hierarchically as follows:

- File-level attributes
 - "Bandpass Filter Frequencies (Hz)" [Array of float]
 - "Notch Filter Frequencies (Hz)" [Array of integer]
 - "Channel Names (standard_1005)" [Array of string]
 - "GND" [string]
 - "Reference" [string]
 - "Sampling Frequency (Hz)" [integer]
 - "Event Codes" [Array of integer]
 - "Event Codes Meanings" [Array of string]
 - "Experiment ID" [string]
- Participant Group *Participant_X00*
- Session Group *Session_0* (only present in dataset I)
- Movement Group, either *Hand Movement*, *Arm Movement*, *Arm and Hand Movement* or *Rest*
 - "Movement Group" [string]
 - "Movement Details" [Array of string]
 - "Participant ID" [string]

Within a movement group, two data objects are stored:

- "eeg_data" [Array of float]: Epoched EEG data, of shape $(n_{epochs} \times n_{channels} \times n_{timepoints})$
- "events" [Array of integer]: Indicating timings of movement onsets, of shape $(n_{epochs} \times n_{timepoints})$

For dataset I, data are subdivided into sessions for each participant, reflecting the original experimental design, where multiple sessions were recorded per participant. The primary data objects are "eeg_data" and "events",

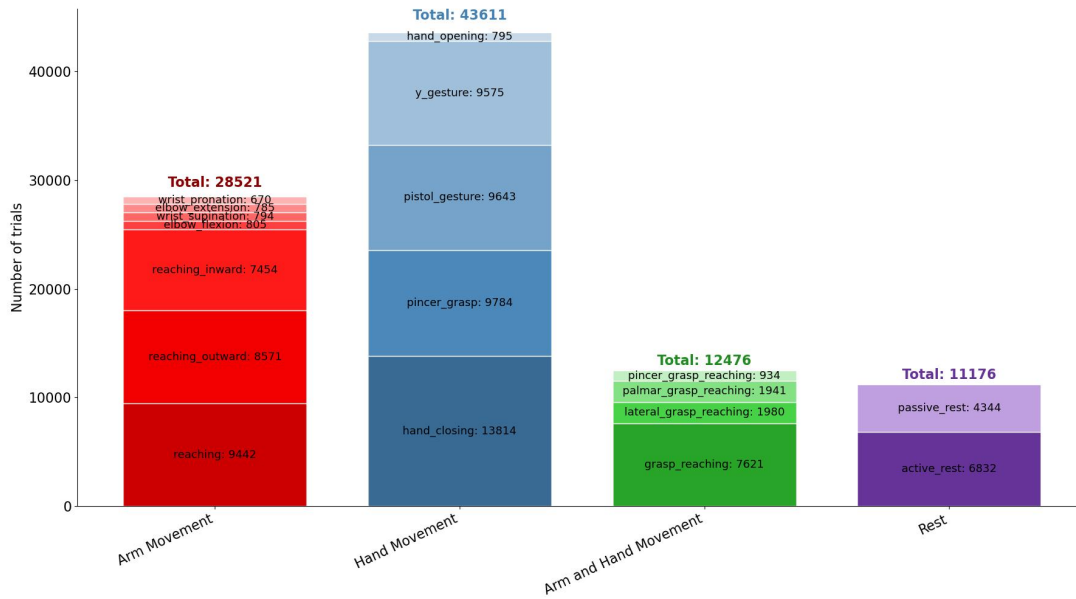


Figure 1: Number of trials per movement group of the processed dataset with occurrence of the detailed movement descriptions.

with "eeg_data" containing the processed EEG epochs and "events" array marking movement onsets (coded as 1 for movement onset, 0 otherwise). Note that although all epochs are centered around movement onsets and span 3 s pre-onset and 5 s post-onset, an epoch may contain multiple movement onset events. This occurs when inter-trial intervals are smaller than the chosen epoch size of 8 s.

RESULTS

Dataset Size: The dataset includes data from 188 participants and comprises 10 HDF5 files from 1.18 GB to 6.97 GB in size, resulting in a total volume of 38.5 GB.

Movement Groups: The EEG trials are organized into four movement groups with detailed movement labels:

- **Hand Movement:** hand closing, hand opening, pincer grasp, pistol gesture, y gesture
- **Arm Movement:** elbow extension, elbow flexion, reaching, reaching inward, reaching outward, wrist pronation, wrist supination
- **Arm and Hand Movement:** grasp reaching, lateral grasp reaching, palmar grasp reaching, pincer grasp reaching
- **Rest:** active rest, passive rest

Table 2: Number of trials per movement group of the processed dataset.

Movement Group	Number of Trials
Arm Movement	28521
Hand Movement	43611
Arm and Hand Movement	12476
Rest	11176
Total	95784

Trials in the Arm and Hand Movement group represent combinations of hand and arm actions (e.g., reaching

combined with grasping). When a further differentiation of grasping or reaching movements was not possible, trials were labeled generically as grasp or reaching. During preprocessing, 22.7% of trials were excluded, resulting in a total of 84,608 movement trials and 11176 rest trials. An overview of the number of trials per movement group is provided in Table 2, while the distribution of detailed movement labels is shown in Figure 1.

EEG Channels: Across experiments, the datasets include recordings with different EEG channel configurations (58-64 channels), covering 94 distinct channels overall, following the extended international 10-05 system [19]. A subset of 31 channels is present in all trials. The availability of individual channels relative to the total number of trials is illustrated in Figure 2.

Data Quality: The effect of data preprocessing

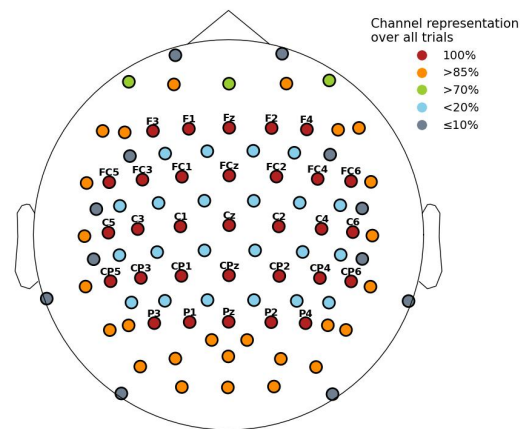


Figure 2: Channel representation with respect to the overall number of trials in the dataset. The channels in red are present in all trials. No channels had availability between 20% and 70%.

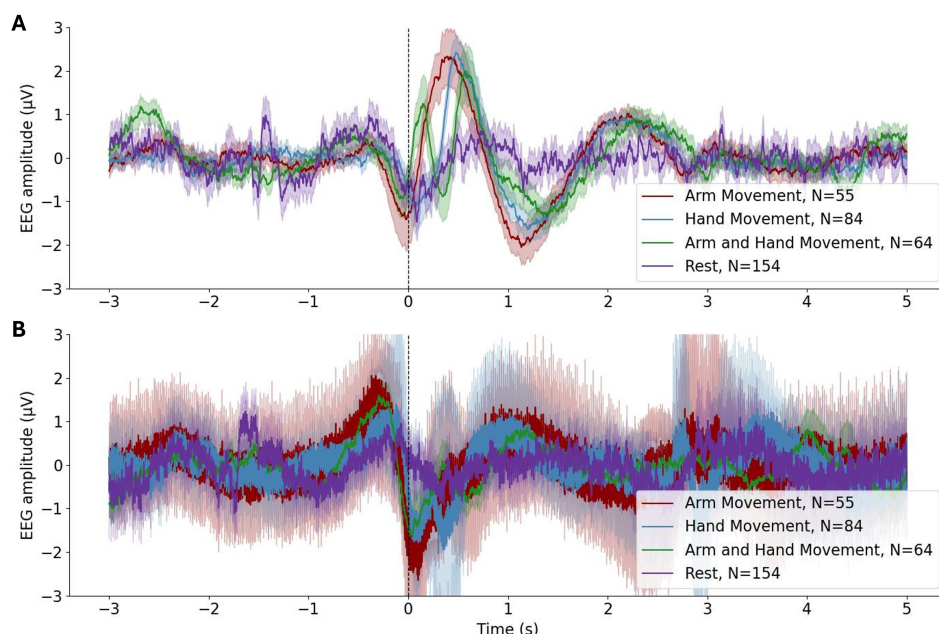


Figure 3: ERP of channel C1 per movement group for the fully and minimally processed data averaged over all trials of all participants with 95% confidence interval. N indicates the number of participants per movement group. Time point 0 s marks the movement onset. (A) ERP of the fully processed dataset. (B) ERP of the minimally processed data.

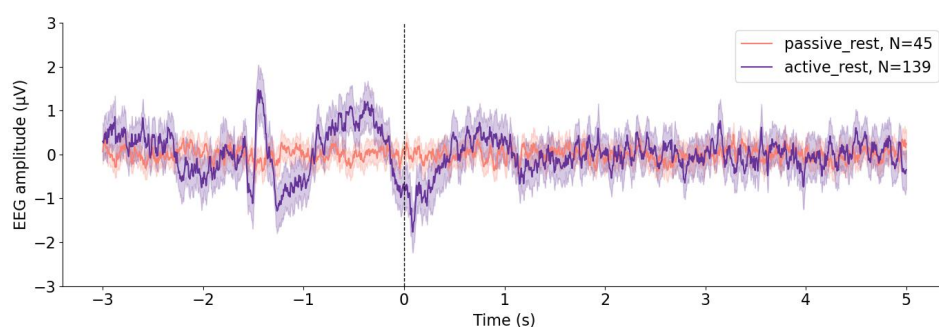


Figure 4: Activity of channel C1 of the fully processed data for movement details active and passive rest for the processed data averaged over all trials of all participants with 95% confidence interval. N indicates the number of participants per rest group. Time point 0 s marks the virtual rest onset.

was evaluated by comparing the event-related potentials (ERPs) derived from the fully processed dataset to those obtained from minimally processed data high-pass filtered at 0.5 Hz using a non-causal 4th-order Butterworth filter, without further artifact correction or trial rejection. For the minimally processed dataset a non-causal filter was used for the purpose of offline analysis and to preserve the temporal morphology of the ERPs by avoiding phase distortions or latency shifts. In Figure 3, the ERPs of channel C1 per movement group for the fully processed dataset (causal filtering) and the minimally processed data (non-causal filtering) are presented averaged over all trials of all participants with 95% confidence intervals. In both cases, characteristic movement-related cortical potential (MRCP) patterns [20] are visible, beginning prior to movement onset. In the fully processed dataset, the waveform exhibits a phase shift. Figure 4

presents the averaged activity of channel C1 of the fully processed dataset for the rest condition, separated into active and passive rest. The averaged passive rest ERP shows lower amplitudes and overall reduced low frequency oscillations. In Figure 5, for channel C1, the EEG signals were band-pass filtered in the alpha band (8 Hz to 12 Hz) and the beta band (13 Hz to 30 Hz), squared to estimate instantaneous power and smoothed over time. The resulting power time series were z-scored within each subject and then averaged across subjects separately for movement and rest trials. Around movement onset, the movement trials exhibit a decrease in band power, followed by an increase approximately 2.5 s after movement onset. This pattern reflects the classical event-related desynchronization (ERD) followed by event-related synchronization (ERS). In contrast, the rest trials do not show such patterns. Topographic maps of EEG amplitudes be-

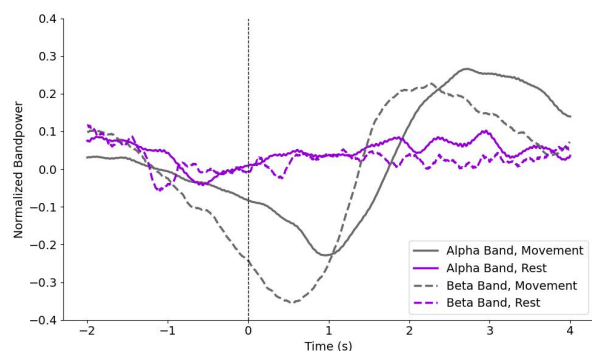


Figure 5: Time-varying bandpower of channel C1 for the alpha band (8 Hz to 12 Hz) and beta band (13 Hz to 30 Hz) averaged over movement trials (gray) and rest trials (purple). Time point 0 s marks the movement onset or virtual rest onset.

fore and after movement onset for the fully processed dataset, depicted in Figure 6, demonstrate increased central activation around movement onset and 0.5 s post-onset for movement conditions, whereas no comparable pattern is observed for rest. For visualization of the topographic maps, a Common Average Re-referencing (CAR) filter was applied, and only the subset of 31 channels present in all trials was used.

DISCUSSION

Dataset Structure: Aggregating nine experiments with 188 participants results in a large and diverse EEG dataset for hand/arm movement decoding. The unified HDF5 organization, structured by experiments, participants, and movement groups along with metadata such as detailed movement descriptions, enables flexible extraction of subsets tailored to specific research questions. Despite heterogeneous EEG channel montages, the 31-channel common subset available across all trials covers central motor areas. Moreover, domain-adaptation techniques [21] may allow the exploitation of the full channel information despite montage differences. The selected epoch length of 8 s, provides sufficient temporal context for subsequent processing steps while minimizing the risk of edge artifacts within the time window of interest.

Data Quality and Analysis: The averaged ERPs shown in Figure 3 demonstrate that the preprocessing pipeline effectively attenuated noise components, while preserving relevant signal characteristics. A clear difference in amplitudes around movement onset between the movement groups and the rest group indicate that MRCP patterns are captured in the dataset. The observed phase shift compared to the minimally processed data and the classical MRCP description [20] stems from the usage of purely causal filtering. This design choice was made to ensure compatibility with online BCI applications, where non-causal filtering is not feasible. Another aspect confirming the data quality is the difference between active and passive rest trials in Figure 4. As expected, the active rest condition exhibits increased activity, which indicates the occurrence of visual evoked potentials (VEPs)

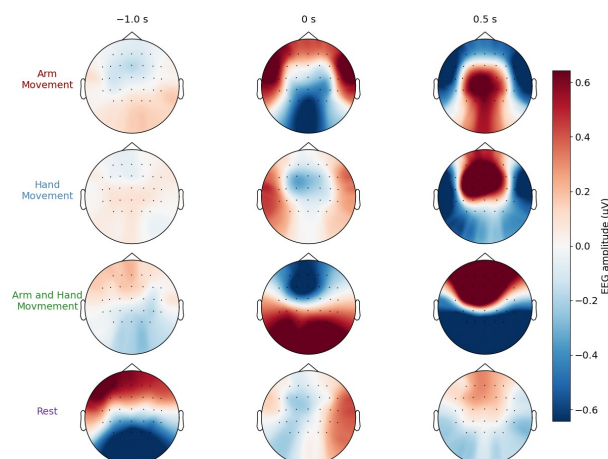


Figure 6: Topographic plots of the grand-average sensor-level EEG amplitude spatially interpolated across a 2D scalp representation of the fully processed dataset 1 s pre movement onset, during onset and 0.5 s post movement onset shown by movement group. For the plots, a CAR filter was applied. Only the subset of channels present in all trials were used.

caused by the visual cue or of contingent negative variation (CNV) signals caused by the expectancy of a motor-related task [22]. In contrast, the passive rest condition, where no cue was used, does not display such patterns. Similar components are visible in the movement conditions of Figure 3, suggesting that the observed signals reflect a combination of movement-related and cue-related activity. The bandpower differences, shown in Figure 5, further confirm the plausibility of the data. Alpha and beta band frequencies play a crucial role in movement-related tasks. ERD around movement onset, followed by ERS, are typically observed in the sensorimotor alpha and beta bands during voluntary movement [23] [24]. The topographic maps in Figure 6 further support the physiological plausibility of the data, revealing increased activation over central regions during and shortly after movement onset in movement conditions but not in rest trials.

Limitations: Despite its size and diversity, the presented dataset has certain limitations. The recordings stem from different experimental paradigms and hardware configurations, which introduces inter-experiment variability that may both be useful but also challenging for downstream tasks. Additional challenges for further machine-learning based usage of the data are the different channel montages and reference electrodes, which may be overcome by re-referencing techniques. Furthermore the dataset consists exclusively of movement execution tasks performed by able-bodied participants. While this supports transfer learning research, the generalization to movement attempts in clinical populations requires further investigation.

CONCLUSION

This work presents a curated, large-scale EEG dataset comprising nine experiments and 188 participants performing hand and arm movement execution tasks. The

data were uniformly processed, structured in a standardized HDF5 format, and extended with detailed metadata to support flexible data selection and transfer learning applications. Signal analyses confirm the presence of physiologically plausible movement-related patterns, while the preprocessing pipeline ensures compatibility with online BCI scenarios. The dataset is intended to facilitate the development and evaluation of generalizable deep learning models for transfer learning in EEG-based movement decoding and to support future research toward clinically applicable BCI systems. The dataset presented in this work will be publicly released in an online repository in the near future.

ACKNOWLEDGMENTS

This work was supported by the AUVA Grant "Neuro-Grasp".

REFERENCES

- [1] Ofner P, Schwarz A, Pereira J, Müller-Putz GR. Upper limb movements can be decoded from the time-domain of low-frequency EEG. *PLOS ONE*. 2017;12(8):e0182578.
- [2] Ofner P, Schwarz A, Pereira J, Wyss D, Wildburger R, Müller-Putz GR. Attempted Arm and Hand Movements can be Decoded from Low-Frequency EEG from Persons with Spinal Cord Injury. *Scientific Reports*. 2019;9(1):7134.
- [3] Motor imagery and EEG-based control of spelling devices and neuroprostheses. In: *Progress in Brain Research*. Elsevier, Jan. 2006, 393–409.
- [4] Müller-Putz GR et al. Applying intuitive EEG-controlled grasp neuroprostheses in individuals with spinal cord injury: Preliminary results from the More-Grasp clinical feasibility study. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*. 2019;2019:5949–5955.
- [5] Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*. 2019;16(3):031001.
- [6] Lopes-Dias C, Sburlea AI, Müller-Putz GR. A Generic Error-related Potential Classifier Offers a Comparable Performance to a Personalized Classifier. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Jul. 2020, 2995–2998.
- [7] Schwarz A, Escolano C, Montesano L, Müller-Putz GR. Analyzing and Decoding Natural Reach-and-Grasp Actions Using Gel, Water and Dry EEG Systems. *Frontiers in Neuroscience*. 2020;14.
- [8] Crell MR, Müller-Putz GR. Handwritten character classification from EEG through continuous kinematic decoding. *Computers in Biology and Medicine*. 2024;182:109132.
- [9] Srisrisawang N, Müller-Putz GR. Simultaneous encoding of speed, distance, and direction in discrete reaching: An EEG study. *Journal of Neural Engineering*. 2024;21(6):066042.
- [10] Schwarz A, Ofner P, Pereira J, Sburlea AI, Müller-Putz GR. Decoding natural reach-and-grasp actions from human EEG. *Journal of Neural Engineering*. 2017;15(1):016005.
- [11] Sburlea AI, Müller-Putz GR. Exploring representations of human grasping in neural, muscle and kinematic signals. *Scientific Reports*. 2018;8(1):16669.
- [12] Crell MR, Kostoglou K, Sterk K, Müller-Putz GR. A novel paradigm for fast training data generation in asynchronous movement-based BCIs. *Frontiers in Human Neuroscience*. 2025;19.
- [13] Suwandjeff P, Müller-Putz GR. EEG Analyses of visual cue effects on executed movements. *Journal of Neuroscience Methods*. 2024;410:110241.
- [14] Egger J, Kostoglou K, Müller-Putz GR. Chrono-EEG dynamics influencing hand gesture decoding: A 10-hour study. *Scientific Reports*. 2024;14(1):20247.
- [15] Gramfort A et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013;7.
- [16] Virtanen P et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17(3):261–272.
- [17] Ablin P, Cardoso JF, Gramfort A. Faster Independent Component Analysis by Preconditioning With Hessian Approximations. *IEEE Transactions on Signal Processing*. 2018;66(15):4040–4049.
- [18] Li A, Feitelberg J, Saini AP, Höchenberger R, Scheltienne M. MNE-ICLabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software*. 2022;7(76):4484.
- [19] Oostenveld R, Praamstra P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*. 2001;112(4):713–719.
- [20] Shibasaki H, Hallett M. What is the Bereitschaftspotential? *Clinical Neurophysiology*. 2006;117(11):2341–2356.
- [21] Zhang K et al. Application of Transfer Learning in EEG Decoding Based on Brain-Computer Interfaces: A Review. *Sensors*. 2020;20(21):6321.
- [22] Walter WG, Cooper R, Aldridge VJ, McCallum WC, Winter AL. Contingent Negative Variation : An Electric Sign of Sensori-Motor Association and Expectancy in the Human Brain. *Nature*. 1964;203(4943):380–384.
- [23] Pfurtscheller G, Silva FH Lopes da. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [24] Müller GR, Neuper C, Rupp R, Keinrath C, Gerner HJ, Pfurtscheller G. Event-related beta EEG changes during wrist movements induced by functional electrical stimulation of forearm muscles in man. *Neuroscience Letters*. 2003;340(2):143–147.

Benchmarking Machine Learning strategies to classify Steady-State Auditory Evoked Potentials.

Henrique Lefundes da Silva¹, Lenaïg Guého^{1,2}, Cyril Plapous¹, Laurent Bougrain^{2,3}, Patrick Hénaff⁴, Rozenn Nicol¹

¹NUX, Orange Innovation, F-22300 Lannion, France

²Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

³Sorbonne Université, ICM, CNRS, Inria, Inserm, Paris, France

⁴Univ Brest, Bretagne INP, CNRS Lab-STICC, Technopole Brest-Iroise - CS 83818, Brest, France

E-mail: laurent.bougrain@loria.fr

ABSTRACT: This study benchmarks classification pipelines for auditory Brain-Computer Interfaces (BCIs), with the objective of detecting the modulation frequency of amplitude-modulated sinusoidal auditory stimuli in electroencephalographic recordings. For this purpose, state-of-the-art Steady-State Visually Evoked Potential (SSVEP) classification algorithms are adapted to the auditory domain. This choice is motivated by the neurophysiological similarities between SSVEP and Steady-State Auditory Evoked Potential (SSAEP), particularly their shared mechanism of neural entrainment to periodic stimuli, despite differences in stimulation frequencies and the cortical regions involved. In this context, both preprocessing strategies and classification models are systematically assessed to identify optimal parameter configurations. Among the filtering techniques examined, a narrow-band Bessel filter yielded the highest accuracies, while the effect of temporal windowing was less pronounced under narrow-band conditions but remained beneficial for broader bandwidths. Riemannian-based methods, particularly tangent-space classifiers, consistently outperformed alternative approaches. Nevertheless, cross-subject classification proved more challenging due to pronounced inter-subject variability. These findings provide practical recommendations for auditory BCIs.

INTRODUCTION

Brain Computer Interfaces (BCIs) enable connected devices control by analyzing brain signals. Reactive BCIs use external stimuli to evoke brain responses [1]. Visual stimuli are commonly used, such as flickering objects at specific frequencies, generating Steady-State Visually Evoked Potentials (SSVEPs) in brain signals, enabling non-invasive detection of the stimulus frequency with ElectroEncephaloGraphy (EEG) [2]. However, SSVEP BCIs are not suitable for all users [3] or situations. Auditory stimuli, using amplitude-modulated sounds, can serve as an alternative or complement, producing Steady-State Auditory Evoked Potentials (SSAEPs) [3, 4]. Thus,

two signals are required: the carrier wave, i.e., the stimulus heard by the users, and the modulating sine wave at a particular frequency. Frequency (and its harmonics) which may be detected in brain activity [3]. Despite their potential, auditory BCIs remain significantly less explored than their visual counterparts.

However, EEG signal classification remains challenging due to substantial intra- and inter-subject variability [5]. In particular, differences in responsiveness to stimulation can affect detection performance across participants, as some individuals exhibit weak or no measurable evoked responses, further complicating reliable classification [6]. These issues can be mitigated through appropriate preprocessing strategies, including filtering, optimal window selection, and dimensionality reduction. Because preprocessing and classification are performed offline, the proposed approach is not constrained by real-time or pseudo-online requirements [7].

In this context, we aim to adapt state-of-the-art SSVEP classification algorithms to the auditory domain. This approach is motivated by the neurophysiological similarities between SSVEP and SSAEP responses, particularly their shared mechanism of neural synchronization in response to periodic stimuli. However, these modalities differ in their typical stimulation frequency ranges, approximately 3 to 80 Hz for the visual domain [8, 9] and about 1 to 200 Hz for the auditory domain [10], as well as in the cortical regions involved [5]. The main contribution of this work is the establishment of a benchmark comprising ten classification strategies for SSAEP-based BCIs. These strategies incorporate preprocessing steps such as filtering and windowing to facilitate the extraction of modulation frequencies from amplitude-modulated auditory stimuli in EEG signals. While recent large-scale benchmarks have addressed motor imagery, SSVEP, and P300 paradigms [11], no publicly available SSAEP dataset currently supports comparative evaluation.

MATERIALS AND METHODS

Experiment and dataset:

The experiment consisted in recording EEG signals from 24 subjects, evenly distributed by age (18–40) and sex (12 women and 12 men), in response to different types of sound stimuli, including both pure tones and natural sounds, at a loudness level of 50 phons (50 dB SPL at 1 kHz) per ear [12]. The experiment was conducted in accordance with the principles of the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects, and all participants provided informed consent. The experiment is conducted in a single session, which involves two 10-minute runs with a 5-minute break, during which subjects are exposed to 50 stimuli: 10 auditory stimuli of four different sounds (a sinus, a Brownian noise, a cicada song and a cat's purr) modulated at 40 Hz [13], and 10 silence stimuli. Stimuli are presented in a pseudo-random order, avoiding consecutive repetition of the same stimulus. At the beginning of the session, subjects complete several assessments to evaluate their hearing ability, handedness, musical knowledge, and mental state. Afterward, they rate their experience and stimulus preferences. EEG data are collected with a 24-channel cap (10-20 system) with passive electrodes using OpenVIBE software [14], sampled at 500 Hz. To reduce noise and artifacts, subjects are asked to remain still with their eyes closed during the recordings. For this study, the dataset includes only brain signals in response to the sinus and the silence stimuli.

Evaluation methods:

Since the dataset contains only a single session, model evaluation is performed in two ways [11]: Within-Session (WS), analogous to a k-fold cross-validation and Cross-Subjects (CS), similar to a Leave-One-Out evaluation. The WS evaluation provides a more statistically robust benchmark and is the most commonly used approach [11].

Preprocessing:

For filtering and temporal windowing analysis, a classifier with minimal hyperparameter tuning is chosen, the Minimum Distance to Mean (MDM) algorithm. WS evaluation is conducted, and accuracy is averaged across all subjects. The classification task is to detect the modulation frequency carried by sinusoids compared to silence, without modulation. Three commonly used EEG filters are considered: two IIR filters (Butterworth and Bessel) and one FIR filter [15]. The FIR filter is included despite not being state-of-the-art, as it remains widely used due to its numerical stability and ease of digital implementation. They are all centered on the modulation frequency, 40 Hz. The IIR filters are set to order 4, while the FIR filter uses a 0.1 Hz transition band (16500 taps, with a sampling frequency of 500 Hz). Evaluation uses 2-second non-overlapping windows. Thus, the optimal filter, identified in the *Bandpass filtering* Section, is applied for the analysis of temporal windowing.

Pipelines:

To identify the best processing chain, multiple techniques are evaluated across three main categories: 1) Spatial Filtering, which includes spatial filtering techniques combined with a classifier; 2) Riemannian, which leverages the geometry of Riemannian manifolds; 3) Deep Learning (DL) methods, particularly Convolutional Neural Networks (CNNs), which learn spatial and temporal filters as feature extraction.

Spatial filtering The state-of-the-art for SSVEP classification mainly consists of linear spatial filtering methods, such as Canonical Correlation Analysis (CCA) [16] and Common Spatial Pattern (CSP) [17]. CSP is often paired with a Linear Discriminant Analysis (LDA) classifier (CSP+LDA), using the log-variance of the spatially filtered signals as features. Despite its simplicity, this approach is robust, leveraging spatial patterns to enhance interpretability.

Riemannian Riemannian pipelines use covariance matrices, which are Symmetric Positive Definite (SPD) and therefore, lie on the Riemannian manifold. This approach offers geodesic distances better adapted to the Riemannian manifold [18]. The algorithm computes the Riemannian distance between the covariance matrix of the EEG segment to assign and the centroids of the covariance matrices of each class, the sample is then assigned to the category associated with the nearest centroid. The distance in the Riemannian space can be calculated by different metrics, in particular, affine-invariant metrics have shown strong performance in SSVEP paradigms [19]. Consequently, those metrics are favored in classifiers such as Minimum Distance to Mean (MDM) and Fisher Geodesic MDM (FgMDM) [18].

A limitation of the Riemannian manifold is that it does not directly support classical Euclidean-geometry classifiers such as Support Vector Machines (SVMs) and Logistic Regression (LR). However, when data are not linearly separable, SVMs can exploit kernel functions (e.g., a Riemannian kernel) to enable linear separation, an approach referred to as RiSVM [20]. Additionally, data can be projected onto the Tangent Space (TS) [18], becoming pipelines such as TS+LR and TS+SVM. However, TS projection produces high-dimensional vectors. Thus, to improve classification performance, dimensionality reduction techniques such as Principal Component Analysis (PCA) [21] or Kernel PCA (KPCA) [22] can be applied (TS+KPCA+LR).

In this study, all Riemannian algorithms are implemented using the *pyriemann* Python library, while machine learning algorithms are developed with *scikit-learn*.

Deep Learning (DL) DL models, such as CNNs, offer the advantage that they do not require manual feature engineering [23]. In EEG analysis, CNNs can extract spatial and temporal patterns directly from the data through their convolutional kernels such as EEGNet [24], ShallowConvNet [23] and GREEN [25]

Due to limited EEG data [26] and the complexity of DL models with thousands of trainable parameters, only simple architectures (i.e. EEGNet, ShallowConvNet and GREEN) are employed. The hyperparameters are adjusted to reduce the number of parameters to ensure stable training and reduce overfitting. These models are selected for their existing implementation in *braindecode* [23] and compatibility with *scikit-learn* pipelines.

RESULTS

Preprocessing:

Bandpass filtering The Bessel filter outperforms both the Butterworth and FIR filters, with optimal performance at a 0.1 Hz bandwidth (Fig. 1). Performance drops by about 10% per 0.01 Hz for narrower bandwidths, while for larger bandwidths, the decline is slower (2% per 0.1 Hz).

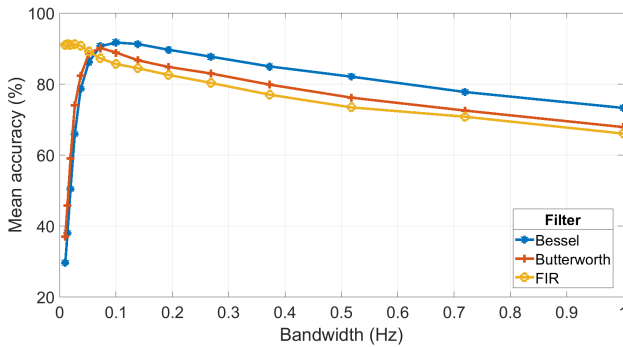


Figure 1: Effect of filter type and bandwidth on classification performance, using 2-second non-overlapping windows. Results are averaged across all subjects.

Temporal windowing A 4th-order Bessel filter with a 0.1 Hz bandwidth centered on the modulation frequency is used. Overlapping windows have little effect on classifier performance (Fig. 2), while increasing the window size to 2000 samples (4 s) improves results. This behavior reflects a fundamental trade-off: longer windows enhance the covariance estimation [27], but reduce the number of available training samples; conversely, shorter windows increase the number of training instances at the expense of less reliable estimates. Importantly, this trade-off is influenced by the filter bandwidth. Larger windows yield an average performance gain of approximately 8% for bandwidths above 0.5 Hz. In contrast, for narrower bandwidths, this advantage diminishes, with window lengths between 1000 and 2500 samples (2–5 s) providing the best performance. Windows between 750 and 1500 samples (1.5–3 s) offer a good balance between sample size and class representation. Therefore, the window size is fixed at 1000 samples (2 s) without overlap for the subsequent analyses.

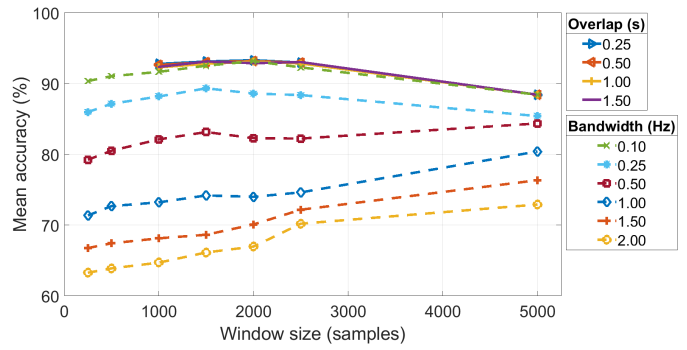


Figure 2: Effect of window size for various overlapping configurations and bandwidths (without overlap), using the optimal Bessel filter (order 4, 0.1 Hz bandwidth). Results are averaged across all subjects.

Artifact rejection No explicit artifact correction or trial rejection procedure was applied in this study. This choice is motivated by both methodological and application-oriented considerations. From an application perspective, the proposed framework is intended for future real-time BCI systems, where online artifact correction and trial rejection remain challenging and may introduce additional latency or instability.

Instead, artifact mitigation is primarily achieved through experimental design and signal processing. Participants were instructed to remain still, seated, and with eyes closed during the experiment, thereby reducing common sources of artifacts such as ocular and muscular activity. In addition, a very narrow band-pass filter centered on the modulation frequency (40 Hz) is employed. This filter strongly attenuates all spectral components outside the frequency band of interest, resulting in substantial suppression of most artifacts.

We also evaluated dedicated artifact rejection methods, such as the Riemannian “potato” approach [19], which removes outlier covariance matrices. However, this method did not yield any measurable improvement in classification performance in our setting. Furthermore, more advanced artifact correction techniques (e.g., channel reconstruction) were avoided, as they may distort the covariance structure and produce ill-conditioned matrices, which is particularly problematic for Riemannian-based classifiers.

Classification:

Since selecting appropriate hyperparameters is crucial for optimal performance, a grid search is performed by evaluating all combinations of a predefined set of values for each parameter (Tab. 1). For WS evaluation, a grid search is conducted using 4-fold inner cross-validation and a 5-fold outer cross-validation, whereas for CS, a simple leave-one-out cross-validation is performed.

Due to its computational cost, especially for DL models, this process is applied to other machine learning models, adjusting DL models based on these results. The fixed hyperparameters for deep learning are listed in Table 2 for Within-Session (WS) and Cross-subjects (CS).

Table 1: Hyperparameters Grid for WS evaluation paradigm.

Step	Hyperparameters	Values
TS	-	-
KPCA	n_components	[2, 8, 16, 24, 32, 40, 80]
	kernel	[linear, rbf]
LR	C	[0.001, 0.01, 0.1, 1, 10]
SVM	C	[0.001, 0.01, 0.1, 1, 10]
	kernel	[linear, rbf]
LDA	-	-
MDM	metric	[riemann, logeuclid]
FgMDM	metric	[riemann, logeuclid]
RiSVM	C	[0.001, 0.01, 0.1, 1, 10]
CSP	n_filters	[2, 4, 8, 12, 16]

Within-Session (WS) The accuracy varies significantly depending on the classifier (Friedman test with Bonferroni correction, $p < 0.001$) (Fig. 3a). The highest accuracy is achieved by the RiSVM pipeline, which performs similarly to other tangent space-based methods, as well as GREEN and MDM algorithms. Lower accuracies are observed with EEGNet, ShallowConvNet, FgMDM, and CSP+LDA. EEGNet consistently yields the lowest performance.

Cross-Subjects (CS) Again, classification results differ significantly across pipelines (Friedman test with Bonferroni correction, $p < 0.001$) (Fig. 3b). The best-performing algorithm is RiSVM. Similar to other tangent space-based algorithms, its accuracy is significantly higher than that of EEGNet, ShallowConvNet, FgMDM, and CSP+LDA, except for TS+SVM, which has a lower accuracy than the other tangent space-based methods and does not differ significantly from CSP+LDA. The lowest accuracy is obtained by EEGNet.

DISCUSSION

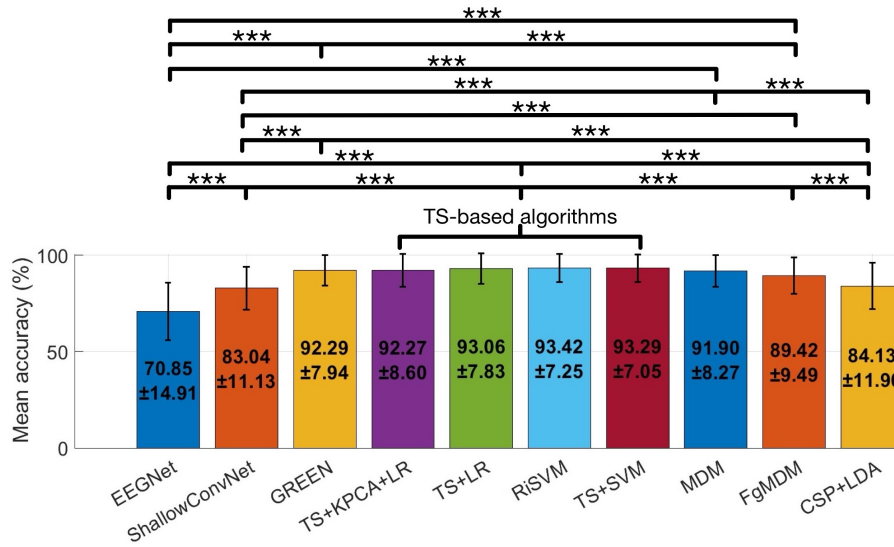
Using an appropriate filter can enhance classification performance by over 15% (Fig. 1). The Bessel filter yields the best results, while the FIR filter performs the worst. However, accurately isolating the modulation frequency is the most influential parameter. Bessel filters offer flat group delay and smooth roll-off, helping to reduce noise and preserve signal quality. Overlapping windows do not impact classification performance (Fig. 2), but window size does. This effect is related to filter characteristics, diminishing as filter quality improves. For bandwidths up to 0.5 Hz, longer windows enhance accuracy, aligning with previous studies indicating that extended windows improve covariance estimation by accumulating relevant patterns [27]. In fact, narrow transition bands and bandwidths lead to FIR filters with long delays, which are problematic for online implementations. Thus, a wider bandwidth can be compensated by using a larger window. However, excessively large windows may cause impractical delays and increase computational load. It should be noted that these parameter settings may not be optimal for all classification algorithms. The underlying assumption

Table 2: Deep Learning Models' Hyperparameters.

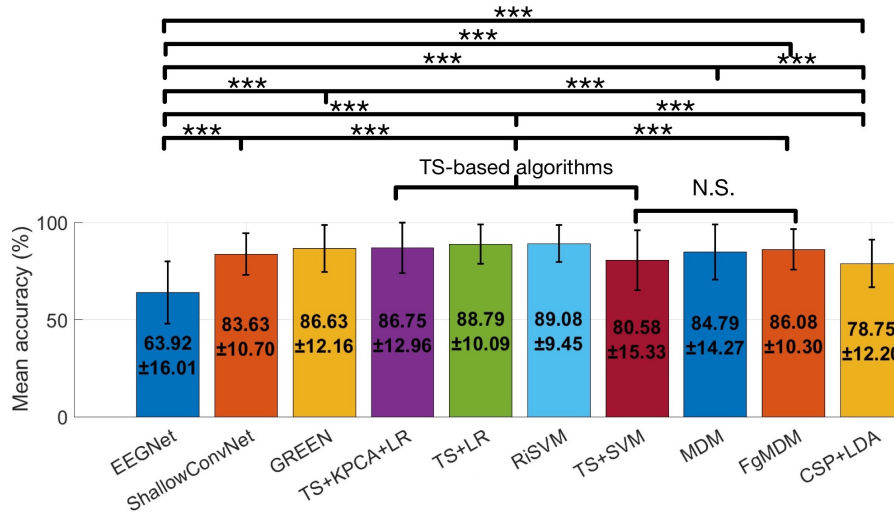
Hyperparameter	Value WS	Value CS
General		
Batch Size	16	16
Loss	BCEWithLogits	BCEWithLogits
Optimizer	Adam	Adam
Epochs	20	30
Learning Rate	0.0005	0.001
EEGNet		
F1	2	2
D	12	12
F2	24	24
kernel_length	32	32
Batch Size	16	16
SCN		
n_filters_time	2	2
n_filters_spat	8	12
pool_time_length	100	100
pool_time_stride	25	25
pool_mode	mean	mean
Batch Size	16	32
GREEN		
n_freqs	5	5
kernel_width_s	0.1	0.1
hidden_dim	32	100
bi_out	24	24
Batch Size	16	32

is that, since most evaluated methods are based on Riemannian geometry (e.g., distance-based approaches or tangent space projections), optimizing performance with the MDM classifier effectively maximizes class separability. This, in turn, tends to increase inter-class distances both on the manifold and after projection into the tangent space, suggesting that improvements observed with MDM are likely to transfer to related approaches.

For both evaluation methods, most pipelines effectively detect the modulation frequency. Riemannian-based methods, especially those based on tangent space, consistently achieve the highest accuracies and outperform others, in line with the MOABB benchmark [11]. GREEN, FgMDM and MDM perform slightly lower, but their results are statistically comparable. Although the latter requires minimal hyperparameter tuning, it offers faster training and maintains comparable robustness. GREEN's performance is expected, as it relies on covariance matrices and TS projection. However, for practical use, tangent space-based algorithms and MDM are preferable to GREEN due to lower computational demands and simpler training procedures. In contrast, deep learning models such as EEGNet and ShallowConvNet, along with CSP+LDA, always achieve the lowest performance due to the limited training data in the WS configuration and the DL methods evaluation simplified to prioritize lightweight, easy-to-tune approaches like Riemannian methods. CS classification (Fig. 3b) includes both intra-subject variability and differences between sessions,



(a) Within-session averaged accuracy across all subjects for each pipeline.



(b) Cross-subject averaged accuracy across all subjects for each pipeline.

Figure 3: Averaged accuracy across all subjects for each pipeline under both evaluation methods. Bar values represent the mean accuracy $\pm$ standard deviation (%). *** indicates $p < 0.001$ (Wilcoxon signed-rank test with Bonferroni correction).

resulting in lower performance compared to WS results across all pipelines, except for ShallowConvNet. Tangent space models, except TS+SVM, continue to outperform other models, demonstrating their ability to generalize across sessions and subjects effectively. The average performance drop across the different evaluations is 5%. To the best of our knowledge, no study has conducted a comprehensive benchmark of preprocessing strategies, particularly band-pass filtering and temporal windowing, in combination with classifier comparison for SSAEP-based BCIs. While similar efforts exist for SSVEP, these findings are not directly comparable to the auditory domain due to differences in stimulation frequencies [8–10] and the cortical regions involved [28].

CONCLUSION

This study benchmarks classification pipelines for auditory BCI, aiming to detect the modulation frequency of

amplitude-modulated sinusoidal sounds in EEG signals. It assesses classification and preprocessing to determine optimal parameters. Applying a narrow-band Bessel filter significantly improves performance and tends to minimize the impact of temporal windowing. On the other hand, this can be compensated for later by broader bandwidths. Tangent space-based classifiers outperform other models. As expected, cross-subject classification remains more challenging due to higher inter-subject variability. In future work, alternatives to sine carriers will be explored, looking for more pleasant sound stimuli. The detectability of the modulation frequency will be assessed using the same classification methods. Additionally, deep learning tools, such as transformer-based models, will be investigated and compared to Riemannian algorithms. Although transformers are challenging to train within sessions, cross-subject experiments with data augmentation could provide sufficient data.

REFERENCES

- [1] Wolpaw JR. Brain-computer interfaces. In: *Handbook of clinical neurology*. Elsevier, 2013, 67–74.
- [2] Vialatte FB, Maurice M, Dauwels J, Cichocki A. Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives. *Progress in neurobiology*. 2010;90(4):418–438.
- [3] Kim DW, Cho JH, Hwang HJ, Lim JH, Im CH. A vision-free brain-computer interface (bci) paradigm based on auditory selective attention. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011, 3684–3687.
- [4] Heo J, Baek HJ, Hong S, Chang MH, Lee JS, Park KS. Music and natural sounds in an auditory steady-state response based brain-computer interface to increase user acceptance. *Computers in Biology and Medicine*. 2017;84:45–52.
- [5] Clerc M, Bougrain L, Lotte F. *Les interfaces cerveau-ordinateur 1: Technologie et applications*. ISTE Group (2016).
- [6] Vidaurre C, Blankertz B. Towards a cure for bci illiteracy. *Brain topography*. 2010;23(2):194–198.
- [7] Carrara I, Papadopoulos T. Pseudo-online framework for bci evaluation: A moabb perspective using various mi and ssvep datasets. *Journal of Neural Engineering*. 2024;21(1):016003.
- [8] Odom JV et al. Visual evoked potentials standard (2004). *Documenta ophthalmologica*. 2004;108(2):115–123.
- [9] Allison BZ, McFarland DJ, Schalk G, Zheng SD, Jackson MM, Wolpaw JR. Towards an independent brain-computer interface using steady state visual evoked potentials. *Clinical neurophysiology*. 2008;119(2):399–408.
- [10] Picton TW, John MS, Dimitrijevic A, Purcell D. Human auditory steady-state responses: Respuestas auditivas de estado estable en humanos. *International journal of audiology*. 2003;42(4):177–219.
- [11] Chevallier S et al. The largest eeg-based bci reproducibility study for open science: The moabb benchmark. 2024.
- [12] Fletcher H, Munson WA. Loudness, its definition, measurement and calculation. *Bell System Technical Journal*. 1933;12(4):377–430.
- [13] Kaongoen N, Jo S. A novel hybrid auditory bci paradigm combining assr and p300. *Journal of neuroscience methods*. 2017;279:44–51.
- [14] Renard Y et al. Openvibe: An open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments. *Presence*. 2010;19(1):35–53.
- [15] Sen D, Mishra BB, Pattnaik PK. A review of the filtering techniques used in eeg signal processing. In: *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*. 2023, 270–277.
- [16] Lin Z, Zhang C, Wu W, Gao X. Frequency recognition based on canonical correlation analysis for ssvep-based bcis. *IEEE transactions on biomedical engineering*. 2007;54(6):1172–1176.
- [17] Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller KR. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*. 2007;25(1):41–56.
- [18] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain-computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*. 2011;59(4):920–928.
- [19] Chevallier S, Kalunga EK, Barthélemy Q, Monacelli E. Review of riemannian distances and divergences, applied to ssvep-based bci. *Neuroinformatics*. 2021;19(1):93–106.
- [20] Barachant A, Bonnet S, Congedo M, Jutten C. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing*. 2013;112:172–178.
- [21] Shlens J. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*. 2014.
- [22] Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*. 1998;10(5):1299–1319.
- [23] Schirrmester RT et al. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*. 2017;38(11):5391–5420.
- [24] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of neural engineering*. 2018;15(5):056013.
- [25] Paillard J, Hipp JF, Engemann DA. Green: A lightweight architecture using learnable wavelets and riemannian geometry for biomarker exploration with eeg signals. *Patterns*. 2025;6(3).
- [26] Lashgari E, Liang D, Maoz U. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*. 2020;346:108885.
- [27] Chevallier S, Kalunga E, Barthélemy Q, Yger F. Riemannian classification for ssvep-based bci: Offline versus online implementations. 2018:371–396.
- [28] Guého L, Bougrain L, Plapous C, Hénaff P, Nicol R. Assessing stimuli detectability and pleasantness for auditory bci. In: *BIOSIGNALS 2026-19th international conference on bio-inspired systems and signal processing*. 2026.

Capacity-normalized Shannon graph complexity as an EEG biomarker for event-level cognitive load estimation in adaptive BCIs

Beatriz Pascual-Roa^{1,2}, Eduardo Santamaría-Vázquez^{1,2,3}, Diego Marcos-Martínez^{1,2},
C. Rubén Ruiz-Gálvez^{1,2}, Víctor Martínez-Cagigal^{1,2,3}, Roberto Hornero^{1,2,3}

¹Biomedical Engineering Group, University of Valladolid, ETSIT, Valladolid, Spain

²CIBER-BBN, Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Madrid, Spain

³Valladolid Health Research Institute (IBioVALL), Valladolid, Spain

E-mail: beatriz.pascual@uva.es

ABSTRACT: Cognitive load (CL) reflects the mental resources required to perform a task. Adaptive brain-computer interfaces (BCIs) could benefit from EEG biomarkers that estimate CL with sufficient temporal granularity to support online adaptation. However, many EEG-based CL studies still define CL according to pre-defined task levels or objective task difficulty, which may not impose comparable cognitive demands across users. This assumption is especially problematic when individuals differ in working-memory capacity. We propose an event-level and capacity-normalized framework based on EEG functional connectivity in frontal and parietal regions. During the Corsi Block-Tapping Test, each encoded item was labeled according to its position relative to each participant's individual maximum working-memory capacity, allowing cognitive demand to be expressed as a percentage of individual capacity. Functional connectivity was quantified using orthogonalized amplitude envelope correlation (AEC) across theta, alpha, low- and high-beta, and gamma frequency bands, while network organization was summarized through Shannon graph complexity (S_C). The approach was evaluated in 62 cognitively healthy older adults (mean age = 70.4 ± 2.9 years). Across all frequency bands, S_C showed a progressive increase as task demands approached individual working-memory capacity. Capacity-normalized CL stages differed significantly after FDR correction ($p < 0.05$), with mostly large effect sizes ($r > 0.5$). These findings suggest that S_C , combined with capacity-normalized event-level labeling, may provide a compact and online-compatible EEG biomarker for CL-aware BCI adaptation in cognitively healthy older adults.

INTRODUCTION

Brain-computer interfaces (BCIs) perform best when they adapt to the user. This becomes challenging when the user operates under high cognitive load (CL). CL refers to the mental resources required to encode, maintain, and manipulate information during a task [1]. When task demands approach the user's limits, performance be-

comes unstable: errors increase and responses slow down [2]. In real-world scenarios, this can compromise safety and reliability. For this reason, closed-loop BCIs require CL estimates that are both temporally precise and individualized to each user, as the same task demand may impose different cognitive burdens depending on individual cognitive capacity and neural dynamics.

CL is often estimated with self-reports, such as NASA-TLX [3], or with behavioral measures such as accuracy and reaction time [1]. These tools are useful, but they are limited for online use. Self-reports are subjective and typically retrospective [4]. Behavioral measures provide indirect information, as they reflect performance outcomes rather than the underlying neural state [4]. In contrast, physiological signals offer a more direct window into brain activity. Electroencephalography (EEG) is particularly interesting in this context. EEG can track fast brain dynamics with high temporal resolution and can be integrated into portable BCI setups [5]. Moreover, EEG is the most commonly used signal in BCI systems, making it a natural candidate for continuous CL estimation within adaptive BCI frameworks.

Recent EEG-based CL research has increasingly explored single-trial decoding, machine-learning models, functional connectivity features, and online-compatible BCI pipelines [6, 7]. These advances show that EEG can support temporally resolved CL estimation beyond classical block-level analyses. However, a remaining challenge is how CL labels are defined. Many studies still assign CL according to objective task difficulty or pre-defined task levels, even though the same task level may represent low demand for one user and near-capacity demand for another. In our previous work, we showed that CL changes within a task level as items accumulate in working memory during the Corsi Block-Tapping Test (CBT) [8]. Frontal theta and parietal alpha mainly reflected general engagement, whereas beta activity tracked faster within-trial variations [8]. These findings suggest that item-level labeling can capture CL progression more precisely than block-level or level-level labels. However, item position alone is still not sufficient for cross-user

comparability, because the same item can represent different proportions of individual working-memory capacity.

Beyond local spectral power, CL also affects how brain regions coordinate [9]. Working memory and executive control rely on fronto-parietal interactions [10]. Functional connectivity can capture these interactions and has been increasingly used for CL estimation, including recent work based on effective or functional connectivity and machine-learning approaches [6, 11]. Graph-theoretical measures can further summarize network organization into compact features suitable for online-compatible pipelines [12]. However, connectivity-based CL studies typically still assign labels from external task levels rather than from each participant's individual cognitive capacity. Thus, combining connectivity features with capacity-normalized event-level labels may provide a more individualized representation of CL.

Two main challenges are therefore important for BCI-oriented CL biomarkers. First, features should capture CL at the time scale of cognitive events, rather than only at the level of task blocks. Second, CL labels should be normalized across users, so that CL levels reflect the proportion of individual cognitive capacity being engaged. A fixed task level does not impose the same CL on all users [13], which is especially relevant in adaptive systems where personalization is central.

The goal of this study is to evaluate whether an EEG biomarker based on network complexity can track capacity-normalized CL at the item level. To address this challenge, we propose a connectivity-based framework that combines event-level CL labeling with normalization based on each participant's Corsi span. Functional connectivity is estimated using orthogonalized amplitude envelope correlation (AEC) and summarized using Shannon graph complexity (S_C). The novelty of the proposed framework lies in combining S_C with a capacity-normalized event-level labeling strategy that expresses each encoded item as a percentage of individual working-memory capacity. This framework is evaluated in 62 cognitively healthy older adults.

MATERIALS AND METHODS

Participants and EEG recordings:

We analyzed 62 cognitively healthy older adults (46 female), aged 65-75 years (70.37 ± 2.85 years). Cognitive status was screened using the Mini-Mental State Examination (MMSE), and participants scoring below 23 points were excluded. Additional exclusion criteria included neurological or psychiatric disorders and current treatment with anxiolytic or antidepressant medication. The study was approved by the local ethics committee (PI 23-3325) and all participants provided written informed consent in accordance with the Declaration of Helsinki. EEG was recorded during the CBT implemented in the MEDUSA© ecosystem [14] using the ITACA CBT application [15]. EEG was acquired with a g.Nautilus PRO

system (g.tec Medical Engineering GmbH, Austria) using 16 active hybrid electrodes at Fp1, Fp2, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P3, Pz, P4, and POz (10-10 system). Conductive gel was used to improve contact quality. The common reference and ground electrodes were placed on the right and left mastoids, respectively. The sampling rate was 250 Hz. Participants followed a longitudinal protocol of nine sessions, with two CBT recordings per session.

Corsi Block-Tapping Test: CBT is a standard task to assess short-term visuospatial working memory [16]. In each trial, blocks illuminate one by one in a fixed order. Participants must reproduce the sequence in the same order. Figure 1 shows a snapshot of the digital CBT implemented within the MEDUSA© ecosystem. The encoding phase presented each item for 500 ms, with a 1000 ms interstimulus interval. Sequence length increased after correct trials, up to a maximum of nine items. In this dataset, the task started at two items. Each level included one series of two trials. A level was passed if at least one trial was correct. This incremental structure increases cognitive demands in a controlled way and allows CL to be examined at the item level as information accumulates within each trial [8].

EEG preprocessing and epoching: Preprocessing was implemented in MEDUSA© Kernel [14]. Signals were filtered with an IIR band-pass filter (0.5-60 Hz). Line noise was removed with a notch filter (49-51 Hz) [17]. We segmented EEG into 1 s epochs aligned to each encoded item. We focused on encoding because it is the stage where CL changes most within a trial [18]. To preserve compatibility with lightweight and online-compatible BCI pipelines, preprocessing was intentionally kept simple. Epochs contaminated by artifacts were automatically rejected. An epoch was discarded if the amplitude exceeded four standard deviations from the mean in at least two samples in any channel [19].

Capacity-normalized event-level labeling: We assigned CL labels at the item level, relative to each participant's visuospatial capacity. We first estimated the individual Corsi span S_i as the maximum sequence length recalled with a success ratio ≥ 0.5 across sessions [16]. This value provides a stable estimate of the participant's upper performance limit and supports cross-subject normalization.

Each trial contains a sequence of length L and item indices $k = 1, \dots, L$. Each encoded item defines one epoch. We assigned each epoch to one of four CL stages based on the fraction k/S_i . The stages cover 0-25%, 25-50%, 50-75%, and 75-100% of individual capacity:

$$\text{Stage}(k; S_i) = \begin{cases} 1, & 0 < \frac{k}{S_i} \leq 0.25 \\ 2, & 0.25 < \frac{k}{S_i} \leq 0.50 \\ 3, & 0.50 < \frac{k}{S_i} \leq 0.75 \\ 4, & 0.75 < \frac{k}{S_i} \leq 1.00. \end{cases} \quad (1)$$

This scheme captures within-trial CL progression and makes CL labels comparable across participants with

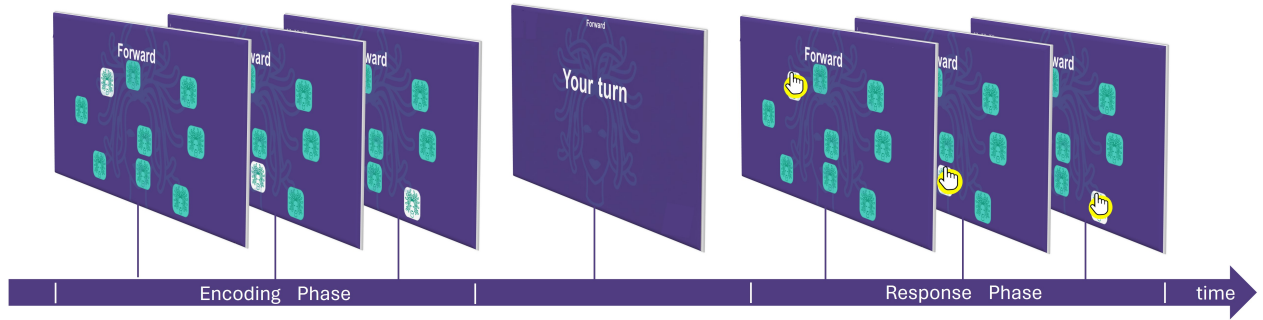


Figure 1: Snapshot of the digital Corsi Block-Tapping Test implemented in the MEDUSA© platform. During the encoding phase, blocks illuminate sequentially. After a brief interval, participants reproduce the sequence in the response phase. Sequence length increases after correct trials, progressively increasing CL.

different spans. Thus, the four stages should not be interpreted as arbitrary task phases, but as capacity-normalized CL stages ranging from low demand to near-capacity demand.

Functional connectivity and S_C : For each epoch, we estimated functional connectivity between all electrode pairs. We used AEC computed from the analytic signal after the Hilbert transform [20]. To reduce spurious zero-lag coupling associated with signal leakage and volume conduction, we used orthogonalized AEC [21]. This produced a weighted 16×16 connectivity matrix A , where A_{ij} reflects coupling strength between electrodes i and j . We summarized network organization using S_C [22]. This metric combines the dispersion of edge weights and the entropy of their distribution:

$$S_C = H \sqrt{\frac{1}{T-1}} \sigma_w, \quad (2)$$

where H is the Shannon entropy of the edge-weight distribution, T is the number of edges, and σ_w is the standard deviation of edge weights. We computed S_C for five frequency bands: theta (4-8 Hz), alpha (8-13 Hz), low beta (13-20 Hz), high beta (20-30 Hz), and gamma (30-60 Hz) [23]. We did not analyze delta because 1 s epochs do not capture its slow cycles reliably [24].

To obtain region-level features, we averaged node-level values within two regions of interest. The frontal region included F7, F3, Fz, F4, and F8, whereas the parietal region included P3, Pz, and P4. These regions were selected to capture broad fronto-parietal network trends associated with working memory and executive control, which are consistently involved in CL modulation [10]. The use of relatively broad ROIs was intended to prioritize robust network-level effects over hemispheric specialization in this event-level framework.

Statistical analysis: We assessed differences in S_C across the four CL stages using the Wilcoxon signed-rank test. Normality was checked with Shapiro-Wilk tests and was typically rejected. We controlled for multiple comparisons using the Benjamin-Hochberg false discovery rate (FDR) procedure with $\alpha = 0.05$ [25]. Effect sizes

were computed with the rank-based correlation coefficient r , suitable for non-parametric comparisons [26]. We interpret $r > 0.5$ as a large effect and $r > 0.3$ as a medium effect [26].

RESULTS

We examined how S_C varied across the four capacity-normalized CL stages in the frontal and parietal regions. The analysis was performed separately for each frequency band. Fig. 2 summarizes these effects. The left panels show boxplots of S_C across stages for each band and region. The right panels present FDR-corrected p -value matrices and corresponding effect sizes for pairwise comparisons. Across bands, S_C increased as the encoded items approached each participant's span, with the largest increase typically occurring between the 50-75% and 75-100% stages. In contrast, the first three stages showed more gradual changes. This trend was visible in both regions and was consistent across subjects. Spatially, frontal regions generally exhibited higher S_C values than parietal regions in theta and alpha bands during low-to-intermediate CL stages (0-25%, 25-50%, and 50-75%). However, in the highest CL stage (75-100%), parietal complexity became comparable to or greater than frontal complexity in these lower-frequency bands. In contrast, low beta, high beta, and gamma bands consistently showed higher frontal complexity across all CL stages. Pairwise comparisons revealed statistically significant differences between CL stages in all frequency bands and regions after FDR correction ($p < 0.05$). In most comparisons, effect sizes were large ($r > 0.5$), indicating strong separability between CL stages.

DISCUSSION

In this study, we combined item-level CL labeling with a capacity-normalized labeling framework based on each participant's Corsi span. For every encoded item, we estimated EEG functional connectivity using orthogonalized AEC and summarized the resulting networks with S_C across canonical frequency bands.

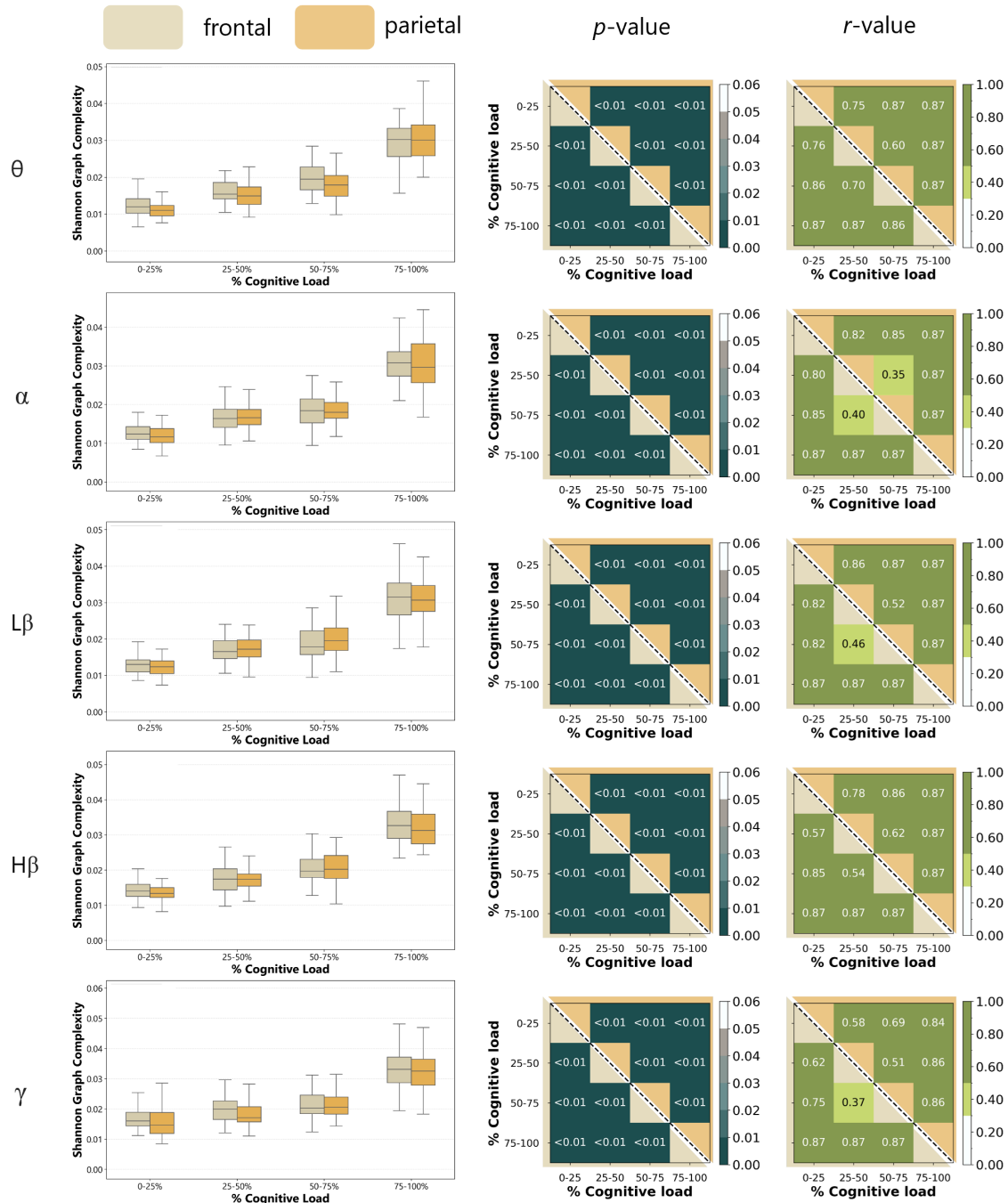


Figure 2: S_C across capacity-normalized CL stages. Each row corresponds to a frequency band: theta (4-8 Hz), alpha (8-13 Hz), low beta (13-20 Hz), high beta (20-30 Hz), and gamma (30-60 Hz). Left: compact boxplots of S_C for each stage, shown separately for frontal and parietal regions. The central line denotes the median and the box spans the interquartile range. Right: matrices of FDR-corrected p -values (left matrix) and effect sizes (r , right matrix) for pairwise stage comparisons. The lower-left triangle corresponds to frontal comparisons and the upper-right triangle to parietal comparisons. Non-white cells indicate significant differences ($p < 0.05$). Darker shades in the effect-size matrices denote larger effects.

Relation to previous EEG-based CL studies: Recent EEG-based CL research has demonstrated the feasibility of single-trial decoding, online-compatible pipelines, and connectivity-based approaches for CL estimation [6, 7]. However, many studies still define CL according to external task difficulty or predefined CL levels, which

may not represent equivalent cognitive demands across users. Connectivity-based approaches using phase- and amplitude-based metrics have also been proposed [11]. In contrast, our framework assigns CL labels at the item level and anchors them to each participant's Corsi span. The resulting stages therefore reflect the proportion

of individual working-memory capacity being engaged, which may improve comparability across users and better match the requirements of adaptive BCIs.

Interpretation of network complexity changes: S_C summarizes the heterogeneity of connectivity weights across the network [22]. Lower values reflect a more homogeneous organization, whereas higher values indicate a broader distribution of coupling strengths. In our data, S_C progressively increased as encoded items approached each participant's working-memory span. This pattern suggests that near-capacity demands may be associated with increasingly heterogeneous functional connectivity configurations. Because S_C is a compact network-level descriptor, these findings should be interpreted as reflecting global changes in connectivity organization rather than specific edge-level mechanisms. Nevertheless, the consistent increase observed across CL stages indicates that capacity-normalized cognitive demands are accompanied by systematic changes in functional network structure. Interestingly, these effects were observed across several frequency bands, which may suggest that CL-related network reorganization is not restricted to a single oscillatory component. From a practical perspective, this broadband sensitivity could be advantageous for future online CL estimation, as it may allow flexible feature selection depending on signal quality and recording conditions. Regional differences were also observed across frequency bands. Theta and alpha bands showed relatively stronger frontal complexity during low-to-intermediate CL stages, whereas parietal complexity became more prominent near individual capacity limits. In contrast, beta- and gamma-band complexity remained predominantly frontal across stages. Although these spatial patterns should not be interpreted as direct evidence of specific neural mechanisms, they may reflect different contributions of fronto-parietal systems as cognitive demands approach individual working-memory capacity. This interpretation is consistent with the established involvement of fronto-parietal networks in working memory and executive control [10].

Implications for online-compatible CL estimation in BCIs:

Closed-loop BCIs require CL estimates that are temporally precise, stable, and individualized to each user. Our approach addresses three practical constraints. First, it produces labels at the level of individual encoded items, which is closer to the temporal scale at which adaptive BCIs may update feedback or task difficulty. Second, it includes capacity normalization, so that high CL reflects proximity to each user's own working-memory limit rather than a fixed external task level [13]. Third, S_C is a single scalar metric computed from a connectivity matrix, which keeps the pipeline lightweight. This facilitates online deployment compared with more complex dynamic approaches. In a CL-aware BCI, S_C could be used as a continuous control variable to adjust task pacing or difficulty, trigger adaptive feedback, or prevent overload during demanding interaction periods.

Limitations and next steps:

This work has several limitations. First, the distribution of epochs across stages can be imbalanced, with fewer samples near capacity. This may affect the stability of estimates in the highest CL stage. Second, the study focuses on a single task and a specific age group. Although participants were cognitively healthy older adults, mental-state properties and EEG biomarkers are known to change with age. Therefore, the present findings should be interpreted within the context of healthy aging. Future work should test the generalization to younger adults, other age ranges, and other CL paradigms. In addition, the present analysis focused on broad frontal and parietal regions to capture robust network-level trends. Future work should investigate hemispheric asymmetries and finer-grained regional effects, as lateralized neural dynamics may provide additional information about CL-related network organization. Third, we used inferential statistics but did not evaluate online classification performance. Although the proposed framework is online-compatible, the present analysis was performed offline and focused on statistical separability between capacity-normalized CL stages. Future work should evaluate online implementation, detection latency, classification robustness under realistic BCI constraints, and subject-independent models using S_C as an input feature.

CONCLUSION

In this study, we addressed the challenge of characterizing CL for BCI-oriented monitoring by leveraging EEG-derived functional brain networks under progressively increasing task demands. We introduced an event-level methodological framework that combines item-level labeling with individual capacity-based normalization and connectivity analysis. This approach allows CL progression to be described relative to each participant's working-memory capacity, rather than only according to objective task difficulty. Our results show that S_C tracks CL across canonical frequency bands in a cohort of 62 cognitively healthy older adults. As task demands approached individual capacity, frontal and parietal connectivity patterns showed progressively higher network complexity. Significant differences between capacity-normalized CL stages, together with mostly large effect sizes, support the sensitivity of this metric to graded changes in cognitive demand. These findings suggest that S_C , when combined with capacity-normalized event-level labeling, may serve as a compact EEG biomarker for CL-aware adaptive BCIs. The metric is derived from functional connectivity, summarizes network organization in a single value, and is suitable for event-level updates. Overall, this work contributes an individualized and online-compatible framework for event-level CL estimation and provides a basis for future adaptive BCI systems that respond to the user's cognitive state.

ACKNOWLEDGMENTS

This work was supported by '0124_EU-ROAGE_MAS_4_E project', co-financed by the EU through the Interreg VI-A Spain-Portugal Programme (POCTEP) 2021-2027, and by the VA140P24 project, funded by the Regional Government of Castilla y León (Consejería de Educación) and EU-ERDF. Support has also been provided by the Centre for Biomedical Research in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN, CB19/01/00012) through the Carlos III Health Institute, co-financed with ERDF funds.

REFERENCES

- [1] Sweller J. Cognitive load during problem solving: Effects on learning. *Cognitive Science*. 1988;12(2):257–285.
- [2] Cowan N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*. 2001;24(1):87–114.
- [3] Hart SG, Staveland LE. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: *Advances in Psychology*. North-Holland, 1988, 139–183.
- [4] Moray N. Subjective mental workload. *Human Factors*. 1982;24(1):25–40.
- [5] Haapalainen E, Kim S, Forlizzi JF, Dey AK. Psychophysiological measures for assessing cognitive load. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp 2010)*. 2010, 301–310.
- [6] Safari MR et al. Classification of mental workload using brain connectivity and machine learning. *Scientific Reports*. 2024;14:10050.
- [7] Zhou Y et al. Cognitive load recognition in simulated flight missions: Exploring temporal dynamics of multiple load states using eeg signals. *Frontiers in Human Neuroscience*. 2025;19:1542774.
- [8] Pascual-Roa B et al. EEG biomarkers of cognitive load: Insights from incremental element encoding in short-term working memory. *Biomedical Signal Processing and Control*. 2026;112:108511.
- [9] Ozcelik E. The neural correlates of cognitive load in learning: An fmri study on graph comprehension. *Learning and Instruction*. 2025;99:102175.
- [10] Kim JS et al. Changes in effective connectivity according to working memory load: An fmri study of face and location working memory tasks. *Psychiatry Investig.* 2012;9(3):283–292.
- [11] Dong M et al. Eeg classification of resting state and arithmetic cognitive workload using functional connectivity of different frequency bands and machine learning techniques. *Brain-Computer Interfaces*. 2025.
- [12] Liu Y et al. Fusion of spatial, temporal, and spectral EEG signatures improves multilevel cognitive load prediction. *IEEE Transactions on Human-Machine Systems*. 2023;53(2):357–366.
- [13] Sweller J. Cognitive load theory and individual differences. *Learning and Individual Differences*. 2024;110:102423.
- [14] Santamaría-Vázquez E et al. Medusa©: A novel python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. *Comput. Methods Programs Biomed.* 2023;230.
- [15] Marcos-Martínez D et al. Itaca: An open-source framework for neurofeedback based on brain-computer interfaces. *Computational Biology and Medicine*. 2023;160(March):107011.
- [16] Kessels RPC, Zandvoort MJE van, Postma A, Kappelle LJ, Haan EHF de. The corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*. 2000;7(4):252–258.
- [17] Urigüen JA, García-Zapirain B. EEG artifact removal - state-of-the-art and guidelines. *Journal of Neural Engineering*. 2015;12(3).
- [18] Pinal D, Zurrón M, Díaz F. Effects of load and maintenance duration on the time course of information encoding and retrieval in working memory: From perceptual analysis to post-categorization processes. *Frontiers in Human Neuroscience*. 2014;8:165.
- [19] Núñez P et al. Exploring non-stationarity patterns in schizophrenia: Neural reorganization abnormalities in the alpha band. *Journal of Neural Engineering*. 2017;14(4):046001.
- [20] Brookes MJ et al. Measuring functional connectivity using meg: Methodology and comparison with fcmri. *NeuroImage*. 2011;56(3):1082–1104.
- [21] Colclough GL, Brookes MJ, Smith SM, Woolrich MW. How reliable are MEG resting-state connectivity metrics? *NeuroImage*. 2016;138:284–293.
- [22] Gomez-Pilar J et al. Quantification of graph complexity based on the edge weight distribution balance: Application to brain networks. *International Journal of Neural Systems*. 2018;28(02):1750032.
- [23] Tatum IV WO. Handbook of EEG interpretation. *Analytical Biochemistry*. 2014;2(1):1–5.
- [24] Cohen MX. Analyzing neural time series data: Theory and practice. MIT Press: Cambridge, MA (2014).
- [25] Sarkar SK. Fdr-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference*. 2004;125(1-2):119–137.
- [26] Pautz N, Olivier B, Steyn F. The use of nonparametric effect sizes in single study musculoskeletal physiotherapy research: A practical primer. *Physical Therapy in Sport*. 2018;33:117–124.

Interpretability of CNN-Based EEG Decoding During Mirror Box Training: A Pilot Study Using CWT Spectrograms and Explainable AI

Adrián Hodúr¹, Roman Rosipal¹, Nina Evetović¹, Zuzana Rošćáková¹, Jozef Jakubík¹

¹Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia

E-mail: adrian.hodur@savba.sk

ABSTRACT: Neural networks are increasingly applied in brain-computer interface research; however, their decision-making process is often considered difficult to interpret. This pilot study investigates whether a convolutional neural network can provide not only accurate but also neurophysiologically meaningful classification of electroencephalogram signals recorded during mirror box training. Two conditions, rest and left hand movement, were analyzed with a focus on interpretability. EEG signals from contralateral electrode C3 were transformed into time-frequency representations using the Continuous Wavelet Transform, and the resulting spectrograms were used as input to the convolutional neural network (CNN). To interpret the model's decisions, Gradient-weighted Class Activation Mapping and other methods were applied, enabling identification of discriminative time-frequency regions. By relating these patterns to established neuroscientific knowledge, we assess whether the model relies on physiologically meaningful features, providing insight into brain dynamics during mirror-induced motor activity and supporting interpretable BCI applications.

INTRODUCTION

Neural networks (NNs) have become a highly popular tool in artificial intelligence due to their ability to model nonlinear relationships and detect complex patterns in classification tasks. However, despite their strong performance, their internal parameters (optimized during the training process) are often difficult to interpret, particularly in deep and highly complex architectures. This limited interpretability presents a challenge in applications where understanding the decision-making process is essential.

Electroencephalogram (EEG) signals, which measure the electrical activity of the brain, are inherently complex and difficult to interpret. The combination of complex NN models and complex neurophysiological signals motivates our effort to develop a NN-based classification model with an emphasis on interpretability. In this study, we classify two labeled conditions recorded during mirror box training: **(1) rest**, during which the patient observes a fixation cross, and **(2) exercise**, during which the patient performs one of four movements with left hand.

To enhance the time-frequency interpretability of EEG

signals, we applied the Continuous Wavelet Transform (CWT) [1], which enables clear detection of temporal changes in frequency bands. This is particularly important because movement or motor imagery are associated with event-related desynchronization (ERD) and event-related synchronization (ERS). ERD corresponds to a decrease in power spectral density in the μ and β sensorimotor rhythms, whereas ERS represents an increase in these bands [2]. During movement, decreased power in the μ and β bands is typically observed, while rest is characterized by stronger μ activity.

For classification, CWT spectrograms were computed from EEG and these spectrograms were represented as images and used as input to a convolutional neural network (CNN) [3] adapted to our dataset and classification task.

To evaluate model interpretability, we employed explainable AI (XAI) methods, primarily Gradient-weighted Class Activation Mapping (Grad-CAM) [4], and for comparison also Local Interpretable Model-agnostic Explanations (LIME) [5] and Occlusion Sensitivity (OS) [6]. These methods highlight input regions that contribute most to classification decisions and were complemented by frequency-based analyses to assess whether the model focuses on neurophysiologically meaningful patterns.

Although the individual components of the proposed approach (CWT, CNN, and XAI methods) are well established, their combination is applied here within a neurophysiologically motivated framework with an emphasis on interpretability. This pilot study, conducted on a single subject, aims not only to distinguish rest and movement conditions but also to determine whether the model relies on physiologically meaningful features, such as frequency-specific synchronization and desynchronization in sensorimotor rhythms. By linking time-frequency representations with model explanations, the approach enables a transparent evaluation of consistency between deep learning decisions and established EEG knowledge in the context of mirror box training. Additionally, the proposed method was compared with the EEGNet architecture [7], which operates directly on raw EEG signals.

MATERIALS AND METHODS

Recording Setup:

EEG signals were recorded using a wireless g.Nautilus

PRO FLEXIBLE system (g.tec Medical Engineering GmbH, Austria), an FDA-cleared device designed for both clinical and research applications. The system was equipped with 28 Ag/AgCl wet electrodes arranged according to the international 10-20 system. The electrodes were integrated into a multi-channel EEG cap connected to a portable amplifier unit, enabling high-quality signal acquisition during experimental procedures. Signals were sampled at 250 Hz. During acquisition, A1 served as the reference electrode and AFz as the ground. For offline processing, the EEG data were re-referenced using a common average reference. Specifically, for each electrode, the mean signal across all recorded electrodes was computed and subtracted from the signal at that electrode, thereby reducing reference-related bias in subsequent analyses.

Raw signal preprocessing:

Raw EEG recordings were first converted to the European Data Format (EDF) and subsequently preprocessed using the BrainVision Analyzer software (version 2.3.03). The preprocessing workflow consisted of several steps aimed at improving signal quality and minimizing artifact contamination. We used a semi-automatic threshold-based artifact detection procedure implemented to identify segments contaminated by high-amplitude or non-physiological activity. All automatically marked segments were visually inspected to ensure accurate classification. Electromyographic (EMG) artifacts were also identified prior to exporting the cleaned dataset for further processing.

Dataset:

EEG data were recorded from a single participant during 3 mirror box (MB) training sessions. Continuous EEG signals were acquired throughout both resting and movement conditions. During the movement periods, each task with left hand was repeated multiple times. Each MB session included four distinct exercises with self-paced pauses between repetitions:

- **Exercise 1: Elbow-Supported Hand Lift** - The participant lifted the hand while keeping the forearm in contact with the table.
- **Exercise 2: Palm Rotation** - The participant rotated the palm upward and downward while the forearm remained on the table.
- **Exercise 3: Ball Roll** - The participant rolled a ball on the table using the palm, lifting the fingers as the ball reached the base of the palm.
- **Exercise 4: Bottle Grasp and Release** - The participant grasped a bottle, released it, and returned the hand to the table.

All movements were performed with the left hand while observing its mirror reflection, creating a visual illusion of movement. This mirror-induced illusion is known to modulate activity in the contralateral sensorimotor cortex [8]. For this reason, EEG signals from electrode **C3** were selected for further analysis.

EEG Preprocessing:

The preprocessing pipeline consisted of five steps: **i)**

Epoching: Continuous EEG signals were segmented into 2-second epochs using a sliding-window approach with 80% overlap between consecutive segments. **ii) Frequency selection:** Only frequency components within the 4-25 Hz range (0.5 ~ Hz resolution) were retained for further analysis. **iii) Time-frequency transformation:** Each epoch was transformed using the CWT (see next section with CWT). From each resulting spectrogram, 125 temporal samples corresponding to the 2-second interval were extracted, forming the input matrices for the CNN. **iv) Initial segment removal:** The first two seconds of each task period were excluded to avoid contamination from head movements and task transition effects. **v) Artifact handling:** All spectrograms containing artifacts were excluded from further analysis. After preprocessing, 2107 spectrogram images with size 43×125 (frequency $\times$ time) were obtained from movement and rest conditions. For validation, data from one specific movement task and its corresponding post-movement rest period were reserved. The remaining movement and rest data were used for CNN training.

CWT for creating input data matrices:

We generated 2-second-long CWT spectrograms within the 4-25 Hz frequency range using the CWT. The CWT provides a complex time-frequency representation of a signal by projecting it onto scaled and translated versions of a mother wavelet. The analytic Gabor-Morlet wavelet [9] was used as mother wavelet. This wavelet provides near-optimal joint time-frequency localization due to its Gaussian envelope and is well suited for oscillatory, non-stationary signals such as EEG. For further analysis, the magnitude of the CWT representing the local signal energy was used. Because lower frequencies require a longer temporal context for accurate magnitude estimation, an additional 100 time points were temporarily appended before and after each 2-second epoch during the CWT computation. These additional samples were used solely to improve the estimation of low-frequency components at the window boundaries and were discarded after the transformation. Consequently, only the original 2-second long spectrograms were retained for further analysis.

Convolutional Neural Network Architecture:

Because the size of our dataset was limited, we designed a simple CNN with a small number of convolutional layers. The architecture was inspired by the well-known AlexNet model [10]. The network was employed to classify CWT-derived EEG spectrograms. The model consists of an input layer followed by convolutional blocks and fully connected layers. The number of trainable parameters was reduced to mitigate overfitting. Each convolutional layer is followed by batch normalization (BN) and the exponential linear unit (ELU) activation function. The combination of BN and ELU activation has been shown to improve performance in deep CNN models applied to EEG data [11]. The architecture of the convolutional blocks after the input layer is summarized in Table 1.

After the convolutional blocks, the feature maps were

Table 1: Overview of the first three convolutional blocks of the proposed CNN. K - filter size, N - number of filters, S - stride, P - padding

Block	Component	Parameters
Conv1	Filters (K/N/S/P)	11×11, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Max 1×5, 1×5, same
Conv2	Filters (K/N/S/P)	5×5, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Max 2×2, 2×2, same
Conv3	Filters (K/N/S/P)	3×3, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Max 2×2, 2×2, same

flattened and passed to a fully connected layer with 16 units, followed by ELU activation and dropout (rate = 0.4) to further reduce overfitting. The final fully connected layer contained 2 neurons, and a softmax function was applied to produce class probabilities.

For comparison, we also evaluated the proposed approach using an EEGNet-based architecture [7] operating directly on raw EEG signals. The input data were band-pass filtered in the 4-25 Hz range and centered around zero mean. The network was adapted to the input size of 1×500 , corresponding to a 2-second EEG segment sampled at 250 Hz. To ensure a fair comparison, the architecture was adjusted while preserving the core design principles of EEGNet. The detailed configuration of the adapted network is summarized in Table 2.

Table 2: Overview of the blocks of EEGNet. K - filter size, N - number of filters, S - stride, P - padding

Block	Component	Parameters
Conv1	Filters (K/N/S/P)	1×125, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Avg, 1×4, 1×4 same
	Dropout (rate)	0,25
Conv2	Filters (K/N/S/P)	1×32, 32, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Avg, 1×8, 1×8, same
	Dropout (rate)	0,25
Final	Full. connected (output)	2
	Soft max classification	

XAI:

To interpret the decisions of the CNN, Grad-CAM [4] was employed. Grad-CAM provides a visual explanation by highlighting the regions of the input that contribute most strongly to a specific class prediction. For comparison, we also used (OS) [6] and LIME [5] methods.

Training options:

For data preprocessing and CNN training, we used MATLAB R2025b. The network was trained using the cross-entropy loss function with class weight correction to account for the imbalance between rest and movement data. The following training options were applied for effective learning. Training was performed using the Adam optimizer [12] with a mini-batch size of 32 and an initial learning rate of 3×10^{-4} . A piecewise learning-rate schedule with periodic decay was employed to ensure

stable convergence during later epochs. L2 regularization with $\lambda = 1 \times 10^{-4}$ was applied to reduce overfitting. Early stopping was implemented based on validation performance, with a patience of 8 and validation checks every 50 iterations. These hyperparameters were selected based on preliminary experiments, which indicated that this configuration provided a good balance between convergence speed, model stability, and generalization performance.

Additional Supplementary Methods:

To assess the neurophysiological validity of the CNN-based analysis we applied traditional EEG analysis methods, including the IRASA method [13] and PARAFAC decomposition [14], to estimate the spatial and spectral characteristics of the EEG signal.

RESULTS

EEG spatial and frequency analysis:

IRASA was applied to separate oscillatory and aperiodic components of the amplitude spectrum at electrode C3 during resting and movement conditions. Baseline recordings included one minute of eyes-closed (EC) and one minute of eyes-open (EO) rest. As shown in Figure 1 (left), a broader oscillatory peak within 8-12 Hz was present at C3 in both conditions. To assess visual alpha activity, we additionally analyzed electrode O1 (Figure 1, right). A pronounced peak around 9 Hz was observed, particularly during the eyes-closed condition, corresponding to occipital α rhythm. Given that occipital alpha can spatially spread toward parietal and central regions due to large-scale thalamo-cortical synchronization [15], this may partly explain the broader and less sharply defined peak at C3 during EC, reflecting overlap between sensorimotor μ and occipital α activity.

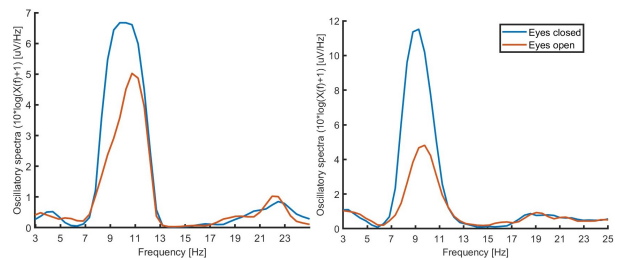


Figure 1: Oscillatory part of the amplitude spectrum as obtained by IRASA method/decomposition. **Left:** Electrode C3. An oscillatory peak is visible between 8 and 12 Hz. EC - blue curve, EO - orange curve. **Right:** Electrode O1. A dominant oscillatory peak is visible around 9 Hz, corresponding to occipital alpha activity. EC - blue curve, EO - orange curve.

During MB training (Figure 2), a clear oscillatory peak around 11 Hz was observed at C3 during rest, corresponding to the μ rhythm. This peak was markedly reduced during left-hand movement, indicating ERD. A comparable desynchronization pattern was also present at O1, reflecting suppression of visual α activity during movement.

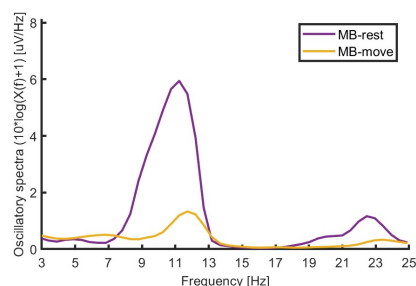


Figure 2: IRASA oscillatory spectra at electrode C3 during MB training. The dark red curve represents the resting/pause condition, while the orange curve corresponds to left-hand movement. A clear reduction of μ amplitude is observed during movement, indicating event-related desynchronization.

PARAFAC analysis further confirmed the presence of a task-related μ component. As illustrated in Figure 3, the extracted component showed a peak frequency of approximately 11.2 Hz and was localized over the left sensorimotor cortex (C3 region).

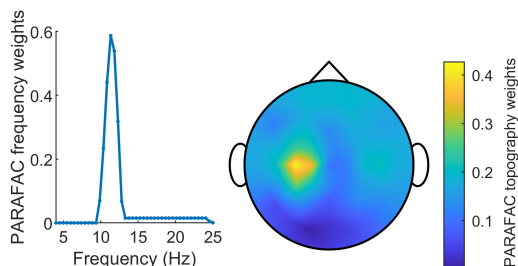


Figure 3: **Left:** Frequency profile of the extracted PARAFAC μ component with a peak at approximately 11.2 Hz. **Right:** Corresponding scalp spatial weights showing predominant activation over the left sensorimotor cortex (C3 region).

CNN, XAI:

After data preprocessing, the CNN model was trained. The training process is illustrated in Figure 4. The blue curve represents the training accuracy over time, while the orange curve indicates the validation accuracy (counted every 50th iteration). As described previously, one of the four exercises, together with its immediately following rest period, was always used as the validation dataset. Therefore, the training procedure was repeated four times, each time using a different exercise-rest pair for validation.

Across the validation runs, the average classification accuracy (rest vs exercise) was approximately 89% for the proposed CNN (AlexNet-inspired) with CWT-spectrogram input and 90% for the EEGNet-based model with raw EEG input. These results indicate comparable performance between the two approaches.

In the next step, the Grad-CAM, LIME and OS method was applied to the trained model using samples from the validation dataset. As an example, two spectrograms were selected. The first was selected from exercise 1 and the second from the rest period after it. In Figure 5, the left panel shows a 2-second spectrogram segment

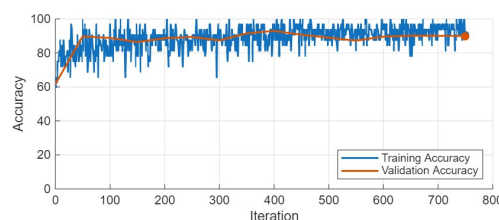


Figure 4: CNN model training process showing training and validation accuracy.

recorded during the exercise 1 and rest period. This samples were correctly classified as the exercise/rest class by the trained AlexNet inspired CNN model. The middle panel shows the Grad-CAM map, and the right panel shows the OS map. Both highlight the time-frequency regions that most influenced the model's decision.

Additional Grad-CAM, LIME and OS maps were generated for multiple spectrograms from all validation datasets. The highlighted regions were often consistent across different samples, with only minor variations. Similar patterns were observed for each validation fold, further supporting the stability of the learned features.

DISCUSSION

The proposed approach achieved classification performance comparable to an EEGNet-based model operating on raw EEG signals. However, a key advantage of the presented method lies in its interpretability. By combining time-frequency representations with XAI techniques, it is possible to directly assess whether the model relies on neurophysiologically meaningful features, providing additional insight beyond classification accuracy.

XAI visualizations provide insight into which spectral regions were most relevant for the CNN classification. In Figure 5 bottom-left panel, a spectrogram segment from the beginning of the exercise period was correctly classified as movement. Since this segment was extracted from the onset of the exercise block, it is reasonable to assume that the subject was actively performing the movement during this interval. Compared to the magnitude observed during the rest condition (top-left panel in Figure 5), this spectrogram shows a clear reduction in μ and β band activity after approximately 0.5s. This pattern is consistent with ERD in the sensorimotor rhythm during motor execution. The Grad-CAM map (Figure 5, bottom-middle) highlights broad regions within these frequency bands, indicating that suppression of μ and β activity played a substantial role in the movement classification. Moreover, occipital α activity is suppressed during the movement, reflecting visual processing of the mirrored hand, and likely contributed alongside μ and β ERD to the classification. These findings are consistent with the IRASA results, which showed reduced oscillatory amplitude in both the μ and occipital α bands during movement. Similarly, the PARAFAC decomposition identified a μ component with a peak frequency of 11.2 Hz, whose amplitude decreased during motor activ-

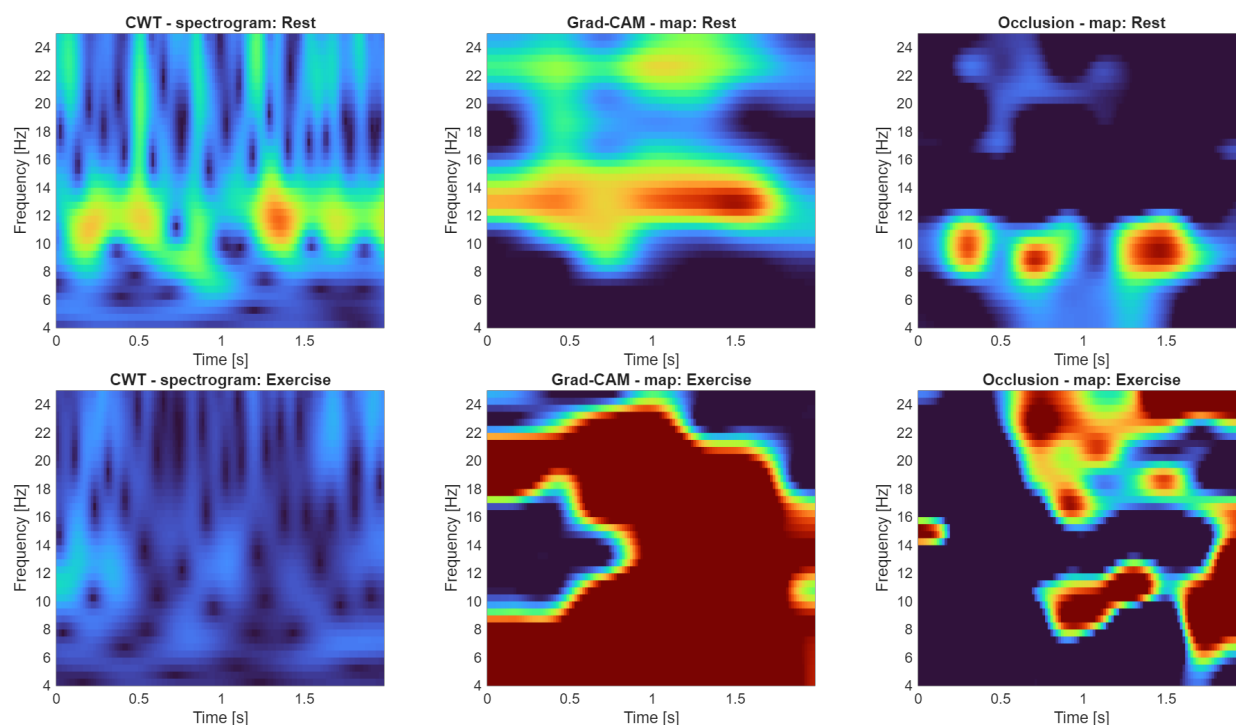


Figure 5: Correctly classified spectrograms. **Left panel:** Original input spectrogram labeled as Rest (top), Exercise (bottom) (color scale (CS): 0-15). **Middle panel:** Grad-CAM activation map highlighting the most influential regions (CS: 0-0.3). **Right panel:** Occlusion Sensitivity (OS) map showing how masking different time-frequency regions affects the model's output (CS: 0-0.01).

ity. Together, these complementary analyses support the physiological plausibility of the CNN decision process. In addition, lower activity in the θ band can be observed during movement compared to rest, and corresponding regions are partially highlighted by the Grad-CAM map. This may reflect reduced low-frequency activity associated with relaxed states or increased task engagement during movement, although this effect is less pronounced than the changes observed in the μ and β bands.

In contrast, Figure 5 (top-left panel) shows a rest-period spectrogram characterized by increased magnitude in the 9-15 Hz range. Based on the IRASA and PARAFAC findings, this activity likely reflects a combination of occipital (visual) α and sensorimotor μ rhythms. During the rest condition, the subject fixated on a visual cross and remained in a relaxed state, which may have contributed to enhanced posterior alpha activity. The corresponding XAI-maps (Figure 5, top-middle panel and top-right panel) emphasizes this frequency range, suggesting that elevated μ and α rhythms was a key discriminative feature for the rest class, with a slight increase in β activity also contributing to the classification.

Despite generally good classification performance, several misclassifications were observed. Analysis of incorrectly classified samples revealed that some rest-period spectrograms with weak sensorimotor μ , visual α , and β activity were classified as exercise. XAI maps visualizations indicated that regions with reduced magnitude in these frequency bands contributed to the incorrect predictions, suggesting that the model relied mainly on changes within the sensorimotor frequency range. One possible

explanation is variability in the subject's behavioral state during rest. Although instructed to fixate on a cross, fluctuations in attention or cognitive engagement may have modulated occipital α activity. In particular, α power over occipital regions has been shown to decrease during externally directed visual attention [16].

Another factor may be the structure of the exercise task itself. The subject performed repetitive movements separated by short pauses, and spectrogram segments extracted during these pauses could exhibit increased μ or α magnitude resembling rest-like patterns. An example is shown in Figure 6, where a spectrogram from the exercise period displays elevated magnitude between 8 and 15 Hz. This likely corresponds to a short pause between repetitions during which partial resynchronization of the sensorimotor rhythm occurred, potentially contributing to incorrect classifications.

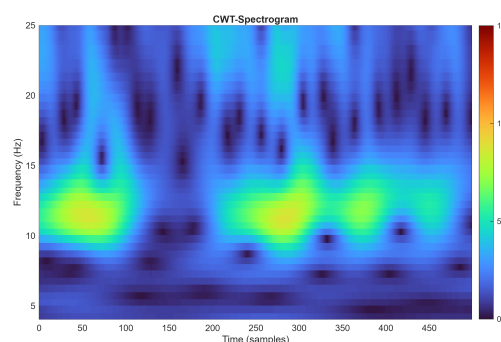


Figure 6: 2-second long spectrogram from exercise period.

CONCLUSION

In this study, we classified CWT-based spectrograms derived from rest and exercise tasks during MB training using a CNN. The model achieved good classification accuracy across multiple validation runs, demonstrating stable performance in distinguishing between movement and rest conditions. Notably, the obtained results were comparable to those achieved by an EEGNet-based model operating directly on raw EEG signals.

Importantly, the proposed approach provides an additional advantage in terms of interpretability. The use of time-frequency representations combined with XAI enables direct assessment of whether the model relies on physiologically meaningful features. The visualizations indicated that the model's decisions were largely consistent with established neurophysiological principles. Movement-related segments were primarily associated with reduced power in the visual α frequency band and with μ and β frequency bands, corresponding to ERD, while rest segments were characterized by increased activity in the α range. These findings were further supported by IRASA and PARAFAC analyses, which independently confirmed modulation of the sensorimotor and occipital rhythms.

Future work will focus on integrating EMG recordings to precisely determine the timing of motor execution. Accurate identification of movement onset and offset would allow removal of spectrogram segments containing transitional pauses between repetitions, improving classification performance and interpretability. In addition, future studies will explore CNN architectures with interpretable kernels, enabling analysis of learned filters in relation to neurophysiological patterns.

Furthermore, combining EMG timing information with EEG analysis would enable a more detailed investigation of ERD dynamics relative to movement onset, providing deeper insight into the relationship between EEG spectral characteristics and CNN-based classification.

FUNDING

This research was supported by the EU NextGenerationEU initiative through the Recovery and Resilience Plan for Slovakia under project No. 09I03-03-V04-00205 (Z.R.) and project No. 09I03-03-V04-00443 (R.R. and A.H.). Additional funding was provided by the European Commission through the HORIZON-MSCA 2022-DN, 101118964-DONUT project. This work was also supported by the Slovak Research and Development Agency under grant APVV-21-0105 (TInVR) and VEGA 2/0094/26.

REFERENCES

[1] Gao RX, Yan R. Wavelets. Springer US: Boston, MA (2011).

- [2] Pfurtscheller G, Lopes Da Silva F. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [3] O'Shea K, Nash R. An introduction to convolutional neural networks. Dec. 2, 2015. arXiv: 1511.08458 [cs]. Accessed: Feb. 26, 2026. [Online]. Available: <http://arxiv.org/abs/1511.08458>.
- [4] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017, 618–626.
- [5] Ribeiro MT, Singh S, Guestrin C. "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, 1135–1144.
- [6] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV). Springer, 2014, 818–833.
- [7] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.
- [8] Garry MI, Loftus A, Summers JJ. Mirror, mirror on the wall: Viewing a mirror reflection of unilateral hand movements facilitates ipsilateral m1 excitability. *Experimental Brain Research*. 2005;163(1):118–122.
- [9] Russell B, Han J. Jean morlet and the continuous wavelet transform. *CREWES Res. Rep*. 2016;28:115.
- [10] Alom MZ et al. The history began from alexnet: A comprehensive survey on deep learning approaches. 2018. arXiv: 1803.01164 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1803.01164>.
- [11] Schirrneister RT et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*. 2017;38(11):5391–5420.
- [12] Kingma DP, Ba J. Adam: A method for stochastic optimization. Jan. 30, 2017. arXiv: 1412.6980 [cs]. Accessed: Feb. 26, 2026. [Online]. Available: <http://arxiv.org/abs/1412.6980>.
- [13] Wen H, Liu Z. Separating fractal and oscillatory components in the power spectrum of neurophysiological signal. *Brain Topography*. 2016;29(1):13–26.
- [14] Rošťáková Z, Rosipal R, Seifpour S, Trejo LJ. A comparison of non-negative Tucker decomposition and parallel factor analysis for identification and measurement of human EEG rhythms. *Measurement Science Review*. 2020;20(3):126–138.
- [15] Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*. 1999;29(2):169–195.
- [16] Magosso E, colleagues. EEG alpha power is modulated by attentional changes during cognitive tasks and virtual reality immersion. *Frontiers in Human Neuroscience*. 2019;13:256.

Fast-Calibrated c-VEP BCI with Comfort-Optimized Stimulation for Robust Control on Screen and Immersive VR Using Semi-Dry EEG

Frédéric Dehais^{1,2}, Emilio Maggi¹, Jules Gomet¹, Juan Torre Tresols¹, Pietro Cimarosto¹, Sébastien Velut¹, Kalou Cabrera-Castillos¹

¹Federation ENAC-ISAE-ONERA, Toulouse, France

²School of Biomedical Engineering, Science and Health Systems Drexel University, PA, USA

E-mail: frederic.dehais@isae-supero.fr

ABSTRACT: Code-modulated visual evoked potentials (c-VEPs) are fast and reliable control signals for reactive brain-computer interfaces (rBCIs), but real-world use is often constrained by visually intrusive flicker and the need for wet EEG. We present a usability-oriented c-VEP rBCI that combines visually comfortable, low-salience textured stimulation, low-rate coding with asynchronous evidence accumulation to allow flexible decision times, and a semi-dry EEG setup to improve wearability while maintaining signal quality. We evaluated the system in two settings: a PC condition consisting of an 11-class keypad typing task on a standard computer screen, and a VR condition consisting of a simple shell game. Despite using only 33 s of calibration data, we achieved 100% accuracy in both conditions. Mean decoding time was approximately 4.3 s in the PC condition and 5.4 s in the VR condition, consistent with additional head motion and rendering constraints in VR. Participants also reported high visual comfort, low visual fatigue, and minimal peripheral distraction from the flicker, supporting the usability of the proposed stimulation. Overall, these results demonstrate fast-calibrated, robust, and visually comfortable c-VEP control that generalizes from conventional displays to immersive VR with a practical semi-dry EEG configuration.

INTRODUCTION

Code-modulated Visual Evoked Potentials (c-VEPs) are attracting increasing interest in the reactive brain-computer interface (rBCI) community because pseudo-random stimulation supports rapid code identification and can substantially reduce calibration compared with conventional steady-state visual evoked potentials (SSVEP) approaches [1, 2]. Despite this potential, most c-VEP rBCIs remain largely confined to controlled laboratory settings. Translation to practical use is still limited by user-centered constraints, notably dependence on wet EEG instrumentation and visually salient flicker that can induce foveal discomfort and distract in peripheral vision [3]. To improve practicality while preserving decoding reliability, we build on the Burst-code c-VEP paradigm,

which concentrates stimulation into brief, aperiodic flashes delivered at a low rate (approximately 3 flashes per second). By time-locking stimulus energy into brief two-frame flashes and spacing bursts sufficiently apart (around 300 ms with added jitter), burst stimulation yields roughly double the evoked-response amplitude compared with conventional faster m-sequence-based stimulation [4]. This provides an important step toward higher signal-to-noise ratio (SNR) outside lab-grade conditions. Burst coding also simplifies calibration to a binary discrimination (flash “1” vs. no-flash “0”) and dramatically reduces calibration time (e.g., 40 s in [5]). Online, decoding correlates a stream of flash probabilities with predefined Burst codes and commits only when accumulated evidence becomes consistent, enabling flexible (asynchronous) decision times while remaining robust to transient errors.

In parallel, to reduce visual load and peripheral distraction while amplifying neural responses, we adopted and adapted subtle, near-imperceptible oriented texture stimulation originally introduced in [6], in place of conventional high-contrast flicker. We hereafter refer to these textures as Stimuli for Augmented Response (StAR). StAR is designed to recruit early visual pathways, from contrast-sensitive retinal ganglion cells to orientation-selective cortical populations, while remaining perceptually unobtrusive. In our prior work, StAR produced substantially larger evoked responses (up to 3×) while remaining perceptually unobtrusive [3]. By combining stronger cortical drive with lower visual salience, textured stimulation supports accurate single-trial detection and minimizes distraction when multiple targets are presented simultaneously. When paired with Burst coding, this approach achieved high performance with low channel-count EEG using eight dry electrodes, reaching 93.6% accuracy in a five-class task and 93.3% mean accuracy in an 11-class keypad task [5]. We refer to this combination of textured stimulation and Burst coding as the StAR-Burst paradigm. Participants also reported improved comfort, describing the textured stimulation as far less noticeable in peripheral vision than conventional high-contrast flicker [3].

A complementary route to further improve SNR and ro-

bustness to motion is the use of semi-dry electrodes. While dry systems can shorten setup time and reduce preparation burden, they may compromise comfort, increase sensitivity to motion-related artifacts, and yield less stable signal quality in some users. Semi-dry EEG electrodes—typically based on saline- or hydrogel-assisted contact—offer a practical compromise by preserving rapid application while achieving electrode–skin impedances and signal quality closer to wet EEG, with reduced cleanup and improved wearability [7–9].

Beyond electrode and stimulus usability, there is growing interest in integrating BCIs into immersive virtual environments. Several studies report that BCI performance in VR can be comparable to, or sometimes better than, conventional screen-based setups across multiple paradigms, including P300 [10, 11], SSVEP [12, 13] and motor imagery [14]. In contrast, c-VEP has received comparatively little attention in VR despite its relevance for reactive interaction. To date, the closest demonstration is a mixed-reality c-VEP speller evaluated in an HMD under largely static viewing conditions, where a 2D interface is overlaid on a controlled passthrough background and participants remain seated to limit motion artifacts [15]. This work constitutes an important first step toward HMD-based c-VEP interaction by addressing display optics and timing constraints.

In the present study, we extend this emerging line of work by considering a more embodied VR scenario. Specifically, our stimuli are rendered as an overlay onto 3D objects and remain stable in the participant’s egocentric view (i.e., updated with head motion), while the task itself requires head movements to select targets. This setting introduces additional challenges for c-VEP decoding, including viewpoint changes, and motion-induced artifacts. Moreover, our prior evaluations of the StAR–Burst paradigm were conducted in largely static, screen-based settings with limited body motion [3]. Here, we therefore test whether fast-calibrated c-VEP decoding with comfort-optimized stimulation can remain reliable under the constraints of immersive interaction.

The objectives of this study are therefore twofold: (i) to replicate the 11-class keypad StAR–Burst benchmark in a controlled setting using semi-dry EEG to quantify potential gains in signal quality and accuracy; and (ii) to implement and evaluate the same decoding principles in a simple VR “shell game” task, explicitly testing robustness under realistic head movements while maintaining usability and visual comfort, again using semi-dry EEG.

MATERIAL AND METHOD

Participants: The study was conducted with 12 healthy participants (2 women, mean age = 32 ± 8 years), primarily staff and students at ISAE-Supaero. The research received ethical approval from the University of Toulouse (CER approval number 2025_1012) and complied with the Declaration of Helsinki. Participants provided written informed consent prior to the experiment. Exclusion

criteria included neurological disorders or psychoactive medication use at the time of the study. All participants had normal or corrected-to-normal vision.

Common Setup: For both the PC (Keypad) and VR (shell-game) tasks, each target was assigned a unique Burst code [4]. Codes consisted of binary sequences with low mutual correlation, and were additionally constrained to limit ERP overlap by enforcing a minimum inter-flash interval of 300 ms plus a jitter between consecutive “1” events within a code.

Instead of conventional high-contrast flicker (e.g., alternating colors) or pattern-reversal checkerboards, which can induce visual discomfort, we used texture-based stimuli composed of small, randomly oriented Ricker wavelet patches (see [3]). Stimuli were presented with a moderate modulation depth of 70%, reducing luminance fluctuations to improve visual comfort while preserving classification performance [4, 16]. This design builds on prior work showing that such textured stimulation can maintain high accuracy—including with low channel-count, non-wet EEG setups—while improving subjective comfort [3]. EEG was recorded using an 8-channel semi-dry system (MBrainTrain) at 500 Hz. Electrodes were placed over occipital and parieto-occipital regions (PO7, O1, Oz, O2, PO8, PO3, POz, PO4) to capture visually evoked potentials (VEPs).

Keypad PC-based BCI:

The PC BCI interface consisted of an 11-class keypad inspired by traditional mobile-phone keypad layouts, with selectable keys for digits 0–9 and the ‘#’ character. Participants were instructed to type predefined four-character sequences by overtly attending to the desired key. The current target sequence was displayed in the bottom-left corner of the screen. After each selection, the chosen key was highlighted in green if correct or red if incorrect, providing immediate performance feedback (click here to see the video showcasing the PC condition). Stimuli were presented on a 24.5” LCD monitor (Iiyama Gold Phoenix G-Master GB2590HSU-B1; 1920×1080; 400cd/m²) at a refresh rate of 60 Hz. The interface also provided *pre-decision visual feedback*: the currently leading candidate was progressively enlarged as the decoder’s confidence increased. This cue externalized the evolving decision state before commitment and, consistent with our prior findings, is expected to improve the user’s perceived sense of control [5].

Shell-game VR BCI:

The VR task was implemented in Unreal Engine¹ using Blueprints to control game logic, animate objects, and transmit flash onsets to the decoding pipeline via LabStreamingLayer (LSL). To reduce unnecessary stimulation and improve ecological validity, stimulation was automatically suspended for objects outside the participant’s current field of view (i.e., outside the view frustum or fully occluded). Although not essential for the present three-target task, this mechanism is intended to support future scenarios with more potential targets than avail-

¹<https://www.unrealengine.com>

able codes by temporarily deactivating hidden items and reallocating their codes to visible targets.

Participants performed a simplified shell game. Three cubes (“cups”) were aligned horizontally in front of the participant, and a sphere (the “ball”) was briefly revealed before being hidden under one cube. The three cubes then followed randomized shuffling trajectories and stopped. During the selection phase, the textured stimulus was rendered as a dynamic material mapped onto each moving cube (i.e., the texture remained locked to the object as the viewpoint changed with head motion) and flashed according to the cube’s assigned Burst code. Once the BCI reached a decision, the ball location was revealed and the selected cube turned green if correct and red otherwise (click here to see the video showcasing the VR condition).

The task was performed using a Meta Quest 3 headset fitted with the Elite head strap. This configuration ensured stable headset placement while maintaining clearance over occipital and parieto-occipital electrodes to limit pressure on the sensors and preserve comfort. The scene geometry was set such that the lateral cubes required head rotation to inspect, ensuring that participants actively moved their head to track targets during the task. *Experimental protocol:* Figure 1 shows a graphical representation of the experimental protocol. Participants were seated comfortably and provided written informed consent prior to the experiment. Then the EEG was set up on their scalp and signal was checked.

For the PC-based keypad experiment, each session began with a passive calibration phase used to train the classification pipeline to detect VEP responses to flash events. Participants fixated a central circle while Burst flickers were presented in five blocks, each comprising three code presentations (2.2 s per presentation). Including cueing and rest periods, calibration lasted 40 s in total and yielded 33 s of usable calibration data. Participants then completed an online typing task using the BCI keypad interface that consists of two sets of five 4-number sequences with a pause between the two sets (a total of 40 targets).

After this experimental condition, participants provided subjective ratings using three 10-point visual semantic differential scales assessing peripheral distraction, visual comfort, and visual fatigue. Furthermore, the participants were also asked to provide rating for a CVS questionnaire, to make sure that the task does not induce adverse effects on participants’ vision.

Participants next performed the VR experiment. It starts with a calibration phase where participants had to focus on a single cube on which the stimuli texture is overlaid and flickered with a random burst code. The 43 s calibration phase consists of 5 blocks composed of 3 repetitions of a 2.2 s code and a 2 s pause after, generating 33 s of usable calibration data. Then, participants completed 10 trials of the “shell game”, trying to guess under which cube the ball was hidden, by focusing on the cube they believed the ball was under.

Finally, as for the previous condition, participants were required to provide the subjective rating to the same questionnaires, visual comfort and CVS-Q.

Decoding Pipeline Implementation:

The online BCI decoding pipeline for the PC and VR conditions were the same and developed in *Timeflux* [17], an open-source Python framework for real-time acquisition, processing, and classification of time-series data. The user interface was implemented as a web application (HTML/CSS/JavaScript) running in a standard browser and connected to the *Timeflux* processing graph for the keypad and in Unreal Engine for the VR. Signal processing and classification were implemented in Python using *imblearn-learn* and *pyriemann*. The overall software architecture follows our previous implementation [3], with minor adaptations to support the present PC and VR configurations.

EEG streams were linked to *Timeflux* via Lab-StreamingLayer (LSL), then preprocessed using an IIR band-pass filter (1–25 Hz) and a band-stop filter (48–52 Hz) to attenuate slow drifts and high-frequency noise. The filtered signal was segmented into 300 ms epochs, time-locked to the display refresh cycle. Epoch timestamps were additionally shifted to compensate for an estimated 80 ms end-to-end latency in stimulus delivery and acquisition.

For each epoch, spatial information was summarized by estimating a covariance matrix across channels [18]. Prior to covariance estimation, stimulus-locked activity was enhanced and dimensionality reduced using xDAWN spatial filtering [19], retaining four spatial filters. Feature extraction was implemented with *pyriemann* (*XdawnCovariances*), which computes Riemannian “super-epochs” by concatenating each spatially filtered epoch with the corresponding average evoked response. The resulting covariance matrices were mapped to the tangent space of the Riemannian manifold using the mean covariance as the reference point. Tangent-space features were then downsampled and classified with a linear discriminant analysis (LDA) model, yielding an epoch-wise probability that a burst-evoked response was present. Because the classifier operates at ~60 Hz, a single evoked response can be detected over multiple consecutive frames. To prevent duplicate detections, probabilities were binarized (threshold = 0.5) and a truncation rule was applied: for any run of consecutive detections, only the first frame was retained.

Calibration was performed to train a binary classifier to discriminate stimulus-locked epochs elicited by a flash (1”) versus a no-flash event (0”). The calibration session lasted 43 s and yielded 33 s of usable EEG after excluding non-stationary segments. EEG was segmented into stimulus-locked epochs; epochs following flash onsets were labeled as class 1, whereas epochs following no-flash events were labeled as class 0. During online operation, the trained classifier produced, for each incoming epoch, a probability of class 1 (flash) versus class 0 (no-flash). Multi-class selection was then achieved by

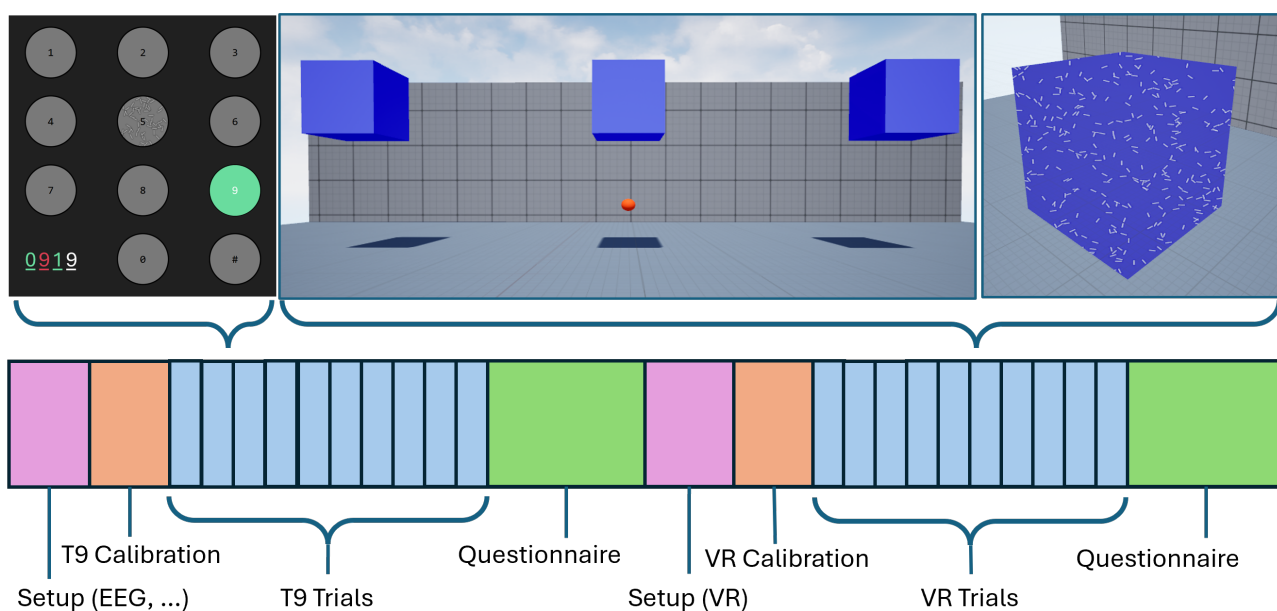


Figure 1: The top row shows the visual setup for each condition: the keypad experiment on the left, with the target sequence displayed in the lower-left corner; the VR experiment in the middle, featuring the three cubes and the ball; and, on the right, an example of a cube overlaid with our textured stimuli. Videos demonstrating the BCI in the PC condition are available here: [PC condition video](#), and in the VR condition here: [VR condition video](#).

code reconstruction: the recent stream of flash probabilities was continuously matched to each target's predefined Burst code. At each video frame, we computed the correlation between the probability stream and each class-specific code, and selected the best-matching candidate. To convert these continuously updated correlations into a robust decision with flexible timing, we used a consecutive-winner strategy with a momentum-decay dynamic [20]. The leading candidate accumulated momentum with an increment that increased as it remained in the lead, while all other candidates decayed by the same increment (bounded by a minimum floor). A decision was emitted when any class exceeded a predefined momentum threshold (see [5, 20] for implementation details).

BCI Metrics: Objective metrics included typing accuracy and decoding time. After each typing task, participants completed a short subjective questionnaire composed of three 10-point semantic differential scales, yielding two questionnaires per participant. The scales assessed perceived visual comfort and visual fatigue, as well as peripheral transparency (i.e., how distractive neighbouring flickers were in peripheral vision).

RESULTS

In both environments, online control reached 100% accuracy. In the screen-based keypad task, mean decoding time was 4.4 s (SD = 2.3 s), whereas in the VR shell-game task it increased to 5.4 s (SD = 0.9 s; Fig. 3). A one-tailed paired t-test indicated a significant increase in decoding time in VR compared with the PC condition ($t = -1.334$, $p = 0.05$, Cohen's $d = -0.536$). Figure 2 shows the ERPs evoked by the flashes during the calibration phase.

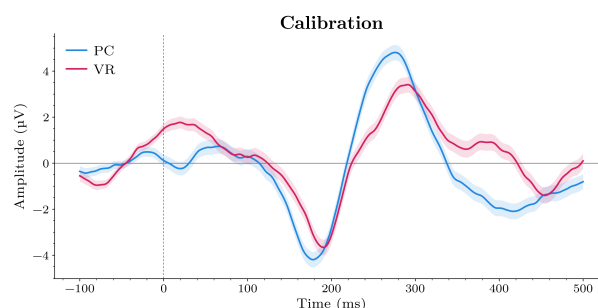


Figure 2: Mean evoked responses to flashes during calibration in the PC (blue) and VR (red) conditions. Shaded regions indicate the standard error.

Subjective ratings remained consistently favorable across conditions. On the PC task, participants reported high comfort (8.2/10, SD = 2.0), good peripheral transparency (7.1/10, SD = 2.5), and low mental fatigue (7.2/10, SD = 3.5). In VR, ratings were slightly lower but comparable (comfort 7.8/10, SD = 1.5; transparency 6.5/10, SD = 2.4; fatigue 6.8/10, SD = 3.4). Paired comparisons did not reveal significant differences between conditions for comfort ($t = 0.95$, $p = 0.18$, $d = 0.27$), mental fatigue ($t = 0.80$, $p = 0.22$, $d = 0.23$), or peripheral distraction ($t = 1.64$, $p = 0.076$, $d = 0.53$).

DISCUSSION

The present study had two objectives: first, to replicate our screen-based 11-class keypad c-VEP benchmark with a *semi-dry* EEG setup in order to quantify potential gains in robustness and performance; and second, to assess whether the same decoding principles generalize to an *immersive* VR task involving embodied interaction (shell

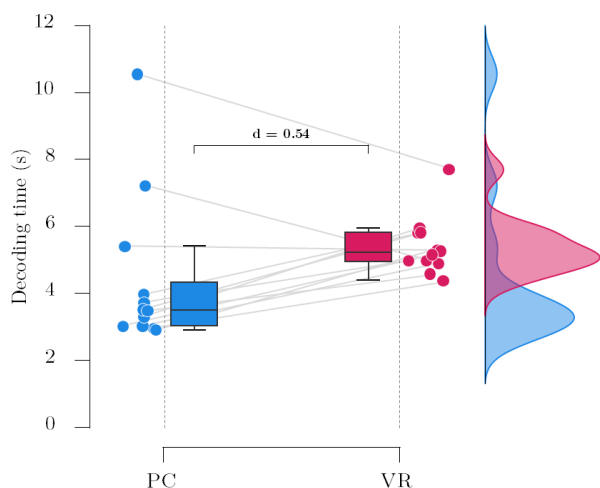


Figure 3: Mean decoding time in s for the PC (blue) and VR (red) experiment.

game) and concomitant motion-related artifacts.

Objective performance and impact of semi-dry EEG: Across both environments, online control reached 100% accuracy after only 33s of calibration data, demonstrating that reliable c-VEP control is achievable with a practical, fast-to-deploy semi-dry EEG configuration. In the screen-based keypad task, this performance exceeds our previous results obtained with a closely matched paradigm using *dry* electrodes (93.3% mean accuracy in [5]). A likely contributor is the higher and more stable signal quality provided by semi-dry contact. Specifically, semi-dry recordings showed larger stimulus-locked deflections (approximately $-4.2\mu\text{V}$ followed by $+4.8\mu\text{V}$) compared with the dry system (about $-3\mu\text{V}$ followed by $+1\mu\text{V}$ in [3]), increasing the discriminability of the elementary flash/no-flash (0/1) responses. This improvement is particularly consequential for the present decoding strategy: because class decisions are reconstructed by matching the inferred 0/1 stream to Burst codes, even modest gains in single-epoch 0/1 classification can propagate into higher reliability in multi-class code identification.

In the immersive VR shell-game task, the system likewise achieved 100% accuracy, supporting the feasibility of visually comfortable c-VEP control in an HMD. However, mean decoding time were statistically increased relative to the screen-based keypad task (5.4 s vs. 4.3 s), despite the VR task involving only three possible targets. This slower decision-making suggests that the effective evidence accumulation rate was reduced in VR. Two factors likely contributed. First, the amplitude of the evoked responses appeared smaller in VR than on the monitor (see figure 2), which reduces the reliability of the 0/1 stream and therefore requires longer accumulation to reach decision thresholds. Second, the headset may have introduced additional constraints on EEG acquisition: pressure from the HMD straps and repeated left–right head movements during target tracking can perturb electrode contact and increase motion-related artifacts affecting on-

line Burst code reconstruction. These integration challenges are consistent with broader observations in VR–EEG systems, where mechanical constraints and motion can impact signal quality and electrode placement [21]. Future work should therefore prioritize tighter mechanical integration of the HMD and EEG (e.g., sensor placement and cabling compatible with the headset strap geometry), as well as motion-aware artifact mitigation during decoding.

Subjective comfort, transparency, and VR-specific factors:

Subjective ratings were overall consistent with our earlier reports that textured, low-salience stimulation improves perceived comfort and reduces peripheral distraction compared with conventional high-contrast flicker [3, 5]. Although mean ratings tended to be slightly lower in VR than on the monitor, paired statistical comparisons did not reveal significant differences between conditions for comfort, mental fatigue, or peripheral transparency. This pattern suggests that the proposed stimulation remains well tolerated when transferred from a standard display to an HMD. At the same time, the small numerical trend toward lower ratings in VR remains plausible given that near-eye displays can impose additional oculomotor and visual demands, which have been associated with digital eye strain and discomfort in VR [22, 23]. Together, these findings indicate that comfort-optimized stimulation generalizes well to immersive VR, while leaving room for further refinements of rendering parameters (e.g., luminance, contrast, texture amplitude depth) and headset fit for longer sessions.

Limitations and future directions: This study has several limitations. First, the sample size was modest; we are currently extending recruitment to confirm the present findings in a larger cohort. Second, although online accuracy was perfect in both conditions, decoding in VR was slower than on the monitor, indicating that responsiveness can still be improved. Promising directions include adaptive decision thresholds driven by real-time signal quality, explicit monitoring and attenuation of motion/EMG artifacts, and improved mechanical integration of semi-dry electrodes with the HMD to minimize pressure-induced contact changes and motion-related impedance fluctuations [21]. Finally, our evidence-accumulation decoder—which correlates a stream of flash probabilities with predefined Burst codes and commits only after sufficient consistent evidence has accumulated—likely contributes to the high accuracy observed here and may also enable *idle-state gating*. When the user is not attending to any target, the incoming evidence should be dominated by “0”-like responses, yielding weak code correlations that remain below the decision threshold. However, this capability must be explicitly demonstrated and quantified. A key next step is therefore to introduce non-control (rest) periods and measure false activation rates, time-to-activation, and overall stability under realistic use conditions.

CONCLUSION

We demonstrated a fast-calibrated (33 s), visually comfortable c-VEP rBCI that achieved 100% online accuracy on a screen-based 11-class keypad task and in an immersive VR shell-game task using a practical *semi-dry* EEG system. Compared with our earlier dry-electrode implementation, semi-dry acquisition likely increased evoked-response amplitude, yielding more reliable code reconstruction leading to 100% accuracy. While VR control remained perfectly accurate, decisions were slower, consistent with motion/headset-induced constraints. Overall, these findings support the feasibility of moving comfort-optimized c-VEP interaction beyond laboratory conditions and into immersive VR, while highlighting the importance of improved EEG-HMD integration for faster, more robust embodied interaction.

REFERENCES

- [1] Martínez-Cagigal V, Thielen J, Santamaria-Vazquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): A literature review. *Journal of Neural Engineering*. 2021;18(6).
- [2] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. *Journal of Neural Engineering*. 2021;18.
- [3] Dehais F, Cabrera Castillos K, Ladouce S, Clisson P. Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry eeg code-vep bci. *Journal of neural engineering*. 2024;21(6):066023.
- [4] Cabrera Castillos K, Ladouce S, Darmet L, Dehais F. Burst c-VEP Based BCI: Optimizing stimulus design for enhanced classification with minimal calibration data and improved user experience. *NeuroImage*. 2023;284:120446.
- [5] Gomel J, Tresols JT, Cimarosto P, Castillos KC, Dehais F. Pre-decision feedback in code-modulated visual evoked potentials brain-computer interface for an 11-class keypad typing task. In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2025.
- [6] Ployart Guillaume ZRSNKS. Visual brain computer interface. number 11684300. 2022.
- [7] Li G, Wang S, Duan YY. Towards gel-free electrodes: A systematic study of electrode-skin impedance. *Sensors and Actuators B: Chemical*. 2017;241:1244–1255.
- [8] Li GL, Wu JT, Xia YH, He QG, Jin HG. Review of semi-dry electrodes for eeg recording. *Journal of Neural Engineering*. 2020;17(5):051004.
- [9] Liu J et al. Ten-hour stable noninvasive brain-computer interface realized by semidry hydrogel-based electrodes. *Research*. 2022;2022:9830457.
- [10] Käthner I, Kübler A, Halder S. Rapid p300 brain-computer interface communication with a head-mounted display. *Frontiers in neuroscience*. 2015;9:207.
- [11] Cattani GH, Andreev A, Mendoza C, Congedo M. A comparison of mobile vr display running on an ordinary smartphone with standard pc display for p300-bci stimulus presentation. *IEEE Transactions on Games*. 2019;13(1):68–77.
- [12] Koo B, Lee HG, Nam Y, Choi S. Immersive bci with ssvep in vr head-mounted display. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). 2015, 1103–1106.
- [13] Liu H et al. A comparative study of stereo-dependent ssvep targets and their impact on vr-bci performance. *Frontiers in Neuroscience*. 2024;18:1367932.
- [14] Škola F, Liarokapis F. Embodied vr environment facilitates motor imagery brain-computer interface training. *Computers & Graphics*. 2018;75:59–71.
- [15] Moreno-Calderón S, Martínez-Cagigal V, Martín-Fernández A, Santamaría-Vázquez E, Hornero R. Toward the integration of mixed reality and brain-computer interfaces based on code-modulated visual evoked potentials. *Biocybernetics and Biomedical Engineering*. 2025;45(3):528–538.
- [16] Ladouce S, Darmet L, Torre Tresols JJ, Velut S, Ferraro G, Dehais F. Improving user experience of SSVEP BCI through low amplitude depth and high frequency stimuli design. *Scientific Reports*. 2022;12(8865).
- [17] Clisson P, Bertrand-Lalo R, Congedo M, Victor-Thomas G, Chatel-Goldman J. Timeflux: An open-source framework for the acquisition and near real-time processing of signal streams. In: BCI 2019-8th International Brain-Computer Interface Conference. 2019.
- [18] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*. 2012;59(4):920–928.
- [19] Rivet B, Souiloumiac A, Attina V, Gibert G. xDAWN Algorithm to Enhance Evoked Potentials: Application to Brain-Computer Interface. *IEEE Transactions on Biomedical Engineering*. 2009;56(8):2035–2043.
- [20] Cimarosto P, Velut S, Castillos-Cabrera K, Tresols JT, Roy RN, Dehais F. Near-invisible c-vep-based passive bci for mental workload monitoring. *Journal of Neural Engineering*. 2026.
- [21] Li H, Shin H, Sentis L, Siu KC, Millán JdR, Lu N. Combining vr with electroencephalography as a frontier of brain-computer interfaces. *Device*. 2024;2:100425.
- [22] Hirzle T et al. Understanding, addressing, and analysing digital eye strain in virtual reality head-mounted displays. *ACM Transactions on Computer-Human Interaction*. 2022;29(4):1–80.
- [23] Cossio S et al. Cybersickness and discomfort from head-mounted displays delivering fully immersive virtual reality: A systematic review. *Nurse Education in Practice*. 2025;85:104376.

Towards Reliable Out-of-Lab pBCIs: Riemannian Geometry Framework for Mitigating Covariate Shift in Real-World Passive Brain-Computer Interfaces

D. Germano¹, V. Ronca^{1,3}, R. Capotorto², S. Ercoli¹, G. Borghini³, G. Di Flumeri³, P. Aricò¹

¹ Department of Computer, Control, and Management Engineering, Sapienza University of Rome, 00185 Rome, Italy

² Department of Anatomical, Histological, Forensic and Orthopaedic Sciences, Sapienza University of Rome, 00185 Rome, Italy

³ Department of Molecular Medicine, Sapienza University of Rome, 00185 Rome, Italy

E-mail: daniele.germano@uniroma1.it

ABSTRACT: Passive brain-computer interfaces (pBCIs) are promising tools for real-world, out-of-lab monitoring of cognitive states (e.g. mental workload). However, their reliability (e.g. classification accuracy) is often compromised by covariate shift caused by changes in headset positioning among different sessions. This study introduces a Riemannian-based correction method designed to mitigate the negative effect of the shift. We evaluated the framework by training a random forest classifier on a source domain to classify two mental workload levels and then tested on a target domain (representing a channels shift) both before and after applying the transformation. The validation involved three levels of analysis: (i) classification performance via area under curve (AUC) analysis, (ii) statistical assessment of manifold alignment through Riemannian distance, and (iii) correlation between geometric correction and AUC recovery. The application of the correction significantly improved workload classification, significantly overcoming AUC scores compared to the uncorrected target data. Statistical analysis revealed a significant reduction in Riemannian distance between the source and target manifolds after correction ($p = 2.36e^{-5}$). Furthermore, correlation analysis showed a strong positive relationship between geometric alignment and performance gain (Pearson correlation $R = 0.628$).

INTRODUCTION

The increasing need of interaction between humans and machines in modern industrial environments, often termed Industry 5.0, necessitates robust and seamless interfaces for monitoring operator state and optimizing performance [1]. Passive brain-computer interfaces (pBCIs) offer a promising, non-invasive solution by continuously assessing cognitive and emotional parameters in real-time [2]. Among the most relevant mental states in real-world applications, particularly in safety-critical environments, is mental workload. Elevated workload levels are strongly associated with performance decrements and increased error rates [3], [4]. While EEG signals have demonstrated effectiveness in workload assessment, particularly through analysis of theta and alpha band activity

in frontal and parietal regions [5], translating pBCI systems from controlled laboratory settings to real-world operational environments presents significant challenges. Unlike the highly controlled conditions of the lab settings, real-world deployment introduces substantial variability in EEG signal characteristics. Factors such as electrode impedance and, most notably, inconsistent sensor positioning on the scalp (i.e. because of changing in headset placement) can significantly impact signal amplitude and spectral features [6]. Maintaining the standardized 10-20 EEG placement protocol is often impractical when users prioritize convenience and rapid deployment, leading to a phenomenon known as covariate shift, where changes in data distribution degrade model performance over time [7]. Although extensively investigated in the broader field of artificial intelligence, covariate shift remains a relatively underexplored problem within the pBCI domain. Numerous studies have demonstrated how the non-stationarity of EEG signals can lead to variations in the distribution of data collected in different recording sessions [8], [9]. Some solutions proposed in the literature require the labeling of new data before starting the classification process [10], [11]. Deep adversarial frameworks offer optimal within- and cross-subject correction accuracy, often reducing calibration to just a few target trials. However, their disadvantages include high computational and implementation complexity and low interpretability, since these models learn complex latent representations that make it difficult to understand which signal features drive the predictions, for example which EEG channels or frequency bands influence the decision. [12], [13], [14]. To overcome these limitations, Riemannian geometry has emerged as a robust framework for BCI. Riemannian geometry methods have been successfully applied to compensate for transfer learning and cross-subject in active BCI [15], [16], [17]. These approaches typically operate on the manifold of symmetric positive definite (SPD) covariance matrices to find invariant representations. This study addresses the challenge of covariate shift resulting from variations in EEG headset positioning among different recording sessions, a common occurrence when users take off and put on again the device. Building upon previous work focused on

detecting such shift [18], we now propose a method for its correction.

Our approach leverages a normalization function applied to EEG features associate to workload, utilizing pre-recorded “eyes open” (EO) data as a stable reference point to recalibrate the data distribution across sessions. This normalization strategy aims to minimize the impact of inconsistent headset placement without requiring labeled data or extensive model retraining, a significant advantage over existing adaptation techniques.

MATERIALS AND METHODS

This study investigates the challenge of covariate shift in electroencephalography (EEG) data arising from inconsistencies in headset positioning across recording sessions, a common issue in ecological passive BCI applications. This study extends a previous research [19], using the same dataset consisting of EEG recordings from 18 right-handed male participants (mean age 25 ± 3 years), recruited from the National University of Singapore (NUS) student and staff population. Participants were selected for their proficiency in video games, general good health, and lack of psychological or pharmaceutical conditions. The experiments were conducted following the principles outlined in the 1975 Declaration of Helsinki, as revised in 2022 [20]. Experiments were approved by the Ethical Committee of the Sapienza University of Rome. The experimental design included six recording sessions distributed over three days, two sessions per day (morning and afternoon), to systematically assess performance changes under controlled conditions. During each session, participants engaged in four runs of the Multi-Attribute Task Battery (MATB) [21], a benchmark task designed to modulate cognitive workload through varying task complexity. Specifically, participants completed two “easy” runs, maintaining a cursor center position with a joystick, and two “hard” runs, simultaneously managing the same cursor control task alongside additional sub-tasks intended to simulate a flight emergency scenario. Each run lasted 2.5 minutes and was preceded by a one-minute eyes-open (i.e. EO) baseline recording to establish resting-state EEG activity. To mitigate potential confounding factors, stimulus parameters were randomized across sessions and participants received comprehensive training to ensure consistent performance and minimize learning effects. This yielded a comprehensive dataset of twelve pairs of “easy-hard” condition recordings, representing the variation in workload across the three experimental days. EEG data was acquired using a Waveguard ANT Neuro amplifier with a sampling rate of 256 Hz from 64 Ag/AgCl electrodes positioned according to the standardized 10-10 international system [22], referenced to the right mastoid electrode and grounded to AFz. To mimic the channels configuration of more practical, low-channel-count wearable EEG systems (e.g. Mindtooth system [23]), data analysis focused on a subset of eight electrodes: five frontals (FPz, AF3, AF4, AF7, AF8) and three parietals (Pz, P3, P4). This selection allowed us to investigate the impact of positional

shift on positions relevant for workload evaluation. To simulate covariate shift stemming from variations in headset placement, we constructed a control dataset and a test dataset, introducing deliberate shifts in electrode selection. Due to the use of pre-recorded high-density EEG dataset, which precluded physical reposition of the headset, we simulated the shift by selecting an adjacent row of electrodes. Specifically, the channels configurations were as follows:

- Control (Source): FPz, AF3, AF4, AF7, AF8, P3, P4, PZ
- Test (Target): FZ, F3, F4, F7, F8, PO3, PO4, POz

Prior to subsequent analysis, the EEG signal underwent a pre-processing pipeline employing a fifth-order Butterworth band-pass filter (2–30 Hz) to remove frequency components not relevant for workload evaluation. Eye blink artifacts were detected and corrected using the o-Clean method [24]. Data segmentation into 1-second epochs was followed by a conservative artifact rejection threshold of $\pm 80 \mu\text{V}$, chosen to minimize data loss while ensuring signal quality. From these cleaned EEG data, we computed across frequency band only, the global field power (GFP) [25], for the theta and alpha frequency bands. GFP serves as a robust estimator of cortical synchronization, providing a time-domain representation of overall EEG activity within the specified frequency ranges. Individual alpha frequency (IAF) was determined from the baseline eyes-open recordings to ensure consistent band definition across participants [26]. To evaluate the mental states, we utilized a random forest classifier implemented in Python’s scikit-learn library [27], an ensemble learning technique that combines multiple decision trees to enhance predictive accuracy and minimize overfitting. The goal of this classifier was to distinguish between the “easy” and “hard” conditions based on the extracted GFP features. To comprehensively assess the impact of task difficulty, we focused our classification efforts on one mental state: workload. For the classification of workload, we leveraged the sensitivity of frontal and parietal cortical regions to cognitive demands [28], [29]. Specifically, features were extracted from the theta band activity recorded from the five frontal electrodes (FPz, AF3, AF4, AF7, AF8) and the alpha band activity from the three parietal electrodes (Pz, P3, P4). This selection was based on a substantial body of existing literature demonstrating a strong correlation between increased theta power in frontal regions and heightened cognitive effort [30], and alpha band modulation in parietal areas reflecting allocation of attentional resources.

To mitigate the covariate shift effect induced by headset repositioning, we employed a Riemannian geometry-based correction function. Unlike standard Riemannian applications that use the covariance matrix directly as the input for the classifier [31], [32], our approach applies a Riemannian transformation to the raw EEG time series. This function operates on the covariance matrices of the EEG signals, computed using all eight selected channels to maximize the information content and accurately capture the underlying neural dynamics. The transformation

function is defined as:

$$T_{mapping} = C_{Source}^{\frac{1}{2}} C_{Target}^{-\frac{1}{2}}$$

Where C_{Target} and C_{Source} represent the Riemannian mean of the covariance matrices for the target and source eyes-open sessions, respectively.

The underlying principle of the proposed approach is that, although the absolute amplitude of EEG signals may vary due to headset repositioning, the relative relationships between signals recorded from electrodes located in a consistent anatomical configuration are expected to remain relatively stable. To exploit this property, we adopt a Riemannian geometry-based framework operating on the covariance matrices of the EEG signals.

Specifically, covariance matrices are computed from 1-second EEG epochs and represented as points on the Riemannian manifold of symmetric positive definite matrices. This representation allows us to measure the distance between the covariance structure of the current session and that of a reference session recorded under a standardized condition. In this study, the reference is derived from an eyes-open resting-state recording. The core of our method consists of a whitening and re-coloring transformation. We first estimate the geometric mean of the covariance matrices from the EO baseline of both the source and target configurations. Then, a transformation matrix is derived to “whiten” the data and “re-color” it according to the source configuration’s statistics. Crucially, this transformation is applied directly to the EEG data of the task session. This ensures that the signal is realigned in the original sensor space before feature extraction and GFP calculation, allowing for a real-time implementation. The choice of eyes-open sessions is motivated by both methodological and practical considerations. First, these recordings do not require task execution or expert annotation, making them naturally unsupervised. This enables a simple recalibration procedure that can be autonomously performed by the user after headset repositioning. In practice, once the headset is removed and worn again, the user only needs to record a short eyes-open resting-state segment (e.g., approximately 30 seconds). The covariance matrices computed from this segment are then used to estimate the transformation matrix required to realign the current session with the reference distribution. To simulate a realistic scenario in which no relationship exists between source and target, the first 50% of the EEG recordings from the eyes-open session in the Source channel configuration was used to compute the covariance matrices C_{Source} , while the remaining 50% of the EEG recordings, acquired in the Target configuration, was used to compute the covariance matrices C_{Target} . To evaluate the effectiveness of the proposed Riemannian-based correction for covariate shift, we conducted three complementary analyses: (i) classification performance evaluation, (ii) Riemannian covariance alignment analysis, and (iii) the correlation between the two metrics.

Classification Performance

A random forest classifier was trained using, each time,

two paired sessions from the source configuration. The features used for training and testing were the GFP values for each configuration. For the testing phase, a separate session was selected to serve as the benchmark in three distinct scenarios:

- Source_test: Unshifted source GFP data.
- Target: GFP data affected by the covariate shift
- Debiased: The target GFP data after applying the Riemannian correction on EEG signal.

By using a session-wise data split, we ensured that the training and test sets were completely independent. Specifically, for each training phase, two recording sessions were used for training while a separate session was reserved exclusively for testing. Because no data from the test session was included during training, this procedure prevented any potential data leakage. This process was repeated across the three sessions in a cross-validation scheme, and the final performance was computed as the average AUC obtained from the three individual test sessions. To evaluate classification performance, we used the area under curve (AUC) score as the primary metric. Additionally, AUC was computed at different decimation time points (i.e. by averaging the model’s predicted values (0 or 1) over various time intervals, from 1 to 60.

At each decimation step, the number of correct assessments is determined by comparing the predicted values with the actual values, both obtained by applying a moving average with a window size of t and a unit step.

For each decimation value, considering all 18 subjects, we performed a one-side Wilcoxon signed-rank test to evaluate whether the AUC distribution after data transformation (Debiased) was statistically different compared to the AUC distribution obtained when using the untransformed test condition (target).

Riemannian covariance alignment

To quantify the alignment between source and target domains, we employed the Riemannian distance as a metric of similarity between covariance matrices. This analysis included all sessions across all subjects. We calculated the distance between source and target covariance matrices, before (Pre-Correction) and after (Post-Correction/Debiased) the transformation. The objective was to statistically determine whether the correction significantly reduced the distance between the two distributions, effectively aligning the target domain with the source manifold. To this end, a two-sided Wilcoxon signed-rank test was used to compare the pre- and post-correction distances. While the sample size ($n=144$) might support parametric approaches, the observed non-normal distribution of the differences made the Wilcoxon test a more robust and appropriate choice.

Correlation between Geometric Alignment and Classifier Gain

Finally, we investigated the relationship between the geometric correction and the resulting improvement in classification accuracy. For each subject, we calculated two metrics:

- $\Delta AUC = AUC_{Debiased} - AUC_{Target}$

- $\Delta D = D_{Target} - D_{Debiased}$

Where D_{Target} and $D_{Debiased}$ are the Riemann Distance between the covariance matrices of the target pre and post transformation.

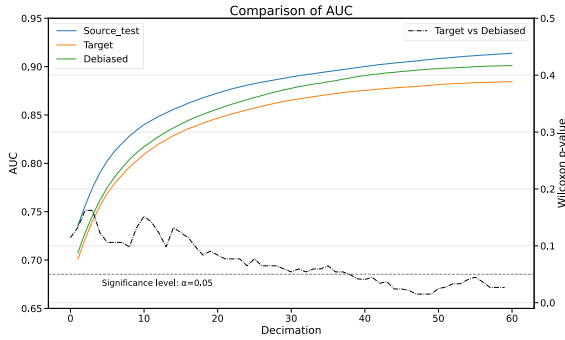


Fig. 1: Comparison of the AUC on the three test configuration: source, target and debiased.

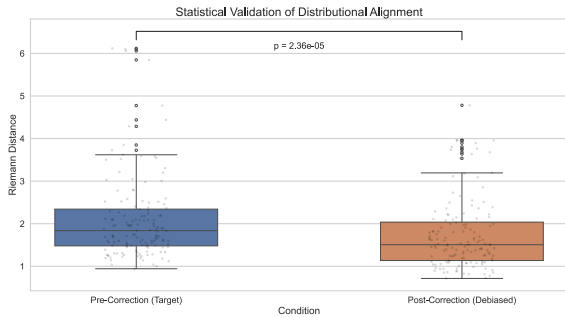


Fig 2: Distribution of Riemann distance between covariance matrix, for each session, of source and target configuration before and after the correction.

RESULTS

The effectiveness of the proposed debiasing transformation was evaluated by comparing the classification performance (AUC) across three conditions: source_test (i.e., test session of Source configuration), target (shifted positioning), and debiased (shifted positioning with Riemannian correction). As illustrated in Fig. 1, the target condition (orange line) exhibited a consistent performance drop compared to the source_test (blue line), confirming that electrode displacement induces a significant covariate shift that degrades the Random Forest classifier's reliability. However, the application of the Riemannian-based normalization (green line) successfully recovered a substantial portion of the lost performance. Notably, the Debiased AUC trend closely follows the source_test trajectory, consistently outperforming the uncorrected target data across all decimation levels. To statistically validate this improvement, a sliding Wilcoxon signed-rank test was performed between the target and debiased distributions (black dashed line). The geometric impact of the Riemannian alignment is summarized in

Fig. 2. The boxplot illustrates the distribution of Riemannian distances between the Source reference mean and the target covariance matrices, both before and after the correction.

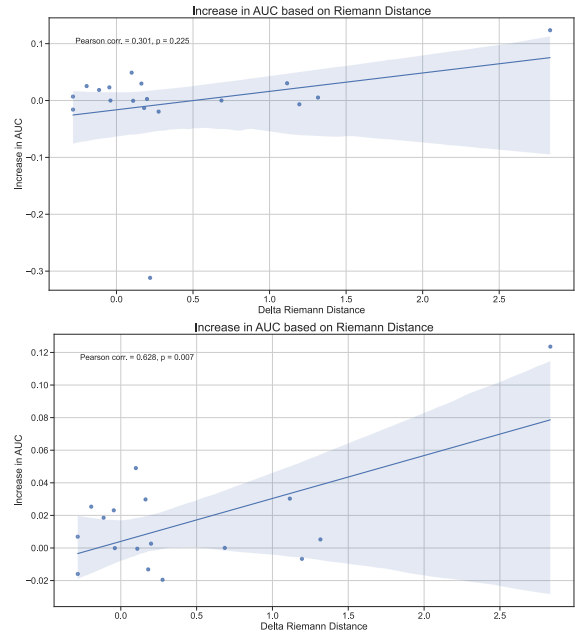


Fig 3: Relationship between Δ Riemann Distance and Δ AUC.

Fig. 3 shows the relationship between the reduction in distance (ΔD) and the improvement in AUC. These results are computed using the entire dataset (top) and after removing a single outlier (bottom). Most data points are located in the upper-right quadrant, and the regression line shows a positive trend. This indicates that subjects who experienced a larger reduction in Riemannian distance also tended to show a higher increase in classification performance. As summarized in Fig. 3, the initial correlation, including all subjects was weak and not statistically significant (Pearson = 0.301, $p = 0.224$). However, by excluding one outlier the correlation became significantly stronger. In the bottom plot of Fig. 3, the Pearson coefficient increases to 0.628 with a significant p-value of 0.007.

DISCUSSION

The results of this study demonstrate that Riemannian geometry could represent an effective way to handle covariate shift in passive BCI systems. The classification results show that shifting channels between tests creates a domain gap that lowers the AUC of the random forest model. However, by applying the debiasing transformation, we were able to move the Target data closer to the original Source distribution, resulting in a significant increase in AUC. Fig.1 reveals that the performance gain becomes statistically significant ($p < 0.05$) after less than 40 seconds, where the p-value drops below the threshold (indicated by the horizontal dashed line). These results

demonstrate that the proposed method effectively realigns the feature distributions on the manifold, mitigating the impact of inconsistent headset placement without the need for additional labelled training data. The statistical analysis of the Riemannian distances provides a clear explanation for this improvement. The significant decrease in distance post-correction ($p = 2.36e^{-5}$) proves that the algorithm successfully aligns the covariance matrices of the two domains. Essentially, the geometric noise caused by the channel shift is reduced, making the data more consistent for the classification model. The most interesting finding is the positive correlation between the Δ metrics. The fact that a larger reduction in Riemannian distance leads to a higher AUC gain suggests that this geometric distance is a reliable predictor of model performance. If we are able to minimize the distance between the two configurations (i.e. source and target) on the Riemannian manifold, we can directly improve the reliability of the BCI. The fact that the p-value dropped from 0.224 to a significant 0.007 proves that the observed trend is not a matter of chance. While the majority of subjects followed a clear linear trend, one subject deviated from this pattern, suggesting that additional factors may occasionally affect the correction. This study has some limitations, mainly due to the relatively small sample size and the use of a laboratory dataset, recorded under optimal conditions. Furthermore, the covariate shift was artificially created and not due to actual incorrect repositioning of the headset by the user. These design choices were intentional, aiming to avoid additional complexities and possible biases associated with EEG signals. However, this work serves as a preliminary step toward addressing covariate shift.

CONCLUSION

This study suggests that Riemannian-based alignment may represent a promising preprocessing strategy for improving the classification AUC of passive BCI applications. However, it should be noted that the analyses presented in this work were conducted in controlled laboratory conditions. Within this context, our findings indicate that the covariate shift induced by changes in electrode positioning can be partially mitigated using Riemannian geometry. This is particularly relevant for passive BCI systems, such as cognitive workload monitoring, where the absence of explicit user commands makes real-time calibration challenging. Unlike active BCI tasks, there adaptation often relies on supervised task-related data, our approach demonstrates that covariate shift can be addressed as a global deformation of signal space. Despite the limited sample size, the results provide statistical evidence supporting the hypothesis of this work. In particular, we observed a significant reduction in the Riemannian distance between configurations, which was associated with an improvement in workload decoding performance. The significant correlation between Δ Riemann Distance and Δ AUC further suggests that reducing the distance between source and target manifolds may contribute to improving cross-session stability. Most of the

18 participants showed improvements in classification performance after geometric correction. This highlights the effectiveness of using a neutral baseline, such as an “eyes-open” state, to define a transformation function that can be applied directly to the EEG time series in real-time. However, the presence of a case in which AUC decreased despite the geometric alignment suggests that additional factors may influence the effectiveness of the proposed approach. In some instances (i.e., 1 participant out of 18), the signal-to-noise ratio in the target configuration may be sufficiently degraded that covariance alignment alone cannot fully recover the lost information. Overall, the results indicate that the proposed method may represent a useful step toward improving the robustness of passive BCI systems under electrode displacement, offering a practical alternative to complex model retraining. Future work will focus on extending the analysis to larger datasets, developing a structured implementation of the proposed framework, and evaluating its performance on real-world datasets collected in operational environments.

ACKNOWLEDGEMENT

This work was funded by the “NeuroUX 5.0” project, within the program BRIC, funded by the National Institute for Insurance against Accidents at Work, CUP B83C24006240005; “TRUSTY” project (GA: 101114838) provided by the HORIZON -SESAR EU program; “TCI: Neurophysiological Framework for Predicting Team Dynamics in High-Responsibility Operational Environments -Team-Computer Interface” provided by the Sapienza University of Rome to Pietro Aricò (CUP B83C24007070005).

REFERENCES

- [1] P. Coelho *et al.*, “Industry 5.0: The Arising of a Concept,” *Proc. Comp. Sci.*, vol. 217, pp. 1137–1144. 2023, doi: 10.1016/j.procs.2022.12.312.
- [2] J. Yang *et al.*, “Human-machine interaction towards Industry 5.0: Human-centric smart manufacturing,” *Digital Engineering*, vol. 2, p. 100013, 2024, doi: 10.1016/j.dte.2024.100013.
- [3] P. Aricò *et al.*, “How Neurophysiological Measures Can be Used to Enhance the Evaluation of Remote Tower Solutions,” *fnhum*, vol. 13, 2019, doi: 10.3389/fnhum.2019.00303
- [4] J. Fan *et al.*, “The Impact of Workload and Fatigue on Performance”. 2017, p. 105. doi: 10.1007/978-3-319-61061-0_6.
- [5] B. Raufi *et al.*, “An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload,” *fninf*, vol. 16, 2022, doi: 10.3389/fninf.2022.861967.
- [6] C. L. Scrivener *et al.*, “Variability of EEG electrode positions and their underlying brain regions: visualizing gel artifacts from a simultaneous EEG-fMRI dataset,” *Brain and Behavior*, vol. 12, no. 2, p. e2476, 2022, doi: 10.1002/brb3.2476.
- [7] A. Kamrud *et al.*, “The Effects of Individual

- Differences, Non-Stationarity, and the Importance of Data Partitioning Decisions for Training and Testing of EEG Cross-Participant Models,” *Sensors*, vol. 21, no. 9, Art. no. 9, 2021, doi: 10.3390/s21093225.
- [8] H. Raza *et al.*, “Bagging Adversarial Neural Networks for Domain Adaptation in Non-Stationary EEG,” in *2019 IJCNN*, 2019, pp. 1–7. doi: 10.1109/IJCNN.2019.8852284.
- [9] H. Raza *et al.*, “Covariate shift-adaptation using a transductive learning model for handling non-stationarity in EEG based brain-computer interfaces,” in *2014 IEEE BIBM*, Nov. 2014, pp. 230–236. doi: 10.1109/BIBM.2014.6999160.
- [10] H. Raza *et al.*, “Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related EEG-based brain-computer interface,” *Neurocomp.*, vol. 343, pp. 154–166, 2019, doi: 10.1016/j.neucom.2018.04.087.
- [11] H. Raza *et al.*, “Adaptive learning with covariate shift-detection for motor imagery-based brain-computer interface,” *Soft Comp.*, vol. 20, no. 8, pp. 3085–3096, 2016, doi: 10.1007/s00500-015-1937-5.
- [12] K. Yin *et al.*, “GITGAN: Generative inter-subject transfer for EEG motor imagery analysis,” *Pattern Recognition*, vol. 146, p. 110015, 2024, doi: 10.1016/j.patcog.2023.110015.
- [13] Z. Liang *et al.*, “A novel deep transfer learning framework integrating general and domain-specific features for EEG-based brain-computer interface,” *Biomed. Signal Proc. Cont.*, vol. 95, p. 106311, 2024, doi: 10.1016/j.bspc.2024.106311.
- [14] H. Zhao *et al.*, “Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 535–545, 2021, doi: 10.1109/TNNLS.2020.3010780.
- [15] F. Yger *et al.*, “Riemannian Approaches in Brain-Computer Interfaces: A Review,” *IEEE Trans. Neu. Syst. Rehb. Eng.*, vol. 25, no. 10, pp. 1753–1762, 2017, doi: 10.1109/TNSRE.2016.2627016.
- [16] A. Bleuzé *et al.*, “Transfer Learning for the Riemannian Tangent Space: Applications to Brain-Computer Interfaces,” in *2021 ICEET*, Istanbul, Turkey: IEEE, 2021, pp. 1–6. doi: 10.1109/ICEET53442.2021.9659607.
- [17] J. Jin *et al.*, “RUNet: A Zero-Calibration Framework for Cross-Domain EEG Decoding via Riemannian and Unsupervised Representation Learning,” *IEEE Trans. on Biomed. Eng.*, pp. 1–17, 2026, doi: 10.1109/TBME.2026.3653024.
- [18] D. Germano *et al.*, “Unsupervised Detection of Covariate Shift Due to Changes in EEG Headset Position: Towards an Effective Out-of-Lab Use of Passive Brain-Computer Interface,” *Applied Sciences*, vol. 13, no. 23, p. 12800, 2023, doi: 10.3390/app132312800.
- [19] D. Germano *et al.*, “Towards the Correction of Covariate Shift in EEG-Based Passive Brain-Computer Interfaces for Out-of-Lab Applications,” *IEEE EMBC*, vol. 2025, pp. 1–7, 2025, doi: 10.1109/EMBC58623.2025.11252974.
- [20] “WMA - The World Medical Association-WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects.” <https://www.wma.net/policies-post/wma-declaration-of-helsinki>
- [21] J. R. Comstock *et al.*, “The multi-attribute task battery for human operator workload and strategic behavior research,” *NAS 1.15:104174*, 1992. <https://ntrs.nasa.gov/citations/19920007912>
- [22] V. Jurcak *et al.*, “10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems,” *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, 2007, doi: 10.1016/j.neuroimage.2006.09.024.
- [23] R. Vincenzo *et al.*, “Beyond the lab: real-world benchmarking of wearable EEGs for passive brain-computer interfaces,” *Brain Inf.*, vol. 13, no. 1, p. 3, 2025, doi: 10.1186/s40708-025-00290-x.
- [24] V. Ronca *et al.*, “o-CLEAN: a novel multi-stage algorithm for the ocular artifacts’ correction from EEG data in out-of-the-lab applications,” *JNE*, vol. 21, no. 5, p. 056023, 2024, doi: 10.1088/1741-2552/ad7b78.
- [25] W. Skrandies, “Global field power and topographic similarity,” *Brain Topogr.*, vol. 3, no. 1, pp. 137–141, 1990, doi: 10.1007/BF01128870.
- [26] W. Klimesch, “EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis,” *Brain Research Reviews*, vol. 29, no. 2, pp. 169–195, 1999, doi: 10.1016/S0165-0173(98)00056-3.
- [27] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [28] M. Qaremoammadlou *et al.*, “Investigating Brain Connectivity and Information Flow in Mental Workload Using EEG and fNIRS Integration,” 2025. doi: 10.48550/arXiv.2511.02457.
- [29] G. N. Dimitrakopoulos *et al.*, “Cognitive Reorganization Due to Mental Workload: A Functional Connectivity Analysis Based on Working Memory Paradigms,” *Appl. Sciences*, vol. 13, no. 4, p. 2129, 2023, doi: 10.3390/app13042129.
- [30] P. Aricò *et al.*, “Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment,” *fnhum*, vol. 10, 2016, 10.3389/fnhum.2016.00539
- [31] J. Ying *et al.*, “Riemannian geometry-based transfer learning for reducing training time in c-VEP BCIs,” *Sci Rep*, vol. 12, no. 1, p. 9818, 2022, doi: 10.1038/s41598-022-14026-y.
- [32] M. Congedo *et al.*, “A New Generation of Brain-Computer Interface Based on Riemannian Geometry,” 2026, doi: 10.48550/arXiv.1310.8115.

Decoding Spoken Syllables from EEG Data via Contrastive Embedding Alignment

T. Cortecchia¹, P. Volpato², S. Tortora^{2,3}, E. Menegatti², L. Tonin^{2,3}

¹Polytechnic University of Turin, Turin, Italy

²Department of Information Engineering University of Padova, Padova, Italy

³Padova Neuroscience Center, Padova, Italy

E-mail: tommaso.cortecchia@polito.it

ABSTRACT: We explore the potential of electroencephalography (EEG)-based brain-computer interfaces as an alternative communication system for people with speech impairments. Instead of considering single words or phonemes, we exploit the syllabic structure of the Italian language as the fundamental building block of speech. We collected data from 21 participants as they pronounced and syllabified a set of sentences. We then trained a neural network to learn a latent representation of EEG segments through contrastive alignment with embeddings extracted from a pre-trained automatic speech recognition model. We framed the evaluation as an information retrieval task, utilizing a top- k retrieval within the EEG latent space, comparing a global model to subject-specific models. We achieved an average Top-10 accuracy of 37.4% for the multi-subject model and of 38.2% for subject-specific models across a 192-class vocabulary. The proposed framework provides a scalable approach to non-invasive speech decoding, enabling greater expressiveness and speed compared to traditional solutions in assistive communication.

INTRODUCTION

Augmentative and alternative communication (AAC) devices complement or replace speech for individuals with severe speech disorders or paralysis; however, they are currently limited by slow conversational rates of around 12-18 words per minute (WPM), far below the average 125-185 WPM rate of natural speech [1]. Brain-computer interfaces (BCIs) for speech decoding are on a promising path to improve speed and accessibility of such systems. While invasive implants have recently enabled accurate, continuous decoding of attempted and silent speech [2-4], their high cost, the need for surgery and clinical support, and the risk of long-term data degradation hinder widespread adoption of these technologies [5].

BCIs based on electroencephalography (EEG) may offer a viable alternative for a low-cost, non-invasive solution suitable for everyday use. Although the research on this technology has seen a growing interest over the past few years, the past works have focused either on passive tasks

like listening or reading [6-9], lacking the active component of a BCI, or on the production of restricted sets of words or phonemes [10, 11] which do not allow for the composition of complex language. To truly enable fluent and naturalistic communication for the end-user, a system should offer a wide vocabulary (or a way to build that) and an intuitive mapping from the mental task to the decoded part of speech. A simple solution would just be collecting data during the pronunciation of sentences. Zhang et al. collected a large dataset of imagined speech for Chinese sentences [12]. Sato et al. investigated large-scale open-vocabulary decoding with 175+ hours of high-density EEG [13]. Craik et al. decoded continuous and discrete acoustic properties of the sound produced during aloud repetition of sentences [14]. With these methods, they are effectively collecting a large vocabulary, but it is difficult to align neural data to its linguistic counterpart, especially in the case of silent speech.

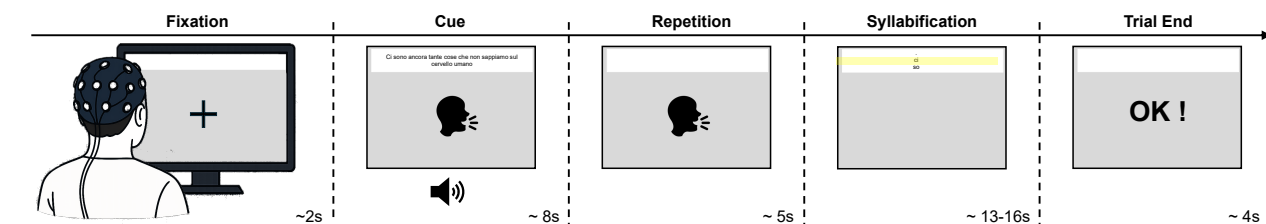
The words of the Italian language, like many other western languages, have the peculiarity of being divisible into smaller units, the syllables. Syllables are small sequences of phonemes that together form a single, distinct sound. Italian counts around 2,700 orthographical syllables (more than 3,700 if we consider phonetics) that contribute to forming more than 150,000 words [15].

In this work, we use the syllabic structure of the Italian language as a building block for a speech BCI. Specifically, we (1) introduce a new protocol that allows for the recording of a broad set of speech units during a natural task, and (2) establish a decoding performance baseline for overt speech decoding by applying state-of-the-art methodologies [9, 16] to the resulting dataset.

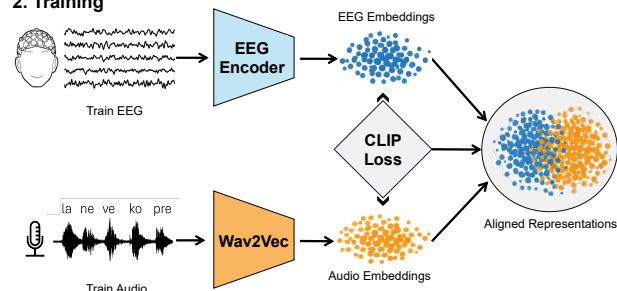
EXPERIMENTAL SETUP

This section introduces the experimental setup, including the recording protocol, the trial structure, and the method used to syllabify the sentences. Although the results we will present in the following sections relate exclusively to the decoding of overt (vocalized) speech, the protocol was designed with the requirements for recording imagined (silent) speech in mind as well. For this reason, we describe the procedures for both conditions here.

1. Data Collection



2. Training



3. Inference

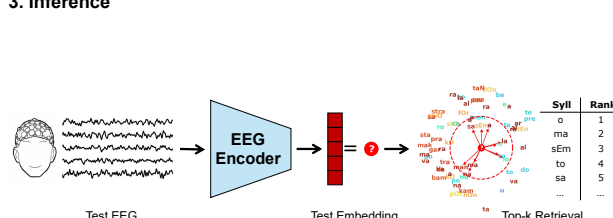


Figure 1: Overview of the proposed framework. (1) *Recording Protocol*: structure of a trial, stage duration is reported below each. Every run consists of 20 repetitions with different sentences. (2) *Cross-modal alignment*: the EEG–audio framework trained to align neural representations with acoustic embeddings using a contrastive objective. (3) *Syllable decoding*: at inference time, acoustic data is discarded, and the target syllable is predicted via a top- k retrieval across the EEG reference set.

When designing the protocol, we kept two key aspects in mind: (1) *Naturalness of the protocol*: Trial-based paradigms work well but are too slow and unnatural, making it difficult to collect a diverse set of classes in a short time. On the other hand, we needed a way to align EEG data with the expected class, especially in the case of silent speech; (2) *Target linguistic unit*: identifying a set of linguistic elements that did not have too many elements (such as words), which would make it impossible to build an effective vocabulary, but also not too elementary (such as letters), because we would come across the same speed limitations as current AAC systems. The perfect trade-off between sample size and production speed can come from dividing words into syllables. Each word allows for the collection of multiple syllables, and the same syllable appears in several different words. Furthermore, by pronouncing segmented words at a measured pace, it is possible to obtain aligned data even in the absence of stimuli. At the same time, syllabification of a sentence is a simple and natural task, and in a single trial it is possible to collect multiple syllables by, for example, syllabifying an entire sentence.

The experiment involved 21 participants (13 male, 8 female), native Italian speakers, each recruited for a single recording session. Each participant performed three runs per condition for the same task. Each run consisted of 20 trials, the same sentences were reused across runs. A total of 60 full trials per condition (20 sentences $\times$ 3 repetitions) was collected this way. A single trial consists of the repetition of the cued sentence, presented both visually on-screen and auditorily via audio playback, that must then be repeated once fluently, at a natural speed, and once performing syllabification, at a cued pace. Each pro-

cedure was repeated twice: first via *overt speech* (vocalized pronunciation), and then via *imagined speech* (mental rehearsal with no acoustic or motor output).

Below we describe in detail each phase of the trial:

- *Fixation (2s)*: A black cross is displayed at the center of the screen. The participant is required to look at the cross and remain still.
- *Sentence presentation (8s)*: The sentence appears in a center-aligned text box while its pronunciation is played using a text-to-speech synthesizer (gTTS, Google). A visual cue representing a person speaking or thinking is displayed. After playback, a 3-s pause allows the subject to finish reading and prepare for repetition.
- *Sentence repetition (5s)*: The sentence is cleared from the screen. The participant then pronounces the sentence at his own natural pace.
- *Syllabification (13-16s)*: After a 1-s pause, the syllabified sentence appears. Syllables scroll upward every 0.6 seconds, with the target highlighted in the middle row. A 3-2-1 dot countdown establishes rhythm, after which the participant articulates the syllables at the given pace.
- *Trial end (4s)*: The text box is cleared, and an on-screen cue signals the end of the trial. A 4-s interval precedes the next trial.

Figure 1 summarizes the trial structure described above. Each run lasted approximately 12 minutes, totaling more than 30 minutes of recording per condition. In addition to EEG data, audio was recorded via microphone during the overt speech block, and we used it to align EEG segments with speech onsets during post-processing. Initial alignment was conducted automatically using Silero

VAD [17] and the Montreal Forced Aligner [18], followed by a manual refinement step using the Praat toolbox [19]. This manual inspection also ensured that any syllables containing pronunciation or timing errors could be accurately corrected or discarded.

All sentences contain a recurrent syntactical structure (subject-verb-complement plus adjectives and articles), such that the effect of sentence length and structure on the recorded brain activity is negligible. Each sentence contains between 11 and 15 words, between 20 and 25 syllables. Syllabification has been automatically processed using phonetic transcriptions from PhonItalia [15], which is also providing the categorization of phonemes according to the International Phonetic Alphabet. The resulting phonetical syllables may differ from the orthographical ones as they reflect the characteristics of articulation and sound, rather than following the typographic rules used in written language (e.g., the orthographical syllabification of the word "pasta" is "pa-sta" while the phonological one is "pas-ta"). Each run allowed for the collection of 462 syllables (192 unique), which enable the construction of up to 22,000 distinct words based on PhonItalia metrics.

In this work, we employed only data related to the syllabification phase, full sentence repetitions were not taken into account. Moreover, the following sections will focus exclusively on the overt speech condition; decoding analyses of the imagined speech data are reserved for future study.

MODEL TRAINING

The high number of syllabic classes, combined with limited training data and the inherently low signal-to-noise ratio of EEG signals, poses a significant challenge for training a deep neural network from scratch. On the other hand, existing models for automatic speech recognition (ASR) have demonstrated high performance in extracting meaningful representations from acoustic data. Therefore, rather than employing a traditional end-to-end classification paradigm, we frame this as a representation learning problem. In this approach, the model learns a mapping from raw EEG data to a structured latent space, specifically aiming to align the embeddings generated by the EEG encoder with those produced by a pre-trained acoustic model. The procedures explained below are an adaptation of the work of [16].

Data processing: Data were recorded with a 64-channel AntNeuro eego system (standard 10-20 montage) at a sampling rate of 2,048 Hz. Channel CPz was used as reference. Frontal channels (AFx, FPx), occipital channels (Ox), and channels M1 and M2 were removed to reduce ocular artifacts and/or due to bad signal quality. The remaining 51 channels cover both the auditory cortex—involved in audio and speech processing—and the primary motor cortex—involved in speech articulation [20, 21]. Preprocessing was kept minimal to comply with real-time closed-loop requirements. The signal was notch

filtered at 50 Hz and band-pass filtered between 1 and 70 Hz. We chose to keep frequencies in the gamma band as they were shown to be effective for phonetic decoding in previous works [22, 23]. Finally, we downsampled the signal to 256 Hz to reduce the computational load.

Both EEG and audio data corresponding to the pronounced syllables were segmented according to the manual annotations. The audio signals were cropped to the exact syllable duration, while the EEG signals were epoched around speech onsets using a fixed window of [0.0, 0.88] s relative to the syllable onset. To extract acoustic features, the audio segments were processed through a pre-trained ASR model. Specifically, we utilized facebook/wav2vec2-large-xlsr-53-italian [24] as a feature extractor, capturing the embeddings from the feature projection layer prior to the transformer blocks. The resulting tensors were averaged across the temporal dimension to obtain a single vector representation for each syllable. For the experimental evaluation, the third run of each subject was reserved as a hold-out test set. The first two runs were split sequentially into training and validation sets using a 90%/10% ratio. To accelerate model convergence, all data underwent channel-wise robust scaling using the training set's median and 10-90% quantiles. We evaluated the architecture on both a multi-subject and a single-subject basis, maintaining consistent data splits across both comparisons to ensure a robust performance baseline.

Architecture: The model is composed of four primary blocks. Data first passes through a *Spatial Attention Layer*, which recombines information across the channel dimension using an attention mechanism; subsequently, a *Subject Layer* applies a participant-specific mapping. These two blocks are described in detail in [16]. In contrast to their approach, we maintain the original channel count during the spatial attention stage rather than enlarging the set. The resulting signals are then processed by an EEGNet-like encoder [25] to extract features from the projected data. The classification head of the standard EEGNet is replaced by a *Temporal Projection* layer (two Conv1d layers interleaved with a BatchNormalization and a GELU activation function). This layer yields a tensor of shape $B \times T \times H$, where B represents the batch size, and T and H match the temporal and hidden dimensions of the Wav2Vec embeddings, respectively. Finally, mirroring the audio embedding process, we average the EEG embeddings across the temporal dimension to obtain a single vector representation per syllable.

Training: The EEG embeddings were aligned to their acoustic counterparts using a contrastive objective inspired by the CLIP loss [26]. This applies a cross-entropy loss over the batch, effectively pulling corresponding EEG and acoustic embeddings together in the shared latent space while pushing non-matching samples apart. Formally, let X be an EEG segment and Y be its corresponding acoustic latent representation (the positive sample). Within a batch, we define $N - 1$ negative acoustic samples, grouping them such that the positive sample

is $\tilde{Y}_N = Y$. The EEG encoder maps X to a latent representation Z . By computing the inner product $\langle \cdot, \cdot \rangle$ followed by a softmax, we train the model using a simplified cross-entropy loss to maximize the probability of selecting the true positive sample:

$$L_{\text{CLIP}} = -\langle Z, Y \rangle + \log \left(\sum_{j'=1}^N e^{\langle Z, \tilde{Y}_{j'} \rangle} \right)$$

The experiments were conducted on a workstation equipped with an AMD Ryzen 7 7700 processor, 16 GB of RAM, and an NVIDIA RTX 3060 GPU with 12 GB of VRAM. The models were trained for a maximum of 1,000 epochs, utilizing an early stopping criterion if the validation loss failed to improve for 75 consecutive epochs. Optimization was performed using the AdamW optimizer with a learning rate of 5×10^{-4} , weight decay of 10^{-4} , and a CosineAnnealing scheduler. Batch size was set to 128. Data and code will be provided in a future release.

RESULTS

We approached the syllable decoding task as an information retrieval problem, effectively implementing a top- k retrieval in the learned latent space. During the evaluation phase, the acoustic modality was completely discarded. Instead, we constructed a reference set exclusively from the EEG embeddings of the training data. For every unseen test sample, we computed its cosine similarity against all individual training embeddings in this reference set, selecting the k nearest candidates to compute our retrieval accuracies. Let $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N\}$ be the set of reference embeddings corresponding to our syllable vocabulary, and $\hat{\mathbf{y}}$ be the predicted EEG embedding. The predicted syllable S is given by:

$$S = \operatorname{argmax}_{i \in \{1, \dots, N\}} \frac{\mathbf{v}_i \cdot \hat{\mathbf{y}}}{\|\mathbf{v}_i\| \|\hat{\mathbf{y}}\|}$$

A prediction is considered correct if the target class is present between the first k predictions of the model.

Table 1 reports the per-subject and average decoding accuracies (Top-1 and Top-10) for both multi-subject and single-subject variants. Overall, the single-subject approach generally outperforms the multi-subject model, achieving an average Top-1 accuracy of 9.5% and a Top-10 accuracy of 38.2%, compared to Top-1 of 9.1%, Top-10 of 37.4% of the multi-subject model. Despite the significantly smaller volume of training data available per model in the single-subject configuration, its superior performance underscores the inherent difficulties of integrating multi-subject EEG data due to high inter-subject variability. When evaluated against a ZeroR baseline, performances exceed chance levels (3,9% Top-1, 22,9% Top-10) in nearly all cases.

In the ablation study (Table 2) we evaluated the individual and combined contributions of the *Spatial Attention Layer* and *Subject Layer* against the base architecture. Results highlight a critical role of the *Subject Layer*,

Table 1: Comparison of retrieval accuracy between single- and multi-subject models

Subj	Multi-subject		Single-subject	
	Top-1	Top-10	Top-1	Top-10
<i>k7</i>	14,3%	44,8%	16,5%	47,0%
<i>i5</i>	15,4%	56,4%	15,2%	57,0%
<i>k3</i>	18,3%	54,8%	15,1%	53,9%
<i>i1</i>	13,4%	45,7%	14,1%	45,0%
<i>k4</i>	10,5%	40,2%	14,0%	50,7%
<i>j8</i>	8,1%	35,7%	13,3%	46,2%
<i>c9</i>	9,5%	39,3%	11,9%	45,3%
<i>j4</i>	9,6%	40,8%	11,6%	40,2%
<i>j5</i>	11,6%	43,7%	10,7%	42,8%
<i>i8</i>	10,2%	41,5%	9,6%	39,1%
<i>f9</i>	11,5%	41,2%	9,4%	37,7%
<i>d7</i>	6,5%	37,4%	8,2%	37,0%
<i>h9</i>	7,2%	34,6%	7,8%	36,7%
<i>i7</i>	6,3%	35,4%	7,8%	37,4%
<i>k2</i>	9,5%	39,3%	6,7%	31,5%
<i>k5</i>	5,0%	30,3%	6,1%	22,4%
<i>j3</i>	5,9%	31,7%	5,7%	33,7%
<i>k6</i>	9,5%	28,4%	4,6%	29,9%
<i>j6</i>	2,6%	21,0%	4,3%	25,8%
<i>k1</i>	2,0%	18,0%	4,1%	23,9%
<i>j7</i>	3,8%	24,2%	3,6%	18,0%
Average	9,1%	37,4%	9,5%	38,2%

which delivers a substantial performance boost across both paradigms. In the multi-subject setting, appending the *Subject Layer* to the base model improves Top-1 accuracy by 6.3% and Top-10 accuracy by 22.7%. This shows how adding the layer enables subject-specific processing that facilitates cross-subject data integration. Interestingly, the *Subject Layer* also improves the single-subject model, improving Top-1 accuracy by 1.5% and Top-10 by 5.67%. Although subject-specific processing is already inherent to the single-subject approach, we hypothesize that this layer acts as an additional linear transformation that effectively enhances the model’s capacity for generalization.

The addition of the *Spatial Attention Layer* by itself fails to provide a performance boost over the base model, instead causing a drop in accuracy for both multi-subject (-0.2% Δ Top-1, Top-10 unchanged) and single-subject (-5.3% Δ Top-1, -15.2% Δ Top-10) configurations. However, when coupled with the *Subject Layer*, spatial attention further boosts multi-subject classification performance, allowing it to reach the full model accuracies reported above. While it introduces a slight degradation in the single-subject context—dropping from 9.6% to 9.5% in Top-1 and from 39.5% to 38.1% in Top-10, this difference is marginal. Consequently, we chose to retain it in the final architecture. Its inclusion maintains architectural alignment for direct comparison with previous methodologies [9, 16], while simultaneously providing a mechanism for improved interpretability—such as the inspection of spatial attention maps—in future work.

To obtain a qualitative assessment of the learned latent

Table 2: Model ablations

model	Multi-subject		Single-subject	
	Top 1	Top 10	Top 1	Top 10
<i>Full model</i>	9,1%	37,4%	9,5%	38,2%
+ <i>Subject Layer</i>	8,2%	36,0%	9,6%	39,5%
+ <i>Spatial Attention</i>	1,6%	13,3%	2,9%	18,6%
<i>Base Model</i>	1,8%	13,3%	8,1%	33,8%

space, UMAP was applied to the test embeddings to reduce their dimensionality for 2D visualization. Data points were colored according to a gradient reflecting sonority and manner of articulation. The emerging topology, resulting from the cross-modal alignment, confirms the efficacy of the contrastive training objective. Additionally, it establishes a robust theoretical foundation for zero-shot or few-shot vocabulary expansion without explicit retraining, although this hypothesis remains to be verified. Figure 2 illustrates the learned space of the top four subjects, ranked according to their Top-10 accuracy in the multi-subject model. Notably, while the underlying topological sorting of phonetic features remains robustly consistent across all speakers, the specific geometric shape of the latent manifold exhibits subject-specific variations, highlighting the model’s capacity to capture both global linguistic structures and individual speaker variability.

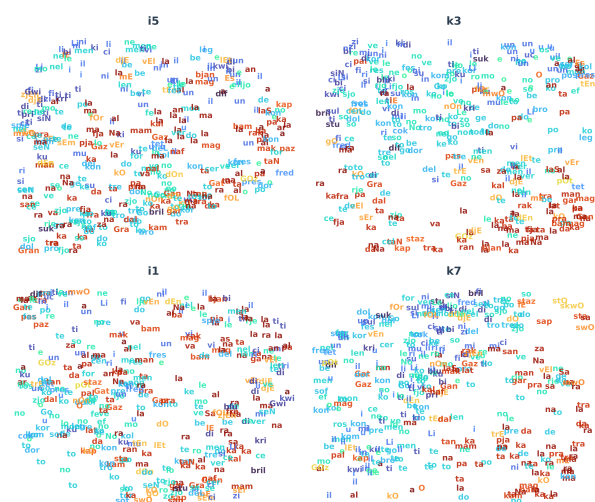


Figure 2: UMAP projections of the multi-subject model’s latent space across four distinct subjects. Data points represent individual syllables, color-coded according to a phonetic sonority gradient, ranging from low-sonority syllables often characterized by occlusive consonants (blue/purple) to high-sonority syllables dominated by open vowels (red/orange). The visualizations give interesting qualitative insights about the model’s learned representation. Syllables with similar acoustic and articulatory properties naturally form cohesive clusters along a continuous phonetic spectrum.

CONCLUSIONS

In this study, we introduced a novel framework for an EEG-based speech BCI leveraging the syllabic structure

of the Italian language. By conceptualizing syllable decoding as a cross-modal representation learning problem, we aligned high-dimensional EEG signals with pre-trained acoustic embeddings via a contrastive objective. We established a robust baseline for overt speech decoding, utilizing a retrieval strategy operating exclusively within the learned EEG latent space during inference. Our evaluations demonstrated that inter-subject variability remains a primary bottleneck for generalized BCI systems, as single-subject models outperformed multi-subject architectures despite relying on significantly less training data. Crucially, our ablation studies highlighted the instrumental role of the *Subject Layer*, which provided substantial accuracy boosts across both paradigms by facilitating subject-specific feature integration. While the *Spatial Attention Layer* did not independently improve decoding accuracy, its inclusion alongside the *Subject Layer* stabilized multi-subject performance and was retained to maintain architectural alignment with current state-of-the-art methodologies.

Despite the promising baseline established by this framework, we acknowledge some critical methodological limitations. First, the current decoding evaluation is restricted to the overt speech condition. While this provides a necessary ground truth for cross-modal alignment, vocalized speech inevitably introduces electromyographic (EMG) artifacts that most likely have inflated the decoding performance. However, for end-users unable to produce intelligible speech, residual articulatory movements may still provide a valuable complementary EMG signal. Capturing this via a non-invasive EEG cap represents a highly accessible and elegant alternative to surgical implants or dense arrays of facial electrodes. Furthermore, it is important to note that even state-of-the-art invasive neuroprostheses rely on attempted speech paradigms, in which the participant actually attempts the articulation, rather than purely imagined speech [3, 4]. This said, it still holds that the true cortical decoding capacity of our non-invasive model must be validated on the imagined speech data in future iterations. A second critical limitation of the current study is the reliance on the manual temporal alignment of EEG data to speech onsets, guided by the synchronized audio recordings. Preliminary tests conducted without this manual alignment resulted in an approximate 12% drop in decoding performance. This discrepancy highlights the necessity of developing a robust, brain-based Voice-Activity Detector (VAD) capable of accurately triggering the decoder in real time without acoustic ground truth.

FUTURE DIRECTIONS

Looking forward, our immediate next steps are to extend this decoding framework to the imagined speech data collected within our protocol, and to implement an automated VAD to ensure accurate, real-time onset detection. We estimate that our syllable-based BCI could achieve a theoretical maximum communication rate of approx-

imately 50 WPM prior to any predictive text enhancements. Although this remains well below the average rate of natural speech, this is more than double the speed of current AAC systems. To achieve full sentence decoding, we will focus on integrating syllable probabilities with N-gram language models and/or Large Language Models. By combining neural decoding with predictive autocompleteness and context-aware language modeling, this framework has the potential to scale into a fast, non-invasive, and highly expressive communication device.

ACKNOWLEDGEMENTS

This publication is part of the project PNRR-NGEU which has received funding from the MUR-DM 118/2023.

REFERENCES

- [1] Elsahar Y, Hu S, Bouazza-Marouf K, Kerr D, Mansor A. Augmentative and alternative communication (aac) advances: A review of configurations for individuals with a speech disability. *Sensors*. 2019;19(8):1911.
- [2] Silva AB, Littlejohn KT, Liu JR, Moses DA, Chang EF. The speech neuroprosthesis. *Nature Reviews Neuroscience*. 2024;25(7):473–492.
- [3] Littlejohn KT et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*. 2025;28(4):902–912.
- [4] Qian Y et al. Real-time decoding of full-spectrum chinese using brain-computer interface. *Science Advances*. 2025;11(45):eadz9968.
- [5] Wyse-Sookoo K et al. Stability of ecog high gamma signals during speech and implications for a speech bci system in an individual with als: A year-long longitudinal study. *Journal of neural engineering*. 2024;21(4):046016.
- [6] Nieuwland MS et al. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*. 2018;7:e33468.
- [7] Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*. 2018;28(5):803–809.
- [8] Accou B, Bollens L, Gillis M, Verheijen W, Van Hamme H, Francart T. Sparrkulee: A speech-evoked auditory response repository from ku leuven, containing the eeg of 85 participants. *Data*. 2024;9(8):94.
- [9] d’Ascoli S et al. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*. 2025;16(1):10521.
- [10] Lopez-Bernal D, Balderas D, Ponce P, Molina A. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in human neuroscience*. 2022;16:867281.
- [11] Bisla M, Anand R. A comprehensive review on state-of-the-art imagined speech decoding techniques using electroencephalography. *Artificial Intelligence in Biomedical and Modern Healthcare Informatics*. 2025;101–126.
- [12] Zhang Z et al. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*. 2024;11(1):1265.
- [13] Sato M, Tomeoka K, Horiguchi I, Arulkumaran K, Kanai R, Sasai S. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *arXiv preprint arXiv:2407.07595*. 2024.
- [14] Craik A, Dial H, L Contreras-Vidal J. Continuous and discrete decoding of overt speech with scalp electroencephalography (eeg). *Journal of Neural Engineering*. 2025;22(2):026017.
- [15] Goslin J, Galluzzi C, Romani C. Phonitalia: A phonological lexicon for italian. *Behavior research methods*. 2014;46(3):872–886.
- [16] Défossez A, Caucheteux C, Rapin J, Kabeli O, King JR. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*. 2023;5(10):1097–1107.
- [17] Team S. Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. <https://github.com/snakers4/silero-vad>. 2021.
- [18] McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M. Montreal forced aligner: Trainable text-speech alignment using kald. In: *Interspeech*. 2017, 498–502.
- [19] Boersma P. Praat: Doing phonetics by computer (version 5.1.05). <http://www.praat.org/>. 2009.
- [20] Hickok G, Poeppel D. The cortical organization of speech processing. *Nature reviews neuroscience*. 2007;8(5):393–402.
- [21] Mantegna F, Orpella J, Poeppel D. Time-resolved hemispheric lateralization of audiomotor functional connectivity during covert speech production. *Cell Reports*. 2025;44(1).
- [22] Jahangiri A, Sepulveda F. The relative contribution of high-gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in eeg data. *Journal of medical systems*. 2019;43(2):20.
- [23] Bhadra K, Giraud AL, Marchesotti S. Learning to operate an imagined speech brain-computer interface involves the spatial and frequency tuning of neural activity. *Communications Biology*. 2025;8(1):271.
- [24] Baeviski A, Zhou H, Mohamed A, Auli M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*. 2020;abs/2006.11477.
- [25] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of neural engineering*. 2018;15(5):056013.
- [26] Radford A et al. Learning transferable visual models from natural language supervision. *CoRR*. 2021;abs/2103.00020.

Mirror-Box Pretraining to Evoke Sensorimotor Activation to Facilitate Motor Imagery Detection in Brain-Computer Interface Stroke Rehabilitation

Nina Evetović¹, Roman Rosipal¹, Adrián Hodúr¹, Zuzana Rošťáková¹

¹Institute of Measurement Science, Slovak Academy of Sciences

E-mail: nina.evetovic@savba.sk

ABSTRACT: Calibration of motor imagery (MI) brain-computer interface (BCI) systems in stroke patients is challenging due to limited voluntary movement and variable imagery ability. We propose mirror-box (MB) training as a physiologically grounded pretraining strategy to induce contralateral sensorimotor activation and extract subject-specific oscillatory components for MI detection. Electroencephalography (EEG) was recorded during six MB sessions in a chronic post-stroke participant. Parallel Factor Analysis (PARAFAC) was applied to decompose multi-channel EEG into spatial, temporal, and spectral components. Trial-pair normalization relative to the corresponding *rest* segment was used to enhance detection of task-related modulation. Distinct μ and β sensorimotor rhythms were identified, showing significant desynchronization during observation and movement compared to *rest*. An individualized linear discriminant analysis classifier achieved $98.8\% \pm 0.6\%$ accuracy in discriminating *rest* from *motor engagement*. These findings support MB pretraining as a robust strategy for physiologically informed MI-BCI calibration and future BCI Head-Mounted Display rehabilitation applications.

INTRODUCTION

Post-stroke upper limb impairment is a major cause of long-term disability, motivating the development of Brain-Computer Interface (BCI) based neurorehabilitation combined with immersive Head-Mounted Display (HMD) systems [1]. Motor imagery (MI) BCI paradigms rely on the modulation of sensorimotor (SM) rhythms, typically in the alpha and beta frequency bands, and on the reliable detection of these modulations for system control [2, 3].

However, training MI-based BCI-HMD systems in post-stroke patients presents substantial challenges. Calibration based on physical movement of the affected limb is often not feasible. MI can be used as an alternative [1], but its effectiveness strongly depends on the patient's imagery ability, which is highly variable and difficult to verify objectively [4].

To address these limitations, we propose Mirror-Box (MB) training as a physiologically grounded pretraining strategy for identifying oscillatory components suitable for MI-BCI control. MB therapy provides visual feedback of the intact limb to induce the illusion of move-

ment in the paretic limb, reliably activating sensorimotor cortical networks [5]. These responses are stimulus-driven and cannot be voluntarily fabricated [6], making MB training a robust approach for extracting subject-specific sensorimotor generators. Furthermore, MB therapy is an established rehabilitation approach for paretic and hemiparetic patients, leveraging visual feedback to engage sensorimotor networks and facilitate motor recovery [7], making it a valuable addition to BCI-based protocols. While not the primary focus of this work, a study from 2024 demonstrated that MB training can serve as a training tool to improve participants' motor imagery (MI) ability [8], which may be regarded as an additional benefit of the proposed framework.

Electroencephalography (EEG) provides a non-invasive means to capture cortical oscillatory activity associated with motor observation and execution [2, 5]. In this study, we applied tensor decomposition to multi-channel EEG recorded during MB training to extract individualized oscillatory components corresponding to sensorimotor activity. Using Parallel Factor Analysis (PARAFAC), we preserved the multidimensional structure of the data across space, time, and frequency, enabling identification of μ and β rhythms modulated during MB training [5, 9]. The aim of this paper is threefold: (i) to extract physiologically interpretable oscillatory components reflecting sensorimotor generators elicited by MB training, (ii) to quantify their discriminative capacity by constructing a linear discriminant analysis (LDA) classifier to differentiate rest from motor-related states, and (iii) to propose MB training as a structured strategy for training and calibrating MI-based BCI models. By using MB training to elicit robust and verifiable sensorimotor activation, we aim to establish a principled framework for feature extraction that can support future BCI-HMD motor neurorehabilitation applications and contribute to a deeper understanding of the neural mechanisms underlying MB therapy.

MATERIALS AND METHODS

Experimental Paradigm and Protocol:

We recorded six MB training sessions. Each experimental session began with baseline resting state recordings consisting of 2 minutes with eyes closed (EC) followed by 2 minutes with eyes open (EO). During the EO condi-

tion, the participant was instructed to fixate on a centrally positioned plus sign (+) to minimize ocular artifacts.

The main part of each session consisted of two MB training blocks. In the first block, referred to as *Visual Left Hand* (VLH), the participant performed movements with the unaffected upper limb while the paretic limb was positioned behind the mirror. The participant observed the reflection of the healthy hand in the mirror, creating the visual illusion of movement in the affected limb. In the second block, *Visual Both Hands* (VBH), the participant again observed the reflection of the healthy hand but simultaneously attempted to move the affected limb in synchrony with the observed movement.

Each block began with a 30 s eyes open resting period. A technician seated opposite the participant demonstrated three to four repetitions of each movement, after which the participant performed the same movement at a self-paced rhythm for approximately 7 to 10 repetitions. Consecutive movement types were separated by a 15 s eyes open resting interval. This procedure was repeated until all 10 movement types, each comprising rest, observation, and movement periods, were completed. Each block concluded with a final 30 s eyes open resting period. After completion of both training blocks, additional 2 minute resting state recordings were obtained under EO and EC conditions.

Each trial was evaluated for separability between *rest* and motor-related conditions (*observe* and *move*), based on mirror neuron mechanisms that support activation during both action observation and execution. These phases were present in both VLH and VBH blocks, but were later unified into functional labels (*rest*, *observe*, *move*) rather than block-specific terminology. Both VLH and VBH were included to increase variability by capturing unilateral and bilateral motor activations, improving generalizability of the learned representations; therefore, all blocks were jointly used for classifier training.

Each mirror box training block included 10 distinct upper limb movements, with following instructions:

- **Elbow-Supported Hand Lift** – Elbow resting on the table; lift the hand while keeping the forearm in contact.
- **Palm Rotation** – Forearm on the table; rotate the palm upward and downward.
- **Palm Lift** – Rest of the hand on the table; lift the palm upward.
- **Finger Spread and Close** – Spread the fingers apart and then bring them together.
- **Fist Closure** – Palm up on the table; close fingers into a fist with the thumb inside.
- **Ball Roll** – Roll a ball on the table using the palm; lift fingers as the ball reaches the bottom of the palm.
- **Bottle Grasp and Release** – Grasp a bottle, release it, then return the hand to the table.
- **Finger Pinch** – Palm up on the table; pinch the fingers together.
- **Thumb-to-Finger Taps** – Palm up; sequentially touch each finger with the thumb from index to little finger

and back.

- **Cube Manipulation** – Palm down; move a cube once using only the thumb and index finger, then move it back using all five fingers.

Participant Characteristics and Dataset Description:

The study included a single participant, a 70-year-old individual with a 10-year history of spastic hemiparesis affecting the right upper limb more prominently. The participant was able to walk independently and demonstrated normal speech and cognitive function. They were capable of following instructions and managing the training session schedule without difficulty.

EEG signals were recorded continuously throughout each session, including baseline resting-state periods with EC and EO, as well as during two MB training blocks. The dataset therefore includes both resting-state and task-related EEG segments, providing sufficient temporal coverage for the analysis of SM rhythms.

EEG Acquisition and Recording Setup:

Continuous EEG activity was recorded using the wireless g.Nautilus PRO FLEXIBLE system, an FDA-cleared device. The system was equipped with 19 Ag/AgCl active wet electrodes arranged according to the international 10–20 system and integrated into an 19-channel g.tec cap and amplifier. The EEG cap was positioned following the manufacturer's instructions. Signals were sampled at 250 Hz.

A1 served as the reference electrode and AFz as the ground. For subsequent signal processing, recordings were re-referenced to the the A2 electrode, resulting in the referencing with the average of both earlobe electrodes, A1 and A2, using the formula $(A1 + A2)/2$.

Signal Preprocessing Pipeline:

The EEG recordings were initially converted to EDF format and preprocessed using Brain Vision Analyzer 2.3.03 [10]. The preprocessing steps included:

- (a) identification and interpolation of channels with missing or corrupted data;
- (b) a semi-automatic threshold-based procedure to detect general artifacts, followed by visual inspection to ensure accurate marking; and
- (c) detection of electromyographic (EMG) artifacts prior to data export.

Subsequently, the data were processed in MATLAB [11], where the continuous EEG signals were segmented into 2 s epochs with a 500 ms moving window. From these epochs, features including fractal, harmonic, and raw spectral components were extracted using an Irregular-Resampling Auto-Spectral Analysis (IRASA) approach [12] with a Hann window.

Parallel Factor Analysis:

Parallel Factor Analysis (PARAFAC) is a tensor decomposition method that extends matrix factorization techniques to multi-way data structures [9]. In contrast to matrix-based approaches that require unfolding higher-order data into two-dimensional representations, PARAFAC preserves the intrinsic multi-dimensional structure of the data, which is particularly advantageous

for EEG signals that naturally exhibit spatial, temporal, and spectral modes [13, 14].

Let $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ denote a three-way tensor constructed from EEG power spectral density estimates across I time segments, J electrodes, and K frequency bins. The PARAFAC model decomposes $\mathbf{X}$ into a sum of F rank-one tensors:

$$\mathbf{X} \approx \sum_{f=1}^F \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f + \mathbf{E}, \quad (1)$$

where $\mathbf{a}_f$, $\mathbf{b}_f$, and $\mathbf{c}_f$ represent the temporal, spatial, and spectral loading vectors of the f -th component, respectively, $\circ$ denotes the vector outer product, and $\mathbf{E}$ is the residual error tensor [9]. Each component therefore corresponds to a tri-linear pattern, often referred to as an “atom,” that jointly characterizes space–time–frequency signatures in the EEG data [14].

Additional constraints may be incorporated to reflect prior knowledge about EEG structure. In particular, non-negativity constraints in selected modes can enhance interpretability when working with amplitude spectral representations [15].

In line with previous applications of PARAFAC to MB therapy and SM rhythm analysis [5], the extracted loading matrices were interpreted as temporal activations, spatial topographies, and spectral profiles of oscillatory EEG components. This space–time–frequency decomposition enables tracking of task-related modulations and longitudinal changes in SM oscillatory activity.

PARAFAC for Sensorimotor Rhythm Identification:

The preprocessed EEG data were arranged into a three-way tensor with dimensions corresponding to channels (spatial), time, and frequency. PARAFAC decomposition was applied to extract task-related oscillatory components across these modes [9, 14].

The resulting loading matrices represented temporal activations, spatial topographies, and spectral profiles of each component (as shown in Fig. 1). Components exhibiting spectral peaks in the alpha (7–12 Hz) and beta (13–30 Hz) bands over sensorimotor regions were identified as SM rhythms and analyzed for task-related modulation during MB training [5].

Each component was visually inspected, and only those localized to the sensorimotor area, within the expected frequency ranges, and consistent with the oscillatory patterns observed in the data were selected. Components not meeting these criteria were discarded.

Individualized Modeling Using Oscillatory SM Components:

After selecting the SM components relevant to motor activity during the task, a linear discriminant analysis (LDA) classifier was constructed. LDA has been widely used in EEG and motor imagery classification due to its simplicity, efficiency, and robust performance in discriminating sensorimotor states from oscillatory EEG features [16]. An individualized LDA classifier was trained to discriminate between *rest* and *active motor engagement*,

where the active class combined both observation and movement segments.

The model was trained using data from 12 separate blocks, corresponding to one cycle of 10 exercises, each comprising rest, observation, and movement segments. To ensure robust performance estimation, a leave-one-out cross-validation approach was applied. In each fold, 11 blocks were used for training and the remaining block was used for testing.

To account for baseline variability and the task structure, trial-by-trial normalization was applied. Specifically, each trial was normalized using the resting-state data from the same trial, providing a precise baseline for detecting changes in sensorimotor activity during observation and movement. This procedure ensured that task-related neural dynamics could be reliably captured across trials.

Statistical Analysis:

For each extracted sensorimotor component, scores were computed by projecting the EEG data onto the atom’s spatial and frequency weights, providing a measure of how strongly the component was expressed in each trial. These scores were then averaged across trials within each block to obtain condition-specific values for *rest*, *observe*, and *move* (12 blocks across six sessions).

Statistical comparisons between *rest* vs *move* and *rest* vs *observe* were conducted using the Wilcoxon signed-rank test, with Bonferroni correction applied to account for multiple comparisons. Significance was reported at corrected confidence levels of $p < 0.05$, $p < 0.01$, and $p < 0.001$. A direct comparison between *observe* and *move* was not performed, as both were grouped as motor-related conditions and evaluated only relative to the *rest* baseline, which is the primary reference in sensorimotor rhythm analysis and the proposed framework.

RESULTS

Detection of Oscillatory SM Rhythms:

Figure 1 presents the participant’s SM oscillatory rhythms extracted using PARAFAC decomposition. Five distinct components were identified, all localized over the sensorimotor cortex. Prior to visualization, spatial weight vectors were zero padded to increase the density of displayed samples and avoid visual broadening due to interpolation. The padding was removed after rendering and did not influence the computed component structure. Two components representing the μ rhythm were separated despite sharing the same spectral peak at 11.8 Hz, indicating spatially distinct sources.

In addition to the μ components, three distinct beta-range components were identified at 13.7 Hz, 16.7 Hz, and 19.5 Hz. The 13.7 Hz component corresponds to SMR, whereas the components at 16.7 Hz and 19.5 Hz represent higher β sub-bands within the sensorimotor rhythm.

Sensorimotor Atom Scores:

In Figure 2, we show the computed atom scores across the *rest*, *observe*, and *move* conditions. For each fre-

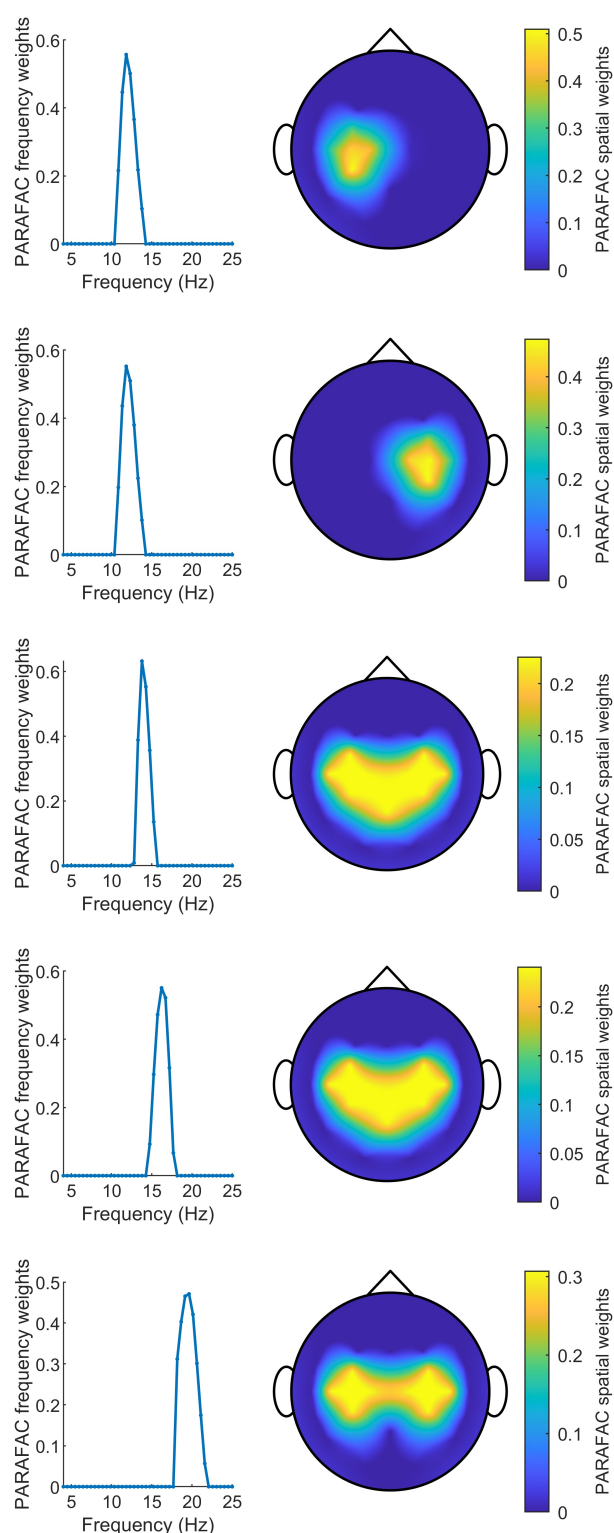


Figure 1: Five SM oscillatory components identified using PARAFAC decomposition. Each component is represented by its frequency weights (left) and the corresponding scalp topography of spatial weights (right). The components correspond, respectively, to the left μ rhythm (11.8 Hz), right μ rhythm (11.8 Hz), SMR (13.7 Hz), β 1 (16.7 Hz), and β 2 (19.5 Hz).

quency band, mean t-scores were calculated for all extracted atoms and averaged across blocks. Boxplots illustrate the distributions of these scores across conditions.

The scores were lower during *move* and *observe* compared to *rest*, indicating that the extracted atoms captured task-related sensorimotor activity and possibly reflected event-related desynchronization (ERD). Wilcoxon signed-rank tests (Bonferroni-corrected) confirmed that these differences were significant for most frequency bands, supporting the reliability of the atoms in reflecting condition-specific modulation of sensorimotor networks. The most statistically significant ERD across trials was observed in the μ rhythm. An example of this desynchronization for a representative movement averaged over all trials is shown in Fig. 3.

LDA Classification of Sensorimotor States:

The individualized LDA classifier demonstrated excellent performance in discriminating *rest* versus *motor engagement*. Across cross-validation folds, the classification accuracy was $98.8\% \pm 0.6\%$, indicating highly consistent predictions. The mean area under the ROC curve (AUC) was 0.990 ± 0.0 , reflecting near-perfect discrimination between classes. The effect size, quantified using Cohen's d , was $d = 2.374 \pm 0.0$, indicating a large separation between rest and motor-related states.

DISCUSSION

In this study, we investigated the use of MB pretraining to identify SM oscillatory components relevant for MI detection in a BCI-HMD paradigm. Our results build on previous work comparing MB-induced EEG activity with signals recorded during robotic splint-assisted movements, we observed overlapping SM generators, supporting the physiological validity of mirror-based activation and demonstrating substantial similarity between atoms extracted from MB training and those elicited by robotic arm movements [5, 17].

Unlike these earlier, more heuristic approaches, the present study applied a LDA classifier to evaluate whether the identified SM components could reliably discriminate motor-related states in new participants. This methodological refinement allowed us not only to confirm the presence of MB-related oscillatory generators but also to quantify their discriminative capacity, providing evidence that these components capture physiologically meaningful SM activity.

Inspection of the extracted components revealed distinct left and right μ rhythms. The right μ component likely reflects contralateral SM activation associated with hand movement. In contrast, the left μ component may represent cortical processes related to MB training, specifically the visual observation of the hand in the mirror, potentially engaging mirror neuron system mechanisms through visually induced motor simulation.

It has been proposed that multiple β rhythms contribute to motor processing and may serve functionally distinct roles. Specifically, different β sub-bands have been

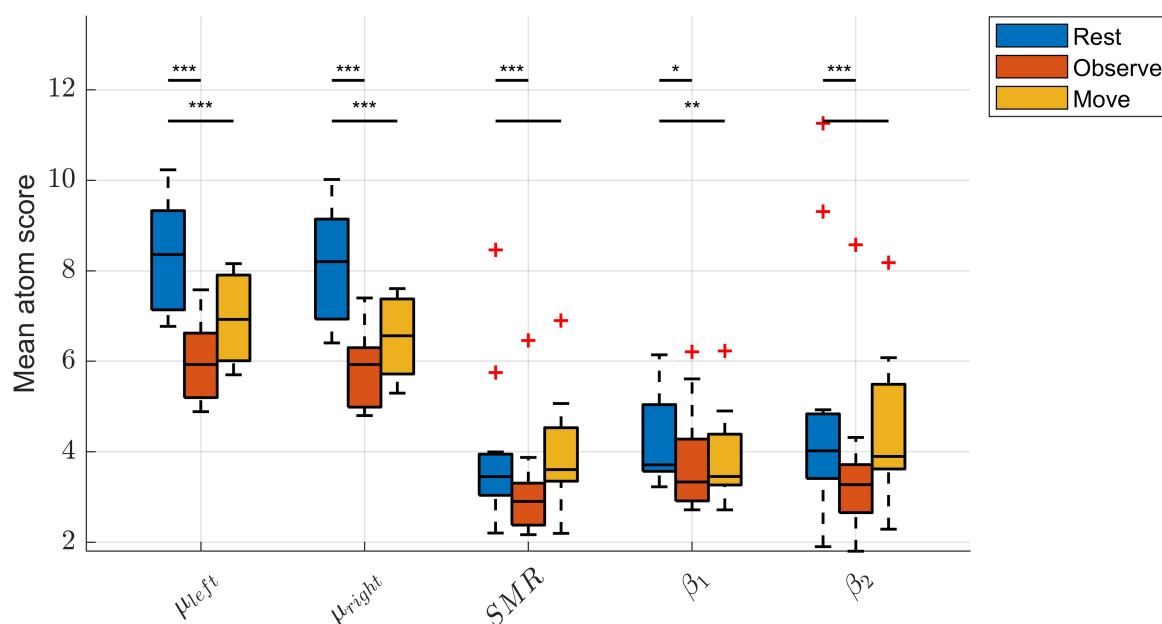


Figure 2: Five SM atom scores for *rest*, *observe* and *move* conditions averaged per block. Asterisks indicate statistically significant differences after correction for multiple comparisons: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

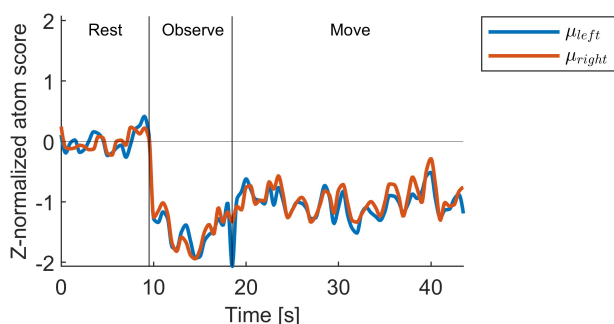


Figure 3: Z-normalized scores of the left and right μ components across *rest*, *observe*, and *move* segments. Scores were normalized relative to the corresponding *rest* baseline. The figure illustrates desynchronization for one representative movement, averaged across all sessions for the same movement task.

reported to exhibit divergent modulation patterns during motor tasks, with some showing ERD while others demonstrate synchronization. These sub-bands may also differ in their peak frequency and spatial distribution, suggesting distinct underlying neural generators and functional relevance.

In our study, the μ rhythm exhibited clear desynchronization during both observation and movement conditions, consistent with engagement of the SM cortex. In contrast, β band activity showed stronger desynchronization during observation than during movement. This difference likely arises from the design of the movement blocks: each movement consisted of approximately ten repetitions, and β rhythms are known to resynchronize after movement, which reduces observable desynchronization in averaged epochs. In comparison, μ rhythms fluctuate but remain desynchronized over longer periods, allowing more reliable detection of task-related modulation. Statistical analysis supported these findings, with signifi-

cant desynchronization observed in both observation and movement conditions relative to rest for the μ rhythms, whereas β band effects were less consistent.

These findings highlight the importance of an appropriate baseline and task protocol to accurately capture oscillatory dynamics. Moreover, our results demonstrate that careful identification of subject-specific SM atoms, combined with individualized classifier training, can yield high-performance detection of *motor engagement*, as reflected in the near-perfect LDA accuracy.

Overall, these results support the use of MB pretraining as a robust strategy for extracting physiologically meaningful EEG features in patients with limited motor function. By providing stimulus-driven activation that does not rely on voluntary movement, MB training enables the identification of SM generators that can subsequently be used to train MI-based BCI models, facilitating reliable detection of motor imagery in rehabilitation contexts. Future work will extend these findings to online BCI-HMD paradigms and investigate their applicability across a larger cohort of post-stroke patients.

CONCLUSION

This study demonstrates that MB pretraining combined with PARAFAC decomposition enables extraction of physiologically meaningful, subject-specific SM (μ and β) components in a chronic post-stroke participant. The identified atoms showed significant task-related desynchronization and enabled near-perfect discrimination between rest and *motor engagement* using an individualized LDA classifier (98.8% accuracy, $AUC = 0.990$).

These findings support MB training as a robust, stimulus-driven strategy for calibrating MI-based BCI models when voluntary movement is limited.

The use of multiple distinct movements was intended to capture a broader range of motor-related neural activations and improve the generalizability of the learned representations. In the current work, these movements were combined into a binary classification framework (*motor engagement* vs. *rest*), consistent with the intended application of pretraining spatial filters. However, extending this approach to multi-class classification between different movements represents a valuable direction for future work and could provide further insight into the separability and specificity of the extracted motor patterns. A direct comparison between VLH and VBH conditions was not performed in the present study, as the primary objective was to distinguish *motor engagement* from *rest* for pretraining purposes. However, comparative analysis between unilateral and bilateral conditions may provide additional insight into lateralized versus distributed sensorimotor activation and will be considered in future work. A key limitation of the present study is the absence of validation in an MI-based setting, which restricts the ability to directly assess the effectiveness of the proposed pretraining strategy in the intended application domain. Additionally, the experimental design limits the investigation of β synchrony and the precise temporal interplay of SM rhythms, particularly due to the repeated movement structure and epoch-based averaging.

Therefore, future work will evaluate the consistency of the identified SM generators in BCI-HMD settings, assess performance without trial-wise normalization, and further examine the temporal dynamics and functional roles of distinct μ and β sub-bands across repeated movements. Overall, this study represents a preliminary step toward improving MI-based BCI performance, with validation in MI contexts identified as a critical next step.

FUNDING

This research was supported by the EU NextGenerationEU initiative through the Recovery and Resilience Plan for Slovakia under project No. 09I03-03-V04-00205 (Z.R.) and project No. 09I03-03-V04-00443 (R.R. and A.H.). Additional funding was provided by the European Commission through the HORIZON-MSCA 2022-DN, 101118964-DONUT project. This work was also supported by the Slovak Research and Development Agency under grant APVV-21-0105 (TInVR) and VEGA 2/0094/26.

REFERENCES

[1] Khan MA, Das R, Iversen HK, Puthusserypady S. Review on motor imagery based BCI systems for upper limb post-stroke neurorehabilitation: From designing to application. *Computers in Biology and Medicine*. 2020;123.

[2] Pfurtscheller G, Neuper C. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*. 1997;239(2-3):65–68.

[3] McFarland D, Wolpaw J. EEG-based brain–computer interfaces. *Current Opinion in Biomedical Engineering*. 2017;4:194–200.

[4] Chen YY, Lambert KJ, Madan CR, Singhal A. Mu oscillations and motor imagery performance: A reflection of intra-individual success, not inter-individual ability. *Human Movement Science*. 2021;78:102819.

[5] Rosipal R, Porubcová N, Barančok P, Cimrová B, Farkaš I, Trejo LJ. Effects of mirror-box therapy on modulation of sensorimotor EEG oscillatory rhythms: A single-case longitudinal study. *Journal of Neurophysiology*. 2019;121(2):620–633.

[6] Oberman LM, McCleery JP, Ramachandran VS, Pineda JA. EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*. 2007;70(13-15):2194–2203.

[7] Gandhi DB, Sterba A, Khatter H, Pandian JD. Mirror Therapy in Stroke Rehabilitation: Current Perspectives. *Therapeutics and Clinical Risk Management*. 2020;Volume 16:75–85.

[8] Gómez DMC, Braidot AAA. Improving motor imagery through a mirror box for BCI users. *Journal of Neurophysiology*. 2024;131(5):832–841.

[9] Bro R. Parafac: Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*. 1997;38(2):149–171.

[10] Brain Products GmbH. *Brainvision analyzer*. Version 2.3.0. Computer software. Gilching, Germany, 2023.

[11] The MathWorks Inc. *Matlab version 24.1 (r2024a)*. Natick, Massachusetts, United States: The MathWorks Inc., 2024.

[12] Wen H, Liu Z. Separating fractal and oscillatory components in the power spectrum of neurophysiological signal. *Brain Topography*. 2016;29(1):13–26.

[13] Cong F, Lin QH, Kuang LD, Gong XF, Astikainen P, Ristaniemi T. Tensor decomposition of EEG signals: A brief review. *Journal of Neuroscience Methods*. 2015;248:59–69.

[14] Miwakeichi F, Martínez-Montes E, Valdés-Sosa PA, Nishiyama N, Mizuhara H, Yamaguchi Y. Decomposing EEG data into space-time-frequency components using parallel factor analysis. *NeuroImage*. 2004;22(3):1035–1045.

[15] Rošťáková Z, Rosipal R. Determination of the number of components in the PARAFAC model with a nonnegative tensor structure: A simulated EEG data study. *Neural Computing and Applications*. 2022;34(17):14793–14805.

[16] Lotte F et al. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*. 2018;15(3):031005.

[17] Rosipal R, Porubcová N, Cimrová B, Farkaš I. Mirror-therapy as a way to start bci robot-assisted rehabilitation: A single case longitudinal study of a patient with hemiparesis. In: 7th International BCI Meeting: BCIs: Not Getting Lost in Translation, Asilomar Conference Center, Pacific Grove, California, USA. 2018, 113.

Unprotected Minds: A Clinical Neuropsychology Perspective on the Ethics of Emerging Neurotechnologies

A.M. Lupaşcu¹, M.L. Flórez Rojas², S. Enriquez-Geppert¹

¹ Department of Clinical and Developmental Neuropsychology, University of Groningen, Groningen, Netherlands

² Faculty of Law, University of Groningen, Groningen, Netherlands

E-mail: a.lupascu.2@student.rug.nl

ABSTRACT: The rapid expansion of EEG-based neurotechnology from clinical and research settings into consumer and workplace environments has created new ethical challenges regarding the interpretation and governance of neural data. While clinical applications operate under strict therapeutic intent and expert oversight, and research is similarly governed by rigorous ethical and regulatory frameworks, non-clinical uses often rely on simplified metrics, unclear consent processes, and inadequate safeguards. This review goes beyond a narrative approach by examining the ethical gaps that arise when EEG-based technologies operate across divergent contexts. Through a biomedical ethics framework, the analysis reveals a critical failure to protect both neural signals and their interpretation once these technologies move outside the clinic. Central to this issue is the critical role of the clinical neuropsychologists, whose expertise enables contextualizing the neural signals while preventing the oversimplification of EEG data, and the overgeneralization of probabilistic brain markers. The findings underscore implications for three key stakeholder groups: neuropsychologists, who can establish interpretative standards; the BCI community, which is critical for ensuring the scientific validity and transparency of EEG-inferred cognitive states; and policymakers, who must address the regulatory gaps by extending protections beyond neural data collection.

INTRODUCTION

The applications of neurotechnology are rapidly shifting from the sterile confinement of the clinical and research setting to the unregulated expanse of the consumer market. Historically, neurotechnology systems utilizing electroencephalography (EEG) appeared as specialized tools, one example being their use in neurorehabilitation to restore motor and cognitive functions in patients with neurological deficits [1]. These systems are defined as technologies that acquire and process neural signals to measure cognitive or symptom-related activity, or generate external control signals. Their operational modes are diverse, from real-time applications like neurofeedback and brain-computer interfaces (BCI) control to continuous EEG-based neurotechnology, which may occur with or without the

user's awareness [2].

However, with improvements in the accessibility of EEG driven by low-cost, wireless headsets, and dry-electrode sensors that eliminate the need for gels, these technologies have recently been propelled into non-clinical and non-research settings. Applications range from wellness devices, gaming, to even workplace surveillance [3, 4]. While these advancements promise a new era of human-computer interaction, where neurotechnology becomes integrated into daily life, they simultaneously expose a critical gap in ethical and legal protection, namely the privacy of the mind itself [5].

Addressing this gap, the current paper investigates the following research question: What are the ethical challenges associated with EEG-based neurotechnology systems in clinical applications compared to their application in non-clinical settings? To date, the discourse on “neurorights” has been led by legal scholars, but the perspective of clinical neuropsychology is urgently needed [6]. With their expertise in interpreting neural data and validating it through behavioral and contextual assessment, clinical neuropsychologists are uniquely positioned as specialists in diagnostics and cognitive evaluation. Their role extends beyond treatment, as they are not only clinicians who assess and interpret brain function but also critical advocates for the ethical protection of neural information, ensuring its responsible use and interpretation. Therefore, the aim of this paper is to identify the ethical gaps in the uses of these systems in these two diverging settings and argue for the inclusion of neuropsychological expertise in neuroethical policy-making.

METHODS

A narrative literature review was conducted to examine the ethical implications of EEG-based neurotechnology across clinical and non-clinical contexts. Papers were selected through searches in databases such as PsycINFO, PubMed, and Google Scholar using combinations of keywords (EEG, neurotechnology, BCI, consumer neurotechnology, and neuroethics). The identified literature was then analyzed using the biomedical ethics framework proposed by Beauchamp and Childress [7].

CLINICAL APPLICATIONS

EEG-based neurotechnology systems are primarily employed to research, assess, restore, or compensate for loss of function [1]. In this context, the purpose of EEG-based neurotechnology is grounded in its research or therapeutic intent. In clinical practice, these systems are useful in both diagnosis and intervention across a wide range of conditions, including epilepsy, stroke, attention-deficit/hyperactivity disorder (ADHD), and depression [8, 9]. For example, neurofeedback represents a thoroughly researched application in which individuals learn to modulate specific patterns of brain activity associated with cognitive or emotional functioning [10, 11]. A meta-analysis of ten studies has shown that neurofeedback can yield sustained symptom remission in the treatment of ADHD [12]. Similarly, BCIs have been developed to assist patients with impairments by translating neural signals into external actions, such as controlling a robotic arm to perform grasping movements [13].

Technical Validity and Interpretation of Brain Signals

Underlying these applications is the strict requirement that data acquisition must occur in highly controlled environments. This necessity can be further explained by the concept of signal-to-noise ratio which is the proportion of meaningful brain activity relative to unwanted interference [14]. In this context, noise refers to any non-neural electrical activity that contaminates recordings, such as power lines interference or biological artifacts like muscle tension or eye blinks [14]. Because the brain's electrical signal is naturally weak, artifacts create a low signal-to-noise ratio, where meaningful data can easily be drowned out by interference. Consequently, clinical settings must employ standardized protocols and carefully managed conditions to minimize these confounding variables, ensuring that the recorded activity reflects actual neural data rather than external noise [14].

Equally important is the role of expert interpretation because neural data does not carry intrinsic meaning and its interpretation is not intuitive. Instead, signals serve as indirect proxies for underlying brain function, requiring contextualized analysis by trained clinicians who understand both the limitations of the measurement and the clinical characteristics of the patient [14, 15]. These factors highlight that neural signals are probabilistic and context-dependent, not transparent windows into one's thoughts and intentions [16]. Ethical and scientific literature consistently emphasizes that current neurotechnology cannot reliably access beliefs, desires, intentions, or individual thoughts [17].

Regulatory Frameworks and Patient Protections

Because these technologies are used to measure and interpret brain activity, they are regulated as medical

devices and must demonstrate safety, efficacy, and clinical relevance. In the European Union (EU), this occurs under the Medical Device Regulation (MDR) 2017/745, which requires clinical evidence, risk assessment, and post-market surveillance [18].

In parallel with device regulation, data protection laws play a central ethical role. Under the EU General Data Protection Regulation (GDPR) 2016/679, neural data falls within the category of “special category data” requiring explicit justification, data minimization, and heightened safeguards [19]. These protections limit secondary uses of neural data and reinforce patient control over how their information is stored, processed, and disclosed. This is a crucial safeguard given the intimate nature of the brain-derived information [20]. While current technology cannot decode complex thoughts, it can nonetheless infer sensitive attributes such as emotional states, cognitive health, and precursors to neurological disease [5].

Informed consent constitutes an additional cornerstone of clinical EEG-based neurotechnology. Patients as well as research participants are typically informed about the purpose of monitoring, the nature of the data collected, potential risks, and anticipated benefits prior to participation [21]. Therefore, clinical guidelines emphasize the responsibility of clinicians to assess decision-making capacity, adapt explanations accordingly, and involve proxies or guardians when appropriate [22].

It is evident that clinical neuropsychologists occupy a distinctive position at the intersection of neural data and human vulnerability. Their specialized training allows them to interpret neural signals within the broader context of the individual, integrating behavioral, cognitive, and situational factors, thereby minimizing the risk of overinterpretation of probabilistic neural markers [23]. Within neuropsychological clinical practice and research, neural data is never treated as a standalone indicator of identity or intent but as one component of a comprehensive assessment focused on patient care and rehabilitation [17].

NON-CLINICAL APPLICATIONS

In recent years, EEG-based neurotechnology has increasingly migrated from research and clinical contexts into non-clinical environments, including workplaces and consumer markets. Technologies originally developed for diagnosis or treatment are now deployed to assess attention, fatigue, engagement, or cognitive performance in healthy individuals [3, 24]. This marks a fundamental shift in purpose, from understanding or restoring impaired function to measuring or improving cognition in everyday life. However, as these systems move into everyday environments, the ethical significance of who interprets neural data and for what purpose demands urgent attention.

In workplace settings, EEG-based neurotechnology has been introduced primarily in high-risk environments. For example, EEG-based headsets have been used in China

to detect in real-time fatigue, stress, or lapses in concentration [24, 25]. Similarly, in mining and construction industries, workers have been documented to use “smart helmets” that are designed to prevent accidents by detecting exhaustion before performance deteriorates [3]. In these contexts, EEG-based neurotechnology is often framed as protective rather than punitive, aiming to identify risks before harm occurs. However, the use of continuous neural surveillance in environments characterized by strong power asymmetries is ethically complex, as workers may have limited ability to reduce participation without facing negative consequences [24].

In the consumer market, the Muse headband is an application serving the goal of self-optimization. It is marketed as a personal meditation tool as the device utilizes dry sensor EEG to provide real-time neurofeedback by translating the user’s brain activity into sounds [26]. The device plays calm weather sounds during periods of focus, but switches to stormy winds when the user loses concentration. Consequently, this technology reframes mindfulness as a form of neuro-optimization for productive efficiency [27].

Technical Limitations and Data Interpretation

In non-clinical applications, concerns about validity and interpretations are particularly salient. Consumer-grade and workplace EEG devices rely on simplified processing of neural data that translates complex activity into reductive metrics, such as a single fatigue score [27]. Furthermore, because neural signals are highly sensitive to movement and environmental noise, non-clinical recordings are more susceptible to artifacts, making them less reliable than their clinical counterparts [28]. Considering these limitations, they are not suited to support decisions about performance, safety, or employability [30]. Moreover, the implications for well-being differ markedly from those observed in healthcare contexts. Continuous monitoring may foster stress and self-censorship, particularly when individuals are aware that their neural states are subject to external scrutiny [5].

Regulatory Gaps

The regulatory framework governing EEG-based neurotechnology changes substantially once these systems move outside clinical environments, creating an uneven level of protection across contexts. In non-clinical settings, the applicability of existing data protection frameworks, including the GDPR, may also become less clear when neural data is collected outside medical contexts [3, 19]. Moreover, while clinical applications fall under the MDR [18], consumer and workplace applications typically fall under broader consumer protection and product safety frameworks, such as the General Product Safety Directive (2001/95/EC) and the Unfair Commercial Practices Directive (2005/29/EC) [30, 31]. These frameworks address the prevention of unsafe products and misleading

commercial claims. However, they were not designed to regulate technologies that monitor neural activity, resulting in a lower regulatory threshold than in clinical contexts.

When artificial intelligence (AI) systems are used to infer cognitive states such as attention, fatigue, or emotional engagement, additional regulatory frameworks may also apply. The EU AI Act (Regulation EU 2024/1689) introduces obligations for high-risk AI systems, particularly in employment and biometric contexts [32]. However, when these systems are not specifically marketed as AI decision making technologies, it is uncertain how consistent this framework will be applied. Consequently, EEG-based neurotechnology that functions simultaneously as a wearable device and an algorithmic inference system may remain subject to fragmented regulatory oversight, despite producing inferences about cognitive states [6]. This raises concerns about both the reliability of neural interpretations and the potential consequences of decisions based on such data.

AN ETHICAL PERSPECTIVE

The moral acceptability of EEG-based neurotechnology depends not only on the usefulness of the technology itself, but on the conditions under which neural data is generated, interpreted, and acted upon. While clinical and non-clinical applications frequently rely on similar tools, they diverge fundamentally in intent, governance, and epistemic reliability. To evaluate these divergences, the following analysis adopts the framework of biomedical ethics proposed by Beauchamp and Childress [7], specifically examining how the transition from clinical to consumer settings challenges the principles of beneficence, non-maleficence, autonomy, and justice (Fig. 1). This medical ethics model was selected because neurotechnology systems originate in clinical practice and involve the measurement of brain activity, which has traditionally been governed by medical standards of care.

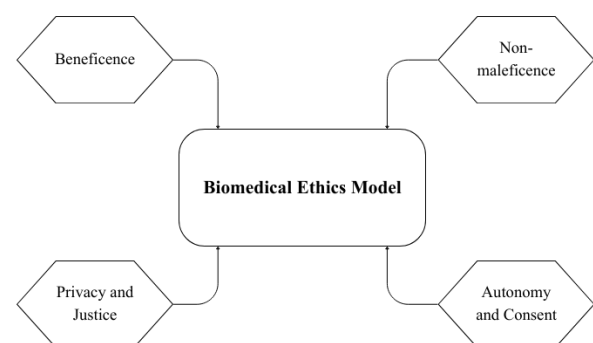


Figure 1: The four principles of the biomedical ethics model.

Beneficence

In clinical contexts, EEG-based neurotechnology systems are anchored in beneficence and epistemic restraint. Despite decades of research and some

efficacious protocols, neurofeedback remains highly variable in its effects, underscoring the need for personalized approaches and deeper investigation into the sources of this heterogeneity [33]. If uncertainty persists even under optimal conditions characterized by validated protocols and expert interpretation, then claims made by non-clinical applications operating outside these constraints warrant heightened skepticism. Unsupported claims based on neural signals may mislead users about their cognitive states and lead them to make decisions about their health based on unreliable information.

Non-Maleficence

This epistemic gap is illustrated by comparative research on two medical-grade and two consumer neurodevices. Ratti et al. [28] found that under specific controlled conditions and when interpreted by trained experts, one of the consumer EEG devices produced signals comparable to the medical-grade equipment. Crucially, this equivalence depended on the laboratory-level control and expert analysis, conditions which are rarely met in real-world consumer and workplace environments. This raises concerns regarding non-maleficence, as users may be penalized or judged based on technically flawed data. When imperfect algorithms generate false results, users face the risk of unjustified punishment based on flawed data.

Autonomy and Consent

Autonomy and consent further differentiate clinical from non-clinical uses. In research and healthcare, informed consent is treated as an ongoing ethical process, shaped by sensitivity to vulnerability and voluntariness [7]. Clinical neuropsychologists and researchers are trained to explain limitations and potential non-response to interventions such as neurofeedback, ensuring that patient's expectations remain realistic. In contexts outside of scientific research and clinical application, by contrast, consent is frequently embedded in employment contracts or terms of use. Thus, consent may be legally valid yet ethically compromised, as individuals face structural pressure that limits their ability to refuse to participate without social or economic cost [3].

Privacy and Justice

The principles of privacy and justice are linked to data ownership. In clinical settings, neural data is strictly defined as a medical record, owned by the patient and protected against discrimination. In consumer and workplace settings, however, neural data is often a proprietary asset, stripping individuals of ownership rights [5]. This shift creates risks of discrimination that can be used to deny employment [24]. Without the protection of medical confidentiality, the secondary use of neural data violates the principle of justice by prioritizing corporate interests over the fair treatment of the individual.

DISCUSSION

The present narrative review examined the ethical challenges that arise as EEG-based neurotechnology expands beyond clinical and research environments into consumer and workplace contexts. Through a biomedical ethics framework, it highlights the urgent need to incorporate the expertise of clinical neuropsychologists whose training in contextualizing neural data and validating it with behavioral and diagnostic insights is essential for mitigating risk of misinterpretation, overgeneralizing, and unethical application. Outside these frameworks, however, EEG data is only translated into reductive metrics. While the emerging neurorights discourse has emphasized protections for mental privacy and cognitive liberty, ethical risks arise also regarding the interpretation of the neural signals [5].

Implications for Neuropsychologists

Clinical neuropsychologists play an important role in ensuring that neural signals are interpreted within appropriate psychological and functional contexts. Their expertise lies in integrating neural indicators while recognizing the limits of neural data and the broader needs of the individual. In clinical practice, EEG is not a standard tool and, when used, plays a secondary role. Its findings are always interpreted alongside neuropsychological data, patient history, and environmental context. This contrasts sharply with the oversimplified or isolated use of EEG in non-clinical settings.

Although guidelines exist for neural interpretation within clinical settings, comparable standards are absent in non-clinical contexts. Neuropsychologists may therefore contribute by helping develop interpretative standards that prevent the overinterpretation of probabilistic neural markers for these emerging applications. In addition, their experience makes them well-suited to provide third-party evaluation when neural metrics are used to inform consequential decisions about individuals' performance, well-being, or cognitive functioning. Beyond clinical practice, their expertise may also contribute to the development of ethical guidelines and policy discussions concerning the interpretation and protection of neural data outside healthcare contexts.

Implications for the BCI Community

These developments carry important implications for the BCI community. As EEG-based neurotechnology becomes more integrated into consumer environments, the challenge lies not only in measuring neural signals but in ensuring that the interpretive claims derived from those signals remain scientifically justified. For the BCI community, maintaining the scientific integrity of neural inferences when EEG leaves the laboratory therefore becomes a shared concern.

In this regard, the principles of open and reproducible science offer a useful framework. Systems that claim to

infer cognitive states should be held to standards comparable to those applied within scientific research. This includes communicating the evidentiary basis for neural inference claims as well as documenting the interpretative pipeline from signal acquisition and preprocessing to algorithmic classification and behavioral inference [34]. As neurotechnology increasingly moves beyond the laboratory, aligning real-world applications with the norms of scientific validation will be essential for maintaining public trust and ensuring BCIs truly benefit the users.

Implications for Policy Makers

Neurotechnology governance should move beyond regulating the collection of neural data and address how such data is interpreted and used to generate inferences about an individual's cognitive states. Current regulatory frameworks remain fragmented, with device safety, data protection, and algorithmic systems governed through separate legal protections. While regulations such as the GDPR provide strong protections for health-related data, neural data generated outside the clinic may be susceptible to different standards [3, 19]. Policymakers should therefore clarify the legal status of neural data and ensure that protections extend to neural inferences and downstream decision-making, particularly in workplace and consumer environments where individuals have limited control over how their cognitive data are interpreted or used. In developing such governance frameworks, policymakers should engage with scientific or clinical experts who possess the expertise necessary to evaluate the validity and limitations of EEG-based systems and the interpretations derived from them.

CONCLUSION

This review highlights that the migration of EEG-based neurotechnology from research and clinical environments into consumer and workplace contexts has created a significant ethical divide. Within clinical practice, neural signals are interpreted within structured frameworks that emphasize therapeutic intent, standardized protocols, and expert oversight. In non-clinical and non-research applications, neural signals are translated into simplified metrics that claim to represent complex cognitive states.

This shift introduces ethical risks that extend beyond questions of data privacy. Neural signals are inherently probabilistic and context-dependent, yet consumers and workplace applications may present them as clear indicators of attention, fatigue, or mental performance. When such interpretations guide decisions about productivity or well-being, the absence of clinical expertise and robust validation becomes ethically problematic.

Addressing these challenges requires collaboration between scientific, clinical, and regulatory communities. Neuropsychologists can contribute expertise in the holistic and contextual interpretation of neural signals,

the BCI community plays a key role in maintaining standards for the scientific validity of EEG systems and the neural inferences derived from them, and policy makers should ensure that the regulatory frameworks address not only the collection of neural data, but also the interpretations and decisions derived from it.

REFERENCES

- [1] Chaudhary U, Birbaumer N, Ramos-Murguialday A. Brain–computer interfaces for communication and rehabilitation. *Nat Rev Neurol* 2016;12(9):513–525
- [2] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain–computer interfaces for communication and control. *Clin Neurophysiol* 2002;113(6):767–791
- [3] Muhl E. The challenge of wearable neurodevices for workplace monitoring: an EU legal perspective. *Front Hum Dyn* 2024;6
- [4] Pauzauskie S, Genser J, Yuste R. Protecting Neural Data Privacy—First, do no harm. *JAMA Neurol* 2024;82(3):212
- [5] Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc and Policy* 2017;13(1):5
- [6] Ienca M. On Neurorights. *Front Hum Neurosci* 2021;15:701258
- [7] Beauchamp TL, Childress JF. *Principles of Biomedical ethics*. OUP USA, (2013)
- [8] Arns M, Heinrich H, Strehl U. Evaluation of neurofeedback in ADHD: The long and winding road. *Biol Psychol* 2014;95:108–115
- [9] Ros T, Baars BJ, Lanius RA, Vuilleumier P. Tuning pathological brain oscillations with neurofeedback: a systems neuroscience framework. *Front Hum Neurosci* 2014;8:1008
- [10] Smit D, Dapor C, Koerts J, Tucha OM, Huster RJ, Enriquez-Geppert S. Long-term improvements in executive functions after frontal-midline theta neurofeedback in a (sub)clinical group. *Front Hum Neurosci* 2023;17:1163380
- [11] Thibault RT, Lifshitz M, Raz A. The self-regulating brain and neurofeedback: Experimental science and clinical promise. *Cortex* 2016;74:247–261
- [12] Van Doren J, Arns M, Heinrich H, Vollebregt MA, Strehl U, Loo SK. Sustained effects of neurofeedback in ADHD: a systematic review and meta-analysis. *Eur Child Adolesc Psychiatry* 2019;28(3): 293–305
- [13] Hochberg LR, Bacher D, Jarosiewicz B, Masse NY, Simeral JD, Vogel J et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* 2012;485(7398):372–375
- [14] Schomer DL, Da Silva FHL. *Niedermeyer's electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, (2018)
- [15] Tatum W, Rubboli G, Kaplan P, Mirsatari S, Radhakrishnan K, Gloss D, et al. Clinical utility of EEG in diagnosing and monitoring epilepsy in

- adults. *Clin Neurophysiol* 2018;129(5):1056–1082
- [16] Poldrack R. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci* 2006;10(2):59–63
- [17] Farah MJ, Hutchinson JB, Phelps EA, Wagner, AD. Functional MRI-based lie detection: scientific and societal challenges. *Nat Rev Neurosci* 2014; 15(2):123–131
- [18] European Parliament and the Council of the European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical Devices. *Official Journal of the European Union* 2017;L177:1-175
- [19] European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* 2016;L119:1-88
- [20] Ienca M, Malgieri G. Mental data protection and the GDPR. *J Law Biosci* 2022;9(1):lsac006
- [21] Burwell S, Sample M, Racine E. Ethical aspects of brain computer interfaces: a scoping review. *BMC Med Ethics* 2017;18(1):60
- [22] Appelbaum PS, Grisso T. Assessing patients' capacities to consent to treatment. *N Engl J Med* 1988;319(25):1635–1638
- [23] Lezak MD, Howieson DB, Bigler ED, Tranel D. *Neuropsychological assessment*. Oxford University Press, (2012)
- [24] Farahany N. *The battle for your brain: Defending the Right to Think Freely in the Age of Neurotechnology*. St. Martin's Press, USA (2023)
- [25] Lopez CAF, Li G, Zhang D. Beyond Technologies of Electroencephalography-Based Brain-Computer Interfaces: A Systematic Review from Commercial and Ethical aspects. *Front Neurosci* 2020;14:611130
- [26] Krigolson OE, Williams CC, Norton A, Hassall CD, Colino FL. Choosing MUSE: Validation of a Low-Cost, Portable EEG system for ERP research. *Front Neurosci* 2017;11:109
- [27] Wexler A, Thibault R. Mind-Reading or misleading? Assessing Direct-to Consumer electroencephalography (EEG) devices marketed for wellness and their ethical and regulatory implications. *J Cogn Enhanc* 2018;3(1):131–137
- [28] Ratti E, Waninger S, Berka C, Ruffini G, Verma A. Comparison of medical and consumer wireless EEG systems for use in clinical trials. *Front Hum Neurosci* 2017;11:398
- [29] Wexler A, Reiner PB. Oversight of direct-to-consumer neurotechnologies. *Science* 2019;363(6424):234–235
- [30] European Parliament and the Council of the European Union. Directive 2001/95/EC on general product safety. *Official Journal of the European Communities*, 2001
- [31] European Parliament and the Council of the European Union. Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market. *Official Journal of the European Union*, 2005
- [32] European Parliament and the Council of the European Union. Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, 2024
- [33] Rogala J, Jurewicz K, Paluch K, Kublik E, Cetnarski R, Wróbel A. The Do's and Don'ts of Neurofeedback Training: A review of the controlled studies using healthy adults. *Front Hum Neurosci* 2016;10:301
- [34] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Du Sert NP, et al. A manifesto for reproducible science. *Nat Hum Behav* 2017;1(1):0021

Circadian alignment modulates mental tasks BCI performance

Pauline Dreyer¹, Simon Kojima¹, Manon Bourdil¹, Loic Bechon¹, Sebastien Rimbert¹, Raphaëlle N. Roy³, Fabien Lotte^{1, 2}

¹Inria Center at the University of Bordeaux, Talence, 33405, France

²LaBRI (Univ. Bordeaux / CNRS/ Bordeaux INP) France

⁴Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France

pauline.dreyer@inria.fr

ABSTRACT: Active brain–computer interfaces (aBCIs) exhibit substantial performance variability both across and within participants. While many sources of variability have been investigated, circadian influences remain largely unexplored in aBCI control. We investigated whether time of day and chronotype (individual’s circadian preference for sleep and activity timing across the day) jointly modulate BCI performance (classification accuracy). Twelve participants completed 6 EEG-based BCI sessions each alternating between morning and afternoon sessions, performing 3 mental tasks (motor imagery, mental calculation, letter/word association). Online performance was analyzed using generalized linear mixed models with binomial distribution and random intercepts for participants and sessions. Overall performance was above chance level, with a significant decrease across runs but no linear trend across sessions. A significant interaction between time of day and chronotype was observed: morning-type participants tended to perform better in the morning, whereas evening-type participants showed relatively higher performance in the afternoon, consistent with the circadian alignment.

INTRODUCTION

Active brain–computer interfaces (aBCIs) allow users to voluntarily modulate their brain activity to control external devices, often via electroencephalography (EEG) [1]. While highly promising for assistive and rehabilitation applications, BCIs are still limited by substantial performance variability [2], both across users and within the same user across sessions [3]. The current literature suggests that performance variability arises from multiple interacting sources [4], including experimental settings such as task design, feedback modality, electrode montage or environmental context [5, 6], as well as stable individual characteristics such as cognitive abilities [7], motor imagery capacity, personality traits, and structural brain properties [8]. Physiological artifacts [9], or psychological states (fatigue, attention, motivation) [10] and intrinsic neurophysiological dynamics [11] (e.g., spontaneous oscillatory power and connectivity fluctuations) further contribute to performance variability over time. Despite this growing understanding of performance vari-

ability, one potentially important factor has received little attention in active BCI research: circadian modulation [12]. Human cognition is regulated by the interaction between homeostatic sleep pressure, which increases with time spent awake, and endogenous circadian processes that follow a near 24-hour rhythm [13]. Together, these mechanisms create daily fluctuations in alertness, attention, inhibition, and working memory. As a result, cognitive performance is not stable from morning to evening. Importantly, these daily changes do not affect all individuals in the same way. The chronotype reflects an individual’s preferred timing of activity and alertness, with some people functioning optimally in the morning and others later in the day [14, 15]. When task timing aligns with chronotype, performance tends to improve and this alignment is known as the synchrony effect [16]. A recent systematic review by Chauhan et al., highlights that this synchrony effect is frequently observed in cognitive tasks involving attention and executive control in healthy adults, and appears to be even more pronounced in older populations, with up to 83% of studies reporting such an effect [17, 18]. Other experimental studies also showed neurophysiological correlates of this effect, indicating that time-of-day and chronotype are associated with measurable changes in EEG oscillatory activity and ERP dynamics [19, 20]. Salehinejad et al.[21] further reported that cortical excitability and neuroplasticity show circadian modulation depending on chronotype, and that performance in several cognitive domains, including motor learning, working memory, and attention, tends to be optimal when tasks are performed at an individual’s preferred time of day.

These findings are highly relevant for active BCI control which requires sustained attention, mental engagement, and reliable modulation of sensorimotor rhythms [22]. All these processes are sensitive to the ongoing brain state and inherently variable throughout the day, leading to a fluctuation of BCI performance accordingly. Despite this theoretical coherence, circadian factors remain largely unexplored in active BCI studies. Most BCI experiments are conducted at a single, often unspecified, time of day without considering chronotype and time-of-day alignment. One of the few study addressing this question is Aricò et al. (2012), who reported preliminary

evidence that chronotype may modulate P300-BCI performance depending on time of day [23].

Given the gap identified in the current literature, and the limited investigation of circadian influences in active BCI research, the present study investigates whether time of day and chronotype modulate active BCI performance. We designed a repeated-measures protocol in which participants completed both morning (AM) and afternoon (PM) sessions on different days. Chronotype was assessed using the reduced French version of the Morningness–Eveningness Questionnaire (MEQ) [14, 24], and the wake up hour was self reported before each session. BCI performance resulted of the classification accuracy of three different mental tasks. These results were analyzed using generalized linear mixed models to account for the hierarchical and repeated-measures structure of the data. We hypothesized that BCI performance would not depend only on time of day, but on its interaction with chronotype, reflecting a synchrony effect in active BCI control. The remainder of this paper presents the experimental protocol, the statistical modeling approach, results, and implications of circadian influence as a potential factor of variability in active BCI performance.

MATERIALS AND METHODS

Participants:

This paper reports the results obtain from twelve (12) participants who completed the full BCI experiment protocol, (5 women / 7 men; age 19-44, $M = 29.2$, $SD = 7.4$). They were required to be between 18 and 60 years old, have normal or corrected vision, and report no history of neurological or psychiatric conditions. No authority relationship existed between the experimenters and the volunteers. All procedures followed the relevant ethical research guidelines set out in the Declaration of Helsinki [25]. Each participant signed a consent form confirming their agreement to participate and to allow the use of their EEG recordings and behavioral data. The research protocol received approval from the Inria ethics committee, the COERLE (Approval number: 2023-13).

EEG data acquisition:

Participants were seated comfortably in front of a computer screen during the experiment. Brain signal activity was recorded using 41 active scalp electrodes positioned according to the international 10–10 system (AF3, AF4, F5, F3, F1, Fz, F2, F4, F6, FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P5, P3, P1, Pz, P2, P4, P6, PO7, PO8). Signals were referenced to the left mastoid, while the ground electrode was placed at the FPz location. Electromyographic (EMG) activity was recorded with 2 active electrodes from both hands (EMGr and EMGI) and electrooculography (EOG) signals was recorded with four active electrodes: two placed above the eyes (FP1 and FP2) and two located below and lateral to the left eye (EOG2 and EOG3). All signals were acquired using an actiCHamp amplifier sys-

tem (Brain Products GmbH, Gilching, Germany) with a sampling rate of 500 Hz. Online signal processing and experimental control were implemented using the open-source BCI software OpenViBE 3.6.0 [26]. The vast majority of electrodes had an impedance below 20k Ω , with a few more difficult electrodes occasionally reaching higher impedances. Throughout each run, the experimenters continuously monitored the EEG signals to ensure stable acquisition quality at all time.

Experimental design:

Each participant completed six experimental sessions conducted on different days in the same room. Sessions alternated between morning (AM) and afternoon (PM) in a counterbalanced order to investigate potential circadian influences on BCI performance. Morning sessions were conducted between approximately 9:30 and 12:10 (mean = 10:48, $SD = 1.33$ h), corresponding to the start of BCI data acquisition following approximately one hour of preparation (equipment setup, instructions, and baseline recordings). Afternoon sessions began between approximately 13:10 and 15:40 (mean = 14:24, $SD = 1.24$ h). Chronotype was assessed prior to the experiment using the reduced French version of the Morningness–Eveningness Questionnaire (MEQ) [14, 24]. Before each session, participants reported their wake-up time. This information was used to compute the duration between wake-up and the start of the BCI acquisition (Time Awake). This variable was treated as a self-reported proxy of time since awakening, providing an indirect estimate of the homeostatic sleep pressure component, which has been shown to influence EEG activity and cognitive performance [13]. Each session included six runs counterbalanced between two experimental contexts. The first corresponded to the A) Classical Graz paradigm, a low-stimulation visual interface [27]. The second environment was a gamified interface, B) Brain-Hero, developed within our research group, enriched with visual elements and background auditory noise designed to approximate real-world conditions. These two contrasted interfaces allowed us to experimentally manipulate the environmental context and assess its potential influence on BCI performance.

The session began with the acquisition phase, with 2 acquisition runs with sham feedback. These 2 runs were used to compute three-band Common Spatial Pattern (CSP) (in theta, alpha, beta bands) and then classified using Linear Discriminant Analysis (LDA) before online control. The training phase then followed and participants performed 4 online runs with real-time feedback: two runs in the Graz interface and two with Brain Hero interface in counterbalanced order across sessions. Within each context, participants performed three mental tasks: (i) kinesthetic motor imagery: Imagination of performing specific non dominant hand movements (e.g., clenching your fist) while focusing on all the sensations associated with the real movement. (ii) mental calculation: mental subtraction tasks of a given subtraction and (iii) letter/word association: generating as many

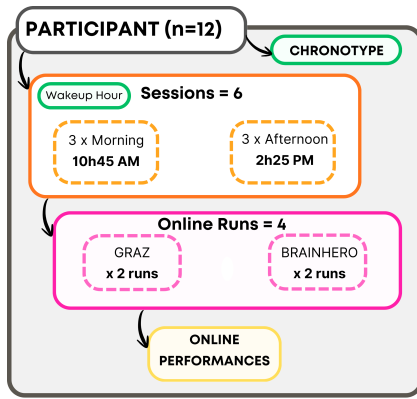


Figure 1: Experimental protocol of the BCI experiment.

words as possible starting with a given letter. Each run consisted of 30 trials (10 per class). Each trial began with a visual cue displayed for 3 s that indicated the required mental task, and participants were instructed to perform the task continuously during the 5-s feedback period. However, due to a technical issue, for some sessions, classifier training was performed using data from the Graz run only. Consequently, contextual differences in this subset may partly reflect calibration constraints rather than purely interface-related effects. Importantly, participants still performed runs in both contexts during the experiment, only the classifier calibration relied exclusively on the Graz data for some sessions.

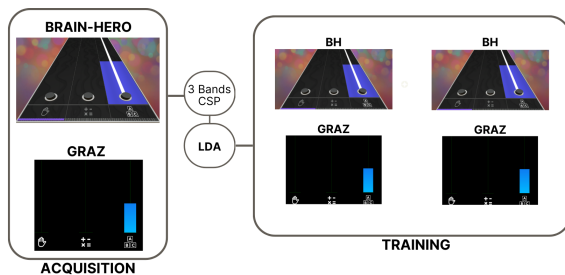


Figure 2: Overview of the acquisition and training procedure. During the acquisition phase, participants performed runs in both contexts (BrainHero and Graz), with sham feedback. EEG data from the acquisition runs were used to compute spatial filters using a three-band CSP and then classified using LDA. For a subset of sessions, classifier training relied on data from the Graz run only but then applied online in both contexts during the training phase with real feedback.

BCI pipeline and performance metric:

The classification model was trained separately for each participant and session using data from both BrainHero and Graz acquisition runs. As mentioned above, in some sessions the model was trained using only the Graz run. First, 39 EEG channels were selected, excluding TP7 and TP8 from the available EEG channel pool. No artifact removal was applied. A fifth-order Butterworth band-pass filter was applied in the forward direction to obtain data in three frequency bands (theta: 4–8 Hz, alpha: 8–12 Hz, and beta: 12–30 Hz). The following processing steps were applied independently to each frequency band.

EEG signals for each task were segmented into epochs. For model training, epochs from 0.5–9.5 s after the onset of the visual cue of each task were extracted. In contrast, for the online experiment, data from 3.0–8.0 s after the visual cue onset (i.e., feedback presentation period) were used.

Next, regularized CSP filters [28] (shrinkage coefficient: 1×10^{-6} , without Tikhonov regularization and with trace normalization) were trained. Since a three-class classification was performed, CSP was trained and applied using a one-vs-all approach, with two filters (one pair) per class. The CSP components of each epoch were then segmented into sub-epochs using a sliding window with a length of 1 s and a step size of $1/16$ s. The data of each sub-epoch were transformed into mean log-power values. Next, the data from the three frequency bands were concatenated for each sub-epoch and used as feature vectors for shrinkage-regularized Linear Discriminant Analysis (LDA) training and classification. These resulted in 18 features per sub-epoch (2 spatial filters, for each of the 3 classes and each of the 3 frequency bands). LDA was trained and applied using the one-vs-all approach. The class of each trial was estimated as follows. For each sub-epoch $n \in \{1, \dots, N\}$, let $a_{n,k}$ denote the probability output of the LDA classifier for class k vs. all. Then, $p_{n,k} = (a_{n,k}) / \sum_{k=1}^N a_{n,k}$ was defined. Using the $p_{n,k}$ obtained from the 128 sub-epochs, the class label of each trial was estimated as $y = \operatorname{argmax}_k \sum_{n=1}^N p_{n,k}$. Trial-wise classification accuracy was the used performance metric.

Statistical modeling:

Statistical analyses aimed to identify the factors associated with variability in online BCI performance, while accounting for the hierarchical and repeated structure of the data. In particular, the analyses tested whether temporal factors (run and session progression, time of day), and participants characteristics (chronotype) were associated with changes in online BCI performance. They were conducted in R (version 4.5.2) implemented in the lme4 package (R). Online BCI performance was modeled using generalized linear mixed models (GLMMs) with a binomial distribution and logit link, appropriate for modeling the number of correct classifications out of a fixed number of trials [29, 30].

The dependent variable, Online BCI performance, was defined as the number of correctly classified trials out of the total number of trials per run (30 trials), modeled as a binomial outcome. The fixed-effects structure included experimental factors defined by the protocol. These included experimental context (BrainHero vs. Graz), run index (within-session progression), session index (between-session progression), Time Awake (TimeAwake: time between waking up and the start of BCI acquisition, within-subject centered), and the interaction between time of day (AM vs. PM) and chronotype (MEQ score, standardized), to test the synchrony effect. Random-effects structures were selected using a hierarchical model-comparison approach based on likelihood-ratio tests (LRTs) and Akaike Information Criterion

(AIC) [29, 30]. The baseline model (Table 1: P0), specified in R as $(1|SUBJECT) + (1|SUBJECT : SESSION)$, included random intercepts for participants and sessions nested within participants to account for the repeated-measures structure of the protocol. Because the study specifically aimed to investigate potential intra-user variability in performance dynamics across runs, a theoretically motivated model including a random slope for run at the participant-session level was additionally tested (Table 1: Prunin). This model assessed whether the effect of run progression varied across sessions within participants. However, it did not significantly improve model fit relative to the baseline model. Additional random slopes for run progression and session progression at the participant level were also explored (Table 1: Prun and Psess, respectively), but neither structure provided a significant improvement in model fit.

Table 1: Comparison of candidate random-effects structures for the GLMM analyses.

Model	AIC	logLik	LRT (χ^2 , p)	Decision
P0	1520.6	-750.31	—	Retained
Prunin	1521.1	-748.57	3.47, $p = 0.18$	Rejected
Psess	1521.5	-749.73	1.15, $p = 0.28$	Rejected
Prun	1522.6	-750.31	0.00, $p = 1$	Rejected

Model adequacy was evaluated using simulation-based residual diagnostics (DHARMA package) and confirmed adequate fit. Multicollinearity among fixed effects was assessed using variance inflation factors (VIFs), all point estimates were below 2.1, indicating no meaningful collinearity. Profile likelihood confidence intervals were computed for fixed effects.

RESULTS

The final model retained random intercepts for participants and sessions nested within participants (Table 1: P0), capturing both inter-user variability in BCI performance and session-to-session variability within participant while providing a parsimonious representation of the hierarchical structure of the data [29]. Overall the average BCI performance from 12 participants across session is above chance level 46.67% at $p < 0.001$ (see Fig. 3). The global median performance reached 66.7% indicating that most participants achieved reasonable control of the system. Because classifier training was partially performed on Graz's run only for a subset of sessions (see: Experimental design), the context was not interpreted as a main outcome variable, but retained in the statistical model as a fixed effect to control for structured sources of variance [29].

Fixed-effect estimates are reported as log-odds coefficients (β), together with odds ratios (OR) and their 95% confidence intervals (Table 2). OR values above 1 indicate increased odds of correct BCI classification, whereas OR values below 1 indicate decreased odds of success.

Table 2: Fixed effects estimates from the GLMM predicting BCI performance. Regression coefficients (β) are reported on the log-odds scale, with corresponding odds ratios (OR) and 95% confidence intervals. Significant effects (95% CI not including 1 and $p < .05$) are highlighted in bold. Positive and negative coefficients indicate increased or decreased odds of success respectively.

Predictor	β	OR	95% CI (OR)	p
Context (Graz)	0.246	1.28	[1.17, 1.40]	<.001
run	-0.062	0.94	[0.90, 0.98]	.004
session	0.040	1.04	[0.99, 1.10]	.144
TimeAwake (z)	-0.094	0.91	[0.80, 1.04]	.162
TIME (PM)	0.130	1.14	[0.87, 1.48]	.338
MEQ (z)	0.121	1.13	[0.84, 1.51]	.416
TIME (PM) $\times$ MEQ(z)	-0.283	0.75	[0.63, 0.91]	.003

Within-session progression showed a significant negative linear effect of run (OR = 0.94, 95% CI [0.90, 0.98], $p = .004$; Table 2), reflecting a progressive decrease in BCI performance across successive runs within a session (see Fig. 3).

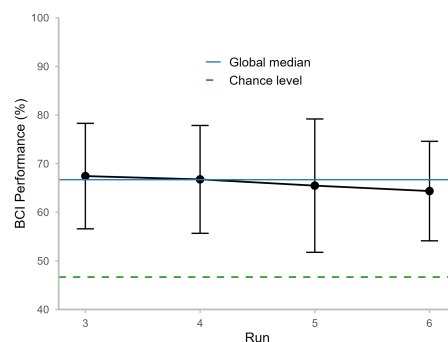


Figure 3: Mean BCI performance for each of the 4 runs in all BCI sessions across the 12 participants. The blue line correspond to the global median (66.7%), and the green one is the chance level (46.67% at $p < 0.001$)

In contrast, no significant linear effect of session was observed (OR = 1.04, 95% CI [0.99, 1.10], $p = .144$). To isolate within-participant variability across sessions, Time awake was centered within each participant and standardized before inclusion in the model. The main effect of time awake was not statistically significant (OR = 0.91, 95% CI [0.80, 1.04], $p = .162$). Neither the main effect of Time of day (PM vs AM) (OR = 1.14, 95% CI [0.87, 1.48], $p = .338$) nor chronotype (MEQ) (OR = 1.13, 95% CI [0.84, 1.51], $p = .416$) reached statistical significance. However, a significant Time $\times$ MEQ interaction was observed (OR = 0.75, 95% CI [0.63, 0.91], $p = .003$).

To facilitate interpretation of this interaction in the GLMM, predicted probabilities were computed on the response scale and visualized across a continuous range of chronotype values (see Fig. 4). This significant interaction indicates that the effect of time of day modulates BCI performance depending on individual chronotype. Specifically, the difference between PM and AM performance becomes increasingly negative as MEQ scores increase. For participants with higher MEQ scores (oriented morning-type), performance tended to be lower in the afternoon compared to the morning. In contrast, par-

ticipants with lower MEQ scores (oriented evening-type) showed relatively higher performance in the afternoon compared to the morning.

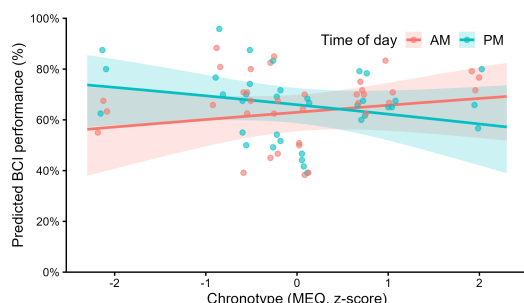


Figure 4: Predicted probability of BCI performance as a function of chronotype (MEQ, z-score) and time of day (AM vs PM), derived from the glmm. Lines represent predicted probabilities (fixed effects only), and shaded areas indicate 95% confidence intervals. The crossing of the AM and PM curves reflects the interaction between chronotype and time of day.

DISCUSSION

The present study investigated whether time of day and chronotype contribute to participant variability in aBCI performance. Using a repeated-measures design with alternating morning and afternoon sessions, we observed that average BCI performance was reliably above chance across participants and sessions, confirming that participants were able to achieve effective BCI control. In line with [31], combining mental tasks considered by the authors as "brain teaser" such as mental calculation or word association with more "dynamic imagery tasks" such as motor imagery, could yield higher feature discriminability. Moreover, in a three-class BCI protocol, the probability that at least one pair of tasks produces separable EEG patterns increases, which may partly explain the above, chance performance observed here for all users.

We observed a small linear decrease in performance across runs within sessions, while no clear linear trend emerged across sessions on different days. The within session decline is compatible with accumulating mental fatigue due to the repeated mental tasks, reduced engagement, and is in line with [32], who reported a within-session decrease in motor imagery EEG separability associated with increasing mental fatigue. The absence of a linear improvement across sessions in this preliminary sample could be explained by the classifier training procedures. The classifier was recalibrated at each session, possibly interfering with a potential learning effect across days. Consequently, the lack of a significant session effect should not necessarily be interpreted as the absence of learning, but rather as a limitation inherent to session-dependent classifier training [33].

The main finding is a significant interaction between time of day and chronotype, whereas neither of them individually showed a robust main effect. More morning-type participants tended to perform better in the morning than

in the afternoon, and more evening-type tended to perform better in the afternoon. This result is consistent with the "synchrony effect" described in cognitive psychology, whereby cognitive performance peaks when tasks are performed at an individual's circadian preferred time of day [13, 16]. Here, the predicted probability curves for AM and PM performance cross at intermediate MEQ scores, which is the characteristic signature of this effect and confirms that time-of-day modulation is not uniform across participants but depends on their chronotype. Our findings extend this effect to aBCI, suggesting that BCI control may be optimized when experimental timing is aligned with a participant's chronotype. The potential benefit of scheduling BCI sessions according to participant chronotype, may be particularly relevant for older adults [18] or in clinical populations where reliable control is critical.

Several limitations should be noted. The current results rely on a small sample ($n = 12$ participants, six sessions each), which limits power and generalizability, despite the repeated-measures design. Larger samples are needed to robustly confirm these effects and test interactions with other psychological or physiological factors. Second, while the reduced version of the MEQ questionnaire is validated and widely used, it reflects subjective behavioral preference rather than a precise marker of biological circadian timing [14, 24]. Additional sleep-related factors, including sleep quality and sleep debt, may also influence performance variability alongside circadian effects. Future studies should therefore implement better control of sleep-wake history. Finally, with the rich database collected here, further analyses could integrate multimodal physiological markers (e.g., ocular activity, heart rate variability) [34], which may provide complementary insights into performance variability and user experience [35]. Additional neurophysiological analyses could also explore how oscillatory activity varies with time of day and chronotype.

CONCLUSION

Our results suggest that circadian alignment between session's timing and individual chronotype contributes to intra participant variability in active BCI performance. Neither time of day nor chronotype alone are a robust main effect. Instead, their interaction, consistent with the synchrony effect, appeared to be the main factor associated with performance variability across sessions. These findings highlight the importance of considering chronotype in the design of BCI experiments, alongside factors that are more commonly controlled such as alertness, engagement or motivation. Taking chronotype in to account may offer a yet underexplored factor to improve user-centered BCI design and adaptive strategies. This consideration is particularly relevant in clinical contexts, where the reliability of BCI control at each session is essential.

ACKNOWLEDGMENTS

This work was funded by ANR project PROTEUS (grant ANR-22-CE33-0015-01). The experiment background sound environment was created by Bernadette Thiboud, a professional film sound editor. We thank all participants for their time and involvement in the study.

REFERENCES

- [1] McFarland D, Wolpaw J. EEG-based brain-computer interfaces. *Cur Opin Biomed Eng.* 2017;4.
- [2] Fairclough SH, Lotte F. Grand challenges in neurotechnology and system neuroergonomics. *Front. Neuroergo.* 2020.
- [3] Saha S, Baumert M. Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review. *Front Comp Neuro.* 2020;13.
- [4] Dreyer P, Roy RN, Lotte F. Building a taxonomy of variability factors in active BCI. *Int. BCI Meeting.* 2025.
- [5] Vukelić M, Gharabaghi A. Oscillatory entrainment of the motor cortical network during motor imagery is modulated by the feedback modality. *NeuroImage.* 2015.
- [6] Roc A et al. A review of user training methods in brain computer interfaces based on mental tasks. *J Neur Eng.* 2021.
- [7] Jeunet C, N’Kaoua B, Subramanian S, Hachet M, Lotte F. Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns. *PloS one.* 2015;10(12):e0143962.
- [8] Gong J et al. White matter connectivity pattern associate with characteristics of scalp eeg signals. *Brain Topogr.* 2017;30(6):797–809.
- [9] Nann M et al. Heart rate variability predicts decline in sensorimotor rhythm control. *J. Neur. Eng.* 2021;18(4).
- [10] Myrden A, Chau T. Effects of user mental state on eeg-bci performance. *Front Hum Neuro.* 2015;9.
- [11] Vidaurre C, Haufe S, Jorajuría T, Müller KR, Nikulin VV. Sensorimotor functional connectivity: A neurophysiological factor related to BCI performance. *Frontiers in Neuroscience.* 2020;14:575081.
- [12] Egger J, Kostoglou K, Müller-Putz GR. Chrono-eeg dynamics influencing hand gesture decoding: A 10-hour study. *Scientific Reports.* 2024;14(1):20247.
- [13] Schmidt C, Collette F, Cajochen C, Peigneux P. A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology.* 2007;24(7):755–789.
- [14] Adan A, Almirall H, Horne & Östberg morningness-eveningness questionnaire: A reduced scale. *Personality and Individual Differences.* 1991;12(3):241–253.
- [15] Chauhan S, Norbury R, Faßbender KC, Ettinger U, Kumari V. Beyond sleep: A multidimensional model of chronotype. *Neuro Biobehav Rev.* 2023;148.
- [16] May CP, Hasher L. Synchrony effects in inhibitory control over thought and action. *J Exp Psych: Hum Percep & Perf.* 1998;24(2).
- [17] Facer-Childs ER, Boiling S, Balanos GM. The effects of time of day and chronotype on cognitive and physical performance in healthy volunteers. *Sports Medicine - Open.* 2018;4(1):47.
- [18] Chauhan S et al. Chronotype and synchrony effects in human cognitive performance: A systematic review. *Chronobiology International.* 2025;42(4):463–499.
- [19] Lara T, Madrid JA, Correa. The vigilance decrement in executive function is attenuated when individual chronotypes perform at their optimal time of day. *PLoS ONE.* 2014;9(2):e88820.
- [20] Venkat N, Sinha M, Sinha R, Ghate J, Pande B. Neuro-cognitive profile of morning and evening chronotypes at different times of day. *Ann. Neuro.* 2020.
- [21] Salehinejad MA et al. Cognitive functions and underlying parameters of human brain physiology are associated with chronotype. *Nat Commun.* 2021;12(1):4672.
- [22] Wolpaw J, Wolpaw EW. *Brain-computer interfaces: Principles and practice.* OUP USA (2012).
- [23] Aricò P et al. On the correlation between bci performance and chronotype. In: *GNB2012.* 2012.
- [24] Bessot N, Margas N. Development and validation of a french version of the morningness-eveningness questionnaire. *Chronob Int.* 2024;41(1):81–92.
- [25] World Medical A. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. *J Postgrad Med.* 2002;48.
- [26] Renard Y et al. Openvibe: An open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments. *Presence.* 2010.
- [27] Pfurtscheller G, Neuper C. Motor imagery and bci. *Proc. IEEE.* 2001.
- [28] Lotte F, Guan C. Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Trans Biomed Eng.* 2010;58(2):355–362.
- [29] Bolker BM et al. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology Evolution.* 2009;24(3):127–135.
- [30] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software.* 2015;67(1):1–48.
- [31] Friedrich EV, Scherer R, Neuper C. The effect of distinct mental strategies on classification performance for brain-computer interfaces. *Int J Psychophysiol.* 2012;84(1):86–94.
- [32] Talukdar U, Hazarika SM, Gan JQ. Motor imagery and mental fatigue: Inter-relationship and EEG based estimation. *J. Comput. Neurosci.* 2019;46:55–76.
- [33] Perdakis S, Millan JdR. Brain-machine interfaces: A tale of two learners. *IEEE Systems, Man, and Cybernetics Magazine.* 2020;6(3):12–19.
- [34] Georgaras E, Vourvopoulos A. Physiological assessment of multimodal bci training. *Res. Biomed. Eng. Technol.* 2025.
- [35] Dreyer P et al. Exploring EOG markers of fatigue during motor imagery BCI use. In: *9th Graz BCI Conference.* 2024.

Introducing simulation-based inference to c-VEP BCI: A feasibility study

Sébastien Velut^{1,2}, Marie-Constance Corsi³, Sylvain Chevallier², Frédéric Dehais¹, Jordy Thielen⁴

¹Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, Toulouse, 31055, France

²INRIA TAU - LISN - Université Paris-Saclay, Gif-sur-Yvettes, 91190, France

³CNRS, Inria, Inserm, AP-HP, Hopital de la Pitié Salpetriere, Paris, 75013, France

⁴Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, 6525 GD, Netherlands

E-mail: sebastien.velut@isae-supero.fr

ABSTRACT: Code-modulated visual evoked-potential (c-VEP) based reactive brain-computer interfaces (BCIs) deliver high information-transfer rates with minimal calibration, yet performance collapses when generalizing across participants. Additionally, only few studies focus on understanding the neural mechanisms underlying c-VEP, which remain poorly understood. In this paper, we introduce the simulation based inference (SBI) framework to the BCI field. SBI allows to offload calibration on simulated synthetic data generated with parameters that may represent a large cohort of different participants. We assess SBI's feasibility to provide competitive classification performance on empirical data as well as inferring response parameters, while being calibrated only on synthetic data. We achieved 82% classification accuracy with a within participant setup using a c-VEP based BCI, and a mean squared error of 0.11 between the empirical visual evoked potential and the inferred one. These results suggest the feasibility of SBI for BCI, and could deepen our understanding of the mechanisms underlying the c-VEP. We present the current limitations of the SBI framework and give future directions to address these.

INTRODUCTION

Among reactive brain-computer interfacing (BCI) paradigms, a recent paradigm, known as the code-modulated visual evoked potential (c-VEP), involves using pseudo-random noise sequences for visual stimulation. Such c-VEP based BCIs have gained more and more attention due to their high information-transfer rates, robustness, and reduced calibration requirements [1].

Despite these advantages, several challenges remain. First, performances often decrease when generalizing between sessions or participants due to inter/intra-subject variability in the EEG. This variability induces the need for calibration in every usage of a BCI. Second, the generative mechanisms underlying the c-VEP, that is the forward model, are still poorly understood, and studies typically rely on discriminative machine-learning models, that is backward models, such as template matching [1] or Riemannian methods associated with linear discriminant analysis [2]. In turn, explainability and physiologi-

cal modeling remain largely underexplored.

To address these challenges, a forward model, like recon-volution [3] can be used to perform inference, recovering response parameters given observed data. Such inference is not trivial, as there can be an infinite number of combinations of parameters that may have caused the observation, that is the problem is ill-defined.

A candidate solution is to use Bayesian methods such as Markov chain Monte Carlo (MCMC) [4] or variational inference [5]. Despite its robustness in arriving at good solutions, such algorithms typically rely on explicit likelihood evaluations, which are computationally expensive or even infeasible in a real-time system such as a BCI [6]. As an alternative to such iterative optimization strategies, simulation-based inference (SBI) [7, 8] embraces a forward model to generate a quasi infinite amount of labeled training data over a wide range of potential parameters. Then, an inverse model can be calibrated to recover these parameters based on the generated synthetic data. If successful, such an SBI framework yields a powerful inverse model that can recover parameters from any arbitrary empirical data, after an initial training period on synthetic data. This so-called amortization could relieve a BCI from its tedious calibration session.

SBI is a field within machine learning that is getting more and more attention because of its wide range of applications, such as in astrophysics and particle physics, and more recently in neuroscience [9–11].

SBI has not yet made its entry into the BCI field, despite its potential to mitigate or even solve several of the large challenges in the BCI field. Particularly, amortized inference could be useful in addressing the inter/intra-subject variability. Additionally, SBI may provide interpretable parameters improving the explainability of the EEG signal and understanding of underlying neural mechanisms. However, SBI depends on several factors that could impact its performance such as the simulation fidelity, the nature of the inference, the nature of the embedding net used before the inference, and the prior specification used to create simulated data [8], which raises non-trivial challenges for EEG applications.

In this study, we investigate the feasibility of applying SBI to c-VEP-based BCI decoding. Specifically, we

showcase the potential of an SBI framework and identify key challenges along the way. Doing so, we analyze both *simulated-to-simulated* and *simulated-to-empirical* scenarios to assess the adequacy and generalization behavior of the model. Beyond performance evaluation, this work explores the potential of SBI as a bridge between decoding accuracy and neural interpretability in c-VEP BCIs. This study does not aim to outperform existing c-VEP classifiers. Instead, we evaluate whether SBI constitutes a viable modeling framework for c-VEP responses by jointly assessing classification performance, parameter recovery, and model adequacy.

MATERIALS AND METHODS

Dataset: We used a previously collected empirical dataset called Ricker patch-based StAR c-VEP (ISAE-SUPAERO Dataset [12]), which received approval from the ethics committee of the University of Toulouse (CER approval number 2023-749) and was conducted in compliance with the Declaration of Helsinki. The EEG data were collected using a dry 8-electrode Enobio system, with a sampling rate of 500 Hz, positioned over the occipital and parieto-occipital sites (PO7, O1, Oz, O2, PO8, PO3, POz, PO4).

This dataset comprises 24 healthy individuals (4 women, 20 men, mean age = 29.3 years, SD = 7.5 years). Participants were instructed to direct their gaze to one of five disks indicated by a 0.5 s cue, after which the disks flashed for 2.2 s according to unique pseudo-random noise-codes. Each participant performed 15 runs, which consisted of 5 trials, one for each of the disks in random order. In summary, the dataset contains a total of 75 trials of 2.2 s, including 15 trials of per class ($N = 5$).

Pre-processing: First, we apply a finite impulse response bandpass filter between 1 Hz and 45 Hz. As the stimulus-onset asynchrony of individual flashes within the stimulus sequences exceeds 0.35 s, we create overlapping epochs of size 0.35 s every 1/500 s (the EEG sampling rate). This gives $E = 15 * (2.2 - 0.35) * 500 = 13875$ epochs per participant per class. Second, to ensure comparable amplitude over participants, we normalize the data by dividing the data by the standard deviation of each participant's training set. Third, to improve the signal-to-noise ratio (SNR) of the captured event-related potential (ERP) response to flashes of the c-VEP, and improve the quality of the signal, we add a spatial filtering step using xDAWN [13]. We select the 4 highest-ranked xDAWN components [14].

To align the different participants into the same feature space and reduce the variability between the subjects, we apply a Riemannian spatial alignment, an effective method for cross-subject analysis [15]. However, to perform the alignment, we first need to create super epochs by combining the average of the spatially filtered epoch with each spatially filtered target response. Let $\mathbf{E}_Z = \frac{1}{E} \sum_{k=1}^E \mathbf{Z}_k$ be the evoked response of the spatially filtered data where E is the number of epochs and

$\mathbf{Z}_k$ is the k^{th} epoch of the spatially filtered target data. $\mathbf{Z}_k \in \mathbb{R}^{C \times T}$, where C is the number of electrode channels, and T is the number of time samples in one epoch. We create the super epochs where the k^{th} sample will be $\mathbf{Z}'_k = \{Z_{k,1}, Z_{k,2}, Z_{k,3}, Z_{k,4}, E_{Z,1}, E_{Z,2}, E_{Z,3}, E_{Z,4}\}$. Combining this preserved information with the application of the spatial filter, intra-epoch variability is mitigated. We compute the evoked response of the spatially filtered data using the 22 s calibration data, and we concatenate it to all epochs, including both the training and test sets.

Then, to perform the Riemannian spatial alignment, we use the participant's calibration phase as a training batch and compute the associated covariance matrices with the Ledoit-Wolf estimator [16]. Then, we measure the Riemannian mean of these matrices, and finally, we transport the samples of the participant with this mean as follows:

$$\hat{\mathbf{Z}}'_i = \bar{\mathbf{R}}^{-\frac{1}{2}} \mathbf{Z}'_i \quad \forall i \in [1, \dots, E] \quad (1)$$

where $\mathbf{Z}'_i \in \mathbb{R}^{C \times T}$ is an element of $\{\mathbf{Z}'_i\}_{i \in [1, \dots, E]}$, $\bar{\mathbf{R}}$ is the Riemannian mean of the covariances matrices of the set $\{\mathbf{Z}'_1, \dots, \mathbf{Z}'_A\}$ and A is the number of samples in the calibration set of the participant. As the forward model presented in the next sections output data with only one channel, we select, at the end, only the first components of the preprocessed data.

Simulation-based inference: For SBI framework we need to define three specific components (see Figure 1). Firstly, we define a forward model that can generate c-VEP responses given a stimulus sequence and a set of VEP parameters. Secondly, we define a prior over the VEP parameters to sample from and input to the forward model and generate a synthetic dataset. Finally, we define a backward model that can retrieve the VEP parameters from any c-VEP data. This backward model composed of an embedding net and a density estimator to retrieve the parameters from the data.

To perform the training of the SBI framework, we first create simulated trials, 15000 trials in our study, by sampling the parameters from the prior and using the forward model with a specific stimulus sequence to generate data. This yields a synthetic dataset of c-VEP responses given ground-truth VEP parameters. Subsequently, we train the backward model (embedding net and density estimator) on this synthetic dataset and its ground-truth labels.

Once trained, we evaluate the performance of the SBI pipeline to regress the parameters given an unseen synthetic dataset, and to classify the attended code sequence given a hold-out empirical dataset.

The SBI framework was implemented using the `sbi` Python library [17]. The following sections describe the three components of the SBI framework in more detail.

SBI Forward Model: The forward model generates an EEG response to a given c-VEP stimulus making use of the deconvolution framework [3]. Specifically, we assume that the c-VEP $s_i(t)$ can be modeled as a convolution of the i -th stimulus $c_i(t)$ and a temporal response function $r(t)$:

$$s_i(t) = c_i(t) * r(t) \quad (2)$$

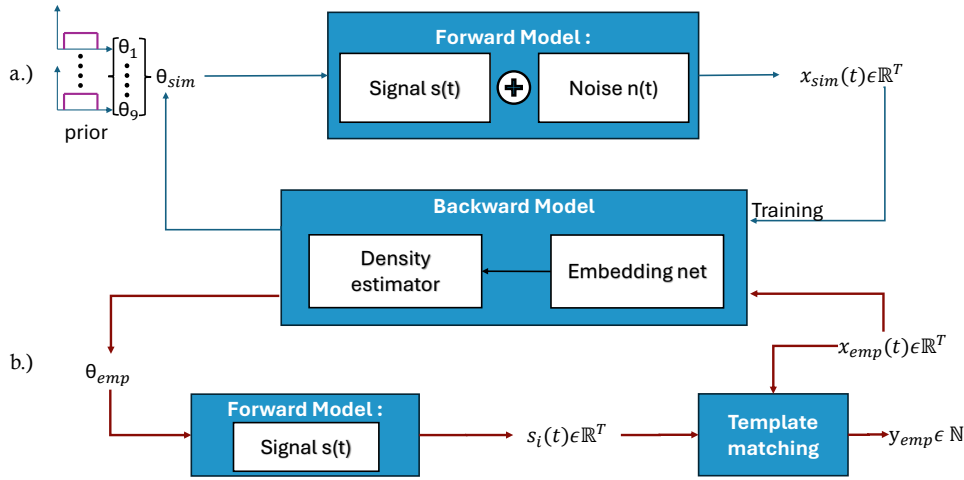


Figure 1: **Schematic of the SBI framework.** (a.) The training loop that samples parameters, generates data using the forward model, uses that to train a backwards model to regress these parameters. (b.) The classification procedure on empirical data to retrieve the attended code sequence. N : Number of codes, T : number of timesteps, θ : parameters, $x(t)$: EEG data, $s_i(t)$: templates, y : labels, $_{sim}$: simulated, $_{emp}$: empirical.

The stimulus sequence $c_i(t)$ is one of the 5 pseudo-random binary noise-codes used during the experiment. We model the temporal response function $r(t)$, which represents the flash-VEP, as the addition of 3 Gaussian components, which ideally represent the N75, P100, and N125 complex of a classical flash-VEP. Each Gaussian component is modeled with 3 parameters: the amplitude, mean and standard deviation of the Gaussian, corresponding to the amplitude, latency and width of the peak. Finally, we add measurement noise $n(t) \sim \mathcal{N}(0, \sigma)$ represented by a zero-mean Gaussian with variance $\sigma = 0.1$ to obtain the simulated data $x_{sim}(t)$:

$$x_{sim}(t) = s_i(t) + n(t) \quad (3)$$

Thus, the forward model is parameterized by a set of 9 parameters θ_{sim} characterizing the underlying VEP model $r(t)$. Given a set of 9 parameters θ_{sim} and a stimulus code $c_i(t)$, the forward model outputs simulated c-VEP source activity $x_{sim}(t)$ of size T , with T corresponding to the number of time-points in our data.

SBI Backward Model: The backward model takes EEG data and retrieves the VEP parameters. It is composed of two components, an embedding net and a density estimator. The aim of the embedding net is to reduce the dimensionality of the input and to create features for the density estimator. The embedding net is important as the density estimator needs clear, informative and low-dimensional features. The density estimator yields a posterior distribution over the VEP parameters from the features created by the embedding net.

As embedding net we used a recurrent neural network (RNN) composed of one hidden layer with 128 units followed by a linear layer with 20 units and a ReLU activation layer. This was motivated by a preliminary

benchmark performed between four different embedding nets: a handcrafted summary statistic function, a convolutional neural network, the encoding part of a trained auto-encoder, and the RNN. The RNN was selected as it performed best on the simulated data when performing the local classifier two-sample test (LC2ST) [18].

As density estimator we used the sequential neural posterior estimator (SNPE) [19]. SNPE is the most widely used density estimator for SBI and can learn the likelihood function from compressed representations of the data. SNPE trains an inference network to directly approximate the posterior distribution $p(\theta|x)$, using samples (θ, x) drawn from the joint distribution $p(\theta, x)$. Unlike methods that focus on predicting a single point estimate of the parameters, SNPE trains the network to predict the parameters of a probability distribution over θ , conditioned on the observation sample. Such networks are trained by minimizing the conditional negative log-likelihood of simulated parameters under the neural posterior estimator. Once the inference network is trained, it can be used to infer θ from any new observed data $x_{obs}(t)$, providing the approximate posterior distribution $p(\theta, x_{obs}(t))$.

A key advantage of this approach is that the network, once trained, can perform inference efficiently for any new observations, either synthetic or empirical, without needing additional simulations or retraining. This amortization feature makes SBI powerful for analyzing many different observations from the same model. This gives a special advantage compared to time-consuming algorithms like MCMC sampling.

SBI Prior Parameters: The SBI framework requires a prior distribution over the VEP parameters to generate synthetic data used to train the backward model. Ideally,

this prior distribution represents the expected range of parameters for the unknown target distribution, here the empirical data.

We define the prior for each parameter as a uniform distribution, limited to a certain range. We informed the lower and higher boundary of this range by analysing the empirical data. That is, for each participant, we fit the VEP model $r(t)$ of the forward model on averaged target epochs. From this fit, we retrieve the 9 parameters, and proceed in one of the following two ways. First, to model a prior across participants, we set the range to the smallest and largest value found across participants, referred to as the *across-pp prior*. Second, to model a prior per individual participant, we apply a small range around the best individual parameters (± 0.02 for the latency, ± 0.2 for the amplitude, ± 0.001 for the width), referred to as the *within-pp prior*.

Analysis: The SBI framework is trained by sampling from the prior, generating synthetic data, and calibrating the backward model using that synthetic data. After training, the SBI framework is evaluated in two ways.

First, we evaluate the regression of the VEP parameters using an unseen simulated dataset. For this analysis, we restrict to synthetic data, because we have access to the ground truth parameters, which is not the case for the empirical dataset. For this analysis, we report the root mean squared error (RMSE).

Second, we evaluate the SBI framework by performing a classification analysis. That is, we apply the backwards model to retrieve VEP parameters and use these to generate template responses using the signal component of the forward model $s(t)$. Specifically, we apply the forward model given each of the stimulus sequences c_i , with $i \in [1, \dots, 5]$, and the predicted parameters θ , but refrain from adding any noise $n(t)$ to the signal $s(t)$ to generate templates $s_i(t)$. Then, we can perform a standard template-matching classifier that maximizes the correlation between data $x(t)$ and templates $s_i(t)$ to retrieve the code index i :

$$\hat{y} = \arg \max_i \text{corr}(x(t), s_i(t)) \quad (4)$$

For each participants, the classification analysis was performed for both prior on both an unseen synthetic dataset of 100 trials, as well as the empirical dataset composed of 75 trials, referred to as *simulated-to-empirical* and *simulated-to-simulated*, respectively.

RESULTS

To assess the performance of retrieving the VEP parameters, we first compute the root mean squared error between the ground-truth parameters used to create the synthetic data in the *simulated-to-simulated* mode for both priors and the associated retrieved parameters. The results are described in Table 1. We see that within-pp RMSE is lower than across-pp RMSE.

Secondly, we assess the performance of retrieving the VEP parameters by computing the mean squared error

between the ground truth VEP response and the inferred VEP response. For the *simulated-to-simulated*, we obtain an average value of 0.08 ± 0.07 for the within-pp prior and 0.29 ± 0.26 for the across-pp prior. For the *simulated-to-empirical*, we get an average value of 0.11 ± 0.06 for the within-pp prior and 0.45 ± 0.17 for the across-pp prior.

Figure 2a shows the VEPs from the *simulated-to-empirical* analysis. We observed that the within-pp prior fits the empirical VEP very well, even for participant P21 which showed the highest MSE value. However, we observed that the VEP responses created with the parameters retrieved from the SBI pipeline trained with the across-pp prior, are far from the empirical value and can explain the low performance with the across-pp prior.

We further assessed the SBI performance by estimating classification accuracy. Figure 3 shows the classification accuracy of both priors in both synthetic to synthetic, as well as synthetic to empirical modes. We observed that the within-pp prior reached a significantly higher accuracy than the across-pp prior (86.2% vs 33.8%, $p < .001$). Moreover, the *simulated-to-simulated* performance is significantly higher than the *simulated-to-empirical* (36.8% vs 30.8% for the across-pp prior, $p = .037$; 89.8% vs 82.6% for the within-pp prior, $p = .028$).

DISCUSSION

To the best of the authors' knowledge, this study is the first to apply simulation-based inference (SBI) in the field of BCI, specifically to the c-VEP paradigm. SBI offers the potential to eliminate the initial calibration step that is typically required for classification models, since, within this framework, training can be performed on a virtually unlimited amount of simulated data.

In this feasibility study, we showed that satisfactory performance can be achieved both in recovering the underlying response parameters and in application-level decoding accuracy, provided that the prior parameters are chosen appropriately. Specifically, priors defined at the individual-participant level yielded high classification accuracy, whereas a universal across-participant prior led to a substantial decrease in performance.

The lower accuracy observed with the across-participant prior may be explained by several neurophysiological factors, including inter-subject and intra-subject variability [20]. In particular, the VEPs estimated from the empirical data showed that not all participants exhibited the canonical three-peak flash-VEP pattern, consisting of a negative peak followed by a positive peak and then a second negative peak [2, 12]. In addition, the latencies of these peaks varied considerably across several participants. This variability may have made it difficult to define a broad across-participant parameter range and also led to overlap in the latency ranges of the three components.

While we aimed to evaluate the full SBI framework, at two stages we relied on empirical data to estimate properties of the framework itself. This warrants caution in

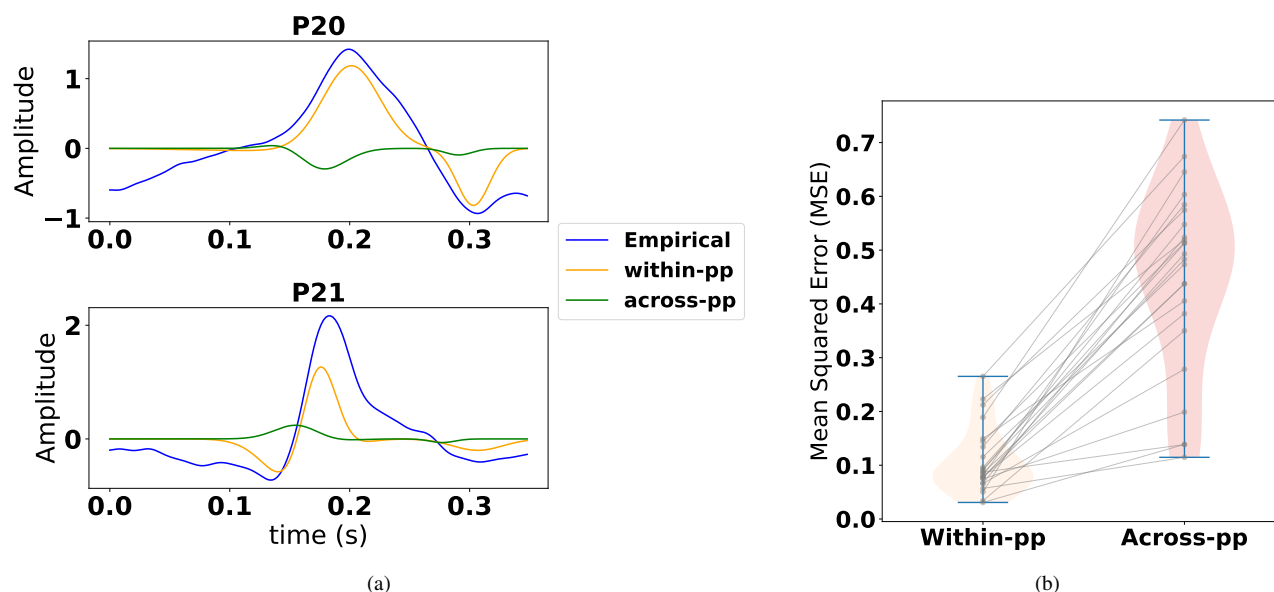


Figure 2: **Regression performance of the SBI framework.** (a) Displayed are the empirical visual evoked potential (VEP) (blue line), predicted VEP using the within-participant prior (yellow line), and predicted VEP with the across-participant prior (green line). Two participants are shown: on the top the participant with the lowest MSE and on the bottom the participant with the highest MSE. (b) Shown is the MSE between the empirical VEP and the simulated VEP created from the parameters retrieved with the within-pp prior (peach color) or with the across-pp prior (coral color). Each dot corresponds to the MSE of one participant. Line connect the points of the same participant.

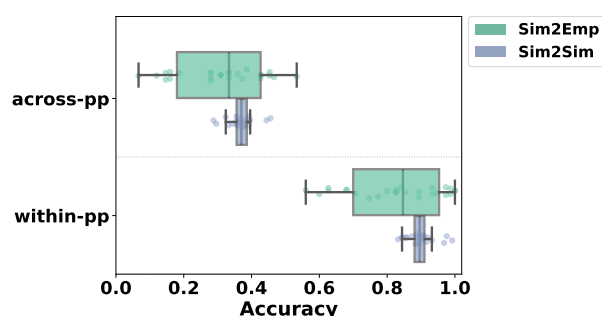


Figure 3: **Classification performance of the SBI framework.** Accuracy for the *simulated-to-simulated* (Sim2Sim) and *simulated-to-empirical* (Sim2Emp) modes for the across-participant and within-participant priors.

the interpretation of the results and limits the extent to which the current framework can be generalized.

First, we simulated a single EEG component extracted with xDAWN and applied Riemannian alignment. This improved performance compared with modeling a single channel. While it required calibration of xDAWN and the Riemannian alignment using only 22 s of data, it violates the amortization of SBI, being dependent on some empirical data for the calibration process.

Second, also the prior distributions were inspired by the empirical data, which under the true SBI framework would not be fully known. Especially constraining to a participant-specific prior considerably improved results. Future work should improve the across-participant setting, as well as test the generalization beyond the currently used dataset, EEG setup, and stimulus protocol. To avoid contradictory information in the training of the

SBI framework, we used only one code when creating the simulated dataset to train the SBI. However, when using a broader prior like the across-participant one, the classification steps tend to classify the data as the same label as the one used in for the training of the SBI, explaining the low performance for the across-pp prior.

In a preliminary analysis, we used the across-pp prior and trained the SBI using all 5 codes, which increased the *simulated-to-simulated* accuracy to 81.3%, but the accuracy remained at 32% for the *simulated-to-empirical*. Future work should integrate all stimulation codes into the SBI training procedure to improve generalization across codes in empirical data. In addition, it will be important to investigate whether the SBI framework can be extended to other experimental protocols.

The forward model used in this study can be considerably improved to reflect realistic EEG data. First, the current model enforces a 3-peak flash-VEP, while some participants showed fewer or more peaks, allowing further finetuning. Second, the current model is fully linear, not modeling any non-linear dynamics like caused by habituation. Third, the noise model considered measurement noise only, ignoring known EEG properties like $1/f$ noise, potential artifacts such as eye blinks, etc. While the current results already show feasibility, it is expected that an improved forward model may further improve results. In the backward model, we used an RNN for the embedding net. We believe that this embedding net could be improved by optimizing its hyperparameters or by developing a new embedding net that better captures features for the density estimator in the c-VEP domain. Furthermore, future work should aim at making the backward

Table 1: **Root mean squared error.** RMSE between predicted parameters and ground-truth used to create synthetic data in the *simulated-to-simulated* mode for both priors.

	Amplitude (a.u.)			Latency (s)			Width (a.u.)		
	N75	P100	N125	N75	P100	N125	N75	P100	N125
across-pp	3.9e-6 ±	9.5e-6 ±	5.6e-6 ±	9.3e-2 ±	7.2e-1 ±	1.3e-1 ±	1.8e-7 ±	3.5e-8 ±	4.1e-8 ±
	3.6e-6	9.6e-6	3.6e-6	8.8e-2	4.1e-1	9.1e-2	1.4e-7	2.6e-8	3.4e-8
within-pp	1.3e-7 ±	1.4e-7 ±	2.2e-7 ±	1.3e-3 ±	1.5e-3 ±	1.6e-3 ±	1.0e-8 ±	7.0e-9 ±	8.1e-9 ±
	8.9e-8	1.1e-7	9.7e-8	9.3e-4	9.8e-4	1.1e-3	4.6e-9	6.0e-9	6.8e-9

model invariant to the stimulus protocol.

While the across-participant setting showed poor performance, the within-participant setting showed promising results. An interesting future direction could be to start the SBI with an uninformed prior for any new user. Subsequently, the prior can be updated by using the predicted parameters from empirical data of the current user. The updated prior can then initiate the calibration of a new SBI, more tailored to that current user. That is, we can perform an active learning SBI framework that adapts in real-time, ultimately improving its performance on unseen data.

CONCLUSION

This study is the first of its kind to demonstrate the feasibility of applying the simulation based inference (SBI) framework in c-VEP based BCI. We show good accuracy in classification of empirical data while having trained the framework on synthetic data. We observed that a prior across participants decreased overall accuracy as compared to a prior within participants, in line with substantial inter-subject variability. We discussed several challenges and how to improve the SBI framework in future work, paving the way for retrieving response parameters and classification of empirical data, while offloading calibration to synthetic data.

REFERENCES

[1] Martinez-Cagigal V, Thielen J, Santamaria-Vazquez E, Perez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): A literature review. *Journal of Neural Engineering*. 2021.

[2] Cabrera Castillos K, Ladouce S, Darmet L, Dehais F. Burst c-VEP based BCI: Optimizing stimulus design for enhanced classification with minimal calibration data and improved user experience. *NeuroImage*. 2023;11.

[3] Thielen J, Broek P van den, Farquhar J, Desain P. Broad-Band visually evoked potentials: Re(con)volution in brain-computer interfacing. *PLOS One*. 2015.

[4] Gilks WR, Richardson S, Spiegelhalter D, Eds. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC: New York (1995).

[5] Zhang C, Butepage J, Kjellstrom H, Mandt S. *Advances in Variational Inference*. Oct. 23, 2018.

[6] Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*. 2010;410–418.

[7] Deistler M et al. “Simulation-Based Inference: A Practical Guide.” arXiv. Tech. Rep. 2025.

[8] Lueckmann JM, Boelts J, Greenberg DS, Gonçalves PJ, Macke JH. Benchmarking Simulation-Based Inference. 2021.

[9] Brehmer J. Simulation-based inference in particle physics. *Nature Reviews Physics*. 2021.

[10] Dax M, Green SR, Gair J, Macke JH, Buonanno A, Schölkopf B. Real-Time Gravitational Wave Science with Neural Posterior Estimation. *Physical Review Letters*. 2021.

[11] Bernardo D et al. Simulation-based inference of developmental EEG maturation with the spectral graph model. *Communications Physics*. 2024.

[12] Dehais F, Cabrera Castillos K, Ladouce S, Clisson P. Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry EEG code-VEP BCI. *Journal of Neural Engineering*. 2024.

[13] Rivet B, Souloumiak A, Attina V, Gibert G. xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface. *IEEE Transactions on Biomedical Engineering*. 2009.

[14] Aristimunha B et al. *Mother of all BCI Benchmarks*. Version 1.4.3. 2025.

[15] Mellot A, Collas A, Chevallier S, Gramfort A, Engemann DA. Geodesic optimization for predictive shift adaptation on EEG data. In: *NeurIPS*. 2024.

[16] Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*. 2004.

[17] Boelts J et al. sbi reloaded: A toolkit for simulation-based inference workflows. *Journal of Open Source Software*. 2025.

[18] Linhart J, Gramfort A, Rodrigues PLC. L-C2ST: Local Diagnostics for Posterior Approximations in Simulation-Based Inference. In: *NeurIPS*. 2023.

[19] Greenberg DS, Nonnenmacher M, Macke JH. Automatic Posterior Transformation for Likelihood-Free Inference. In: *ICML*. 2019.

[20] Velut S, Thielen J, Chevallier S, Corsi MC, Dehais F. Neurophysiological screening of individual variability for robust decoding in c-VEP-based BCI. *Imaging Neuroscience*. 2026.

BCI-based Intelligent Tutoring Systems: A Literature Review and Meta-Analysis on Neuroadaptive ITS

J. Petit¹, M. J-M. Macé¹, A. Lécuyer², L. Pillette¹

¹Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

²Inria, Univ. Rennes, IRISA, CNRS, Rennes, France

E-mail: jimmy.petit@irisa.fr

ABSTRACT: Intelligent tutoring systems (ITS) are artificial systems that provide personalised instruction, relying on student models to monitor cognitive states, such as mental workload. Recent studies have increasingly enhanced this monitoring through the use of passive brain-computer interfaces (BCI). In this review, we focused on ITS that use BCI metrics to adapt their interventions in real time. Following PRISMA 2020 guidelines we searched PubMed, Web of Science, Scopus, OpenAlex, and Semantic Scholar databases. We conducted a risk-of-bias assessment using RoB 2.0, and performed a meta-analysis to evaluate the impact of BCI-based ITS on learning outcomes. We found a moderate-to-large pooled effect size ($SMD = 0.79$) with a 95% confidence interval $[-0.08; 1.65]$ suggesting a potential benefit of BCI-based ITS. However, the wide confidence interval that includes zero indicates substantial uncertainty. As this field is still emerging, with most studies published after 2015, further research is needed to strengthen the evidence base and support more definitive conclusions.

INTRODUCTION

An intelligent tutoring system (ITS) is a computer-based instructional system that provides personalised guidance and feedback by adapting to a learner's needs, performance, and cognitive state, much like a human would in a learning context. It can provide abstract knowledge for formal education, e.g. mathematics [1] or biology [2], or applied skill sets for professional settings. For instance, the ITS called "CODES" provides a web-based environment for cooperative music prototyping [3]. An ITS consists of four basic components: *Domain model*, *Student model*, *Tutoring model* combined within a *User Interface* [4]. The *Domain model*, also known as *Expert model*, is a collection of concepts, rules, and problem-solving strategies of the domain to be learned. This *Domain model* can serve several purposes, for instance, it can be used as a standard to evaluate the trainee's performance or be used to detect errors. The *Student model* is a model that contains information on the trainee, such as their cognitive states and emotional states. It aims at monitoring these states during the step-by-step learning of the trainee. The model can also flag when the student model diverges from the domain model to provide a

correction. The *Student model* is often seen as the core model of the ITS. Finally, the *Tutor model* receives input from the two other models and provides decisions about tutoring strategies or interventions through the use of the *User interface*. To implement a *Student model*, an ITS designer often relies on monitoring in real-time the trainee's levels of mental workload or engagement [5, 6]. Various physiological sensors, including for instance heart rate variability, oculomotor and electrodermal activity, have been used to detect changes in these mental states [7, 8]. However, brain imaging techniques, such as electroencephalography (EEG), are considered to offer a more direct assessment of mental states in real-time [9]. As a result, passive BCI, hereafter simply referred to as BCI, have gained increasing attention within the ITS community over the years.

In a large meta-analysis comprising 107 effect sizes from 14,321 participants, Ma *et al.* [10] provided strong evidence that, in certain contexts, ITS can effectively complement and even substitute for other instructional modalities. Their findings indicated that ITS was associated with higher achievement compared with teacher-led large-group instruction, non-ITS computer-based instruction, and textbook- or workbook-based learning. Instructional formats that did not appear to benefit from ITS, however, were individualized human tutoring and small-group instruction.

Previous literature reviews at the intersection of neurophysiological assessment and learning have focused on different aspects. For instance, Xu *et al.* [11] have published a survey investigating EEG-based BCI for emotion recognition, such as happiness or sadness, in ITS. They focused on the accuracy of the system, however, its integration to actual ITS for inclusion in a *Student model* was not discussed. Another recent review of ITS focused on ADHD symptoms management through the use of serious-games [12]. The authors found that such ITS promote the acquisition of a wide range of cognitive and socioemotional skills among children with ADHD, including visuospatial working memory, attention and inhibition control, planning/organizing and problem solving. Overall, prior reviews have not examined how neurophysiological data have been incorporated into ITS nor whether such information yields additional benefits for learning outcomes.

In this literature review, we focus on studies of ITS-based learning that implement real-time adaptation driven by neurophysiological data reflecting the learner's state, such as engagement, attention, or mental workload. In particular, we examine the details of the adaptation strategies, signal processing methods and experimental design choices used to assess these systems. Through a meta-analysis, we seek to evaluate the effect of BCI-based adaptation in ITS on learning outcomes.

MATERIALS AND METHODS

Study selection In this literature review we included any ITS combined with a BCI. To that end, we incorporated ITS terminology from the classification of educational agents from Chou *et al.* [15] and main neurophysiological data acquisition methods. This resulted in the following query applied to title and abstract content, details of the exact query applied to each database are given in supplementary materials¹:

```
[ "Intelligent Tutoring"
OR "Adaptive Tutoring"
OR "Learning Companion*"
OR "Co-learning Companion*"
OR "Co-learning Agent*"
OR "Emotional Companion*"
OR "Educational Companion*"
OR "Pedagogical Companion*"
OR "Emotional Agent*"
OR "Educational Agent*"
OR "Pedagogical Agent*"
OR "Tangible Interface*"
OR "Tangible Interaction*" ]
AND
[ "Brain Activity"
OR "Electroencephalography" OR "EEG"
OR "Magnetoencephalography" OR "MEG"
OR "fNIRS" OR "fMRI"
OR "Neurofeedback" OR "NF"
OR "Brain-Computer Interface*" OR "BCI" ]
```

Following PRISMA 2020 guidelines, we conducted this search across PubMed, Web of Science, Scopus², OpenAlex, and Semantic Scholar² databases on 16/12/2025. The initial search yielded 325 references. After removing non-experimental studies and duplicates, 115 papers were screened, of which 8 studies met the inclusion criteria and were included in the analysis. Conference and journal articles were included if they: (i) involved both an ITS and measures of brain activity (ii) engaged participants in a learning task (iii) used indices derived from brain activity to adapt the ITS's behavior or pedagogical content. Studies were excluded if they: (i) were not published in English (ii) were proof-of-concept reports based on a single participant (iii) described experimental results

or ITS already presented in a more detailed article. In cases of duplicate reporting, only the most detailed and comprehensive publication was included. No restrictions were set with regards to study type or publication date. The details of the identification, screening, and study inclusion are presented in Figure 1. After inclusion in the review, we extracted the following information: protocol design, sample size, the names of the ITS, their embodiment, the adaptation strategy employed, the brain signal modalities and systems used, the number of available channels, the learning domain and outcome measures, as well as details of the signal-processing pipeline, including the types of temporal and spatial filters applied and the indices derived from brain activity.

Risk of Bias We assessed the risk of bias for all studies included in the literature review using the revised Cochrane Risk of Bias tool (RoB 2.0). When applicable, we applied either the version designed for randomised trials or the supplementary guidance for crossover trial designs. Our assessment focused on the per-protocol effect, examining adherence to the intervention as well as any deviations or implementation failures that could have influenced the reported outcomes.

Meta-analysis To be included in the meta-analysis, studies had to meet two additional criteria: (i) they reported a measurable learning outcome associated with the ITS, and (ii) they included both a control group and an intervention group. Among these studies, we distinguished two types of control groups based on how the control condition was defined. Passive control groups referred to studies comparing a BCI-based adaptive ITS (intervention) with either a static ITS that provided no additional adaptation or with no ITS at all (control). Active control groups, by contrast, compared the BCI-based adaptive ITS with an ITS delivering interventions similar in form to those of the adaptive system, but generated randomly or pseudo-randomly, and therefore unrelated with the neurophysiological data. This second type of control allows for a more balanced comparison by controlling for the amount of intervention and isolating the specific contribution of *adaptivity*. To account for this heterogeneity, the meta-analysis included control-group type as a moderator. This approach yields an overall effect size, subgroup-specific effects, and a test of between-group differences.

One of the included studies had a cross-over design and was analysed as a parallel-group trial [16]. This conservative approach, treating the two conditions as independent groups, ignores the pairing and therefore inflates the variance rather than the effect size, resulting in an under-weighted contribution of this study rather than an artificial gain in precision. Another study included multiple control groups that shared the same experimental group [17]. In accordance with the recommendations of the *Cochrane Handbook for Systematic Reviews* [18], we divided the shared group size equally among the comparisons to avoid double weighting and maintain the statistical independence of the effects.

¹https://gitlab.inria.fr/jipetit/passive_bci_and_its_review

²Accessed via Publish or Perish 8 software.

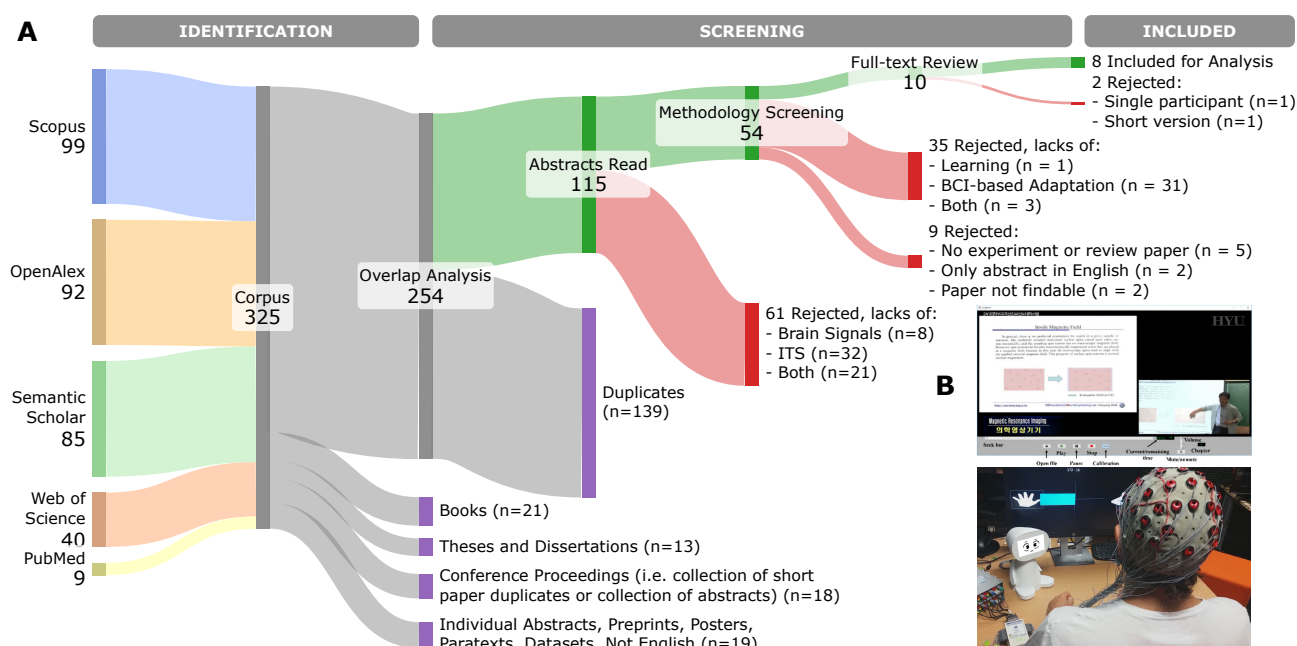


Figure 1: A. PRISMA diagram. B. Visual of two ITS: Top – User interface of the ITS in [13]. This ITS is not embodied. Down – ITS Peanut [14]. All figures have been reproduced with the appreciated authorisation of their respective authors. More visuals and details of the overlap analysis are presented as a Venn diagram in the supplementary material³.

To account for variability in assessment scales of the learning outcomes and variability in experimental conditions, we used the standardised mean difference (SMD) as the effect-size metric and conducted a meta-analysis using a random-effects model. Heterogeneity was assessed using I^2 and τ^2 . Given the small number of available effect sizes ($k = 5$ and $k = 3$ in subgroups), these estimates should be interpreted with caution.

RESULTS

Eight studies were included in the literature review, of which six met the inclusion criteria. At the same time, two did not because they lacked a control group (criterion (ii)) [19, 20]. The methodological characteristics of these studies are summarised below, with additional details provided in the supplementary material³.

ITS Characteristics ITS representations varied widely, ranging from humanoid robots and tangible companions [14, 16, 17, 20, 21] to static facial images animated through lip-sync software [19]. In [13] and [9], the authors did not embody the ITS with either a physical or virtual avatar. The former ITS is a video player that provides a lecture on MRI; the latter is an ITS called MENTOR designed to teach mathematical notation called reverse Polish notation through a series of pedagogical activities. Despite this diversity, their shared purpose is to support learning by adapting instructions to the learner's cognitive or emotional state. Most systems rely on an engagement index computed from the ratio $\beta/(\alpha + \theta)$ [9, 16, 17, 19, 21], sometimes complemented by proprietary metrics of stress [21]. In [13], the

authors developed an *ad-hoc* “attention index” in a preliminary study involving 32 participants. Candidate indices included $\beta/(\alpha + \theta)$, β/α , $(\beta + \theta)/\alpha$, and θ/α , computed for each available channel (Fp1, Fpz, Fp2, F3, Fz, F4, C3, Cz, and C4). Participants were instructed to rapidly identify a “d2” target (the letter *d* with two dots above or below it) among visually similar distractors (a “d” with only one dot, or a “p” replacing the “d”). High task performance was interpreted as reflecting a high attentional state. Based on these performances, the authors selected the fourth-highest candidate, θ/α at Cz, even though stronger candidates, such as $\beta/(\alpha + \theta)$ at C3, were identified, Cz was ultimately chosen due to manufacturing considerations. Additionally, moving average techniques were commonly applied to smooth the index over time and reduce false positive ITS adaptations caused by transient fluctuations. For example, [16] combined an exponentially weighted moving average with a 4-s moving average window, while [9] used a 40-s moving-average window. Only 2 studies did not use an engagement or attention index. In [20], the authors implemented four types of filters, each designed to highlight specific aspects of the user's brain activity on a humanoid tangible companion. For example, to emphasise vision-related activity, they applied a narrow band-pass filter centred on the α frequency band (8–12 Hz), followed by spatial filtering that retained channels located over and around the occipital cortex (P3, Pz, P4, PO3, POz, PO4, O1, Oz, and O2). The resulting α -band amplitude topomaps were then projected onto the tangible companion. In [14], the authors implemented a standard MI-based BCI pipeline consisting of temporal filtering (8-30 Hz band-pass) combined with spatial filtering using common spatial

³https://gitlab.inria.fr/jipetit/passive_bci_and_its_review

pattern filters. Band-power features were estimated via log-variance, and classification was performed using a shrinkage linear discriminant analysis.

All of these neurophysiological signals drive a variety of adaptation strategies: repeating or simplifying content when engagement drops [19, 21], inserting multimedia segments to enliven attention or provides supplementary information [13], modulating a robot's gestures or prosody [16, 17], or selecting between worked examples and problem-solving tasks [9]. Two systems stand apart from these cognitive-state-driven approaches. Frey *et al.* [20] implemented a tangible augmented-reality tutor that does not adapt instructional content but instead visualizes the learner's brain activity in real time through multiple EEG-based filters, aiming to foster conceptual understanding of brain processes. Similarly, Pillette *et al.* [14] adapted their system's behavior based on BCI performance during a mental-imagery tasks, providing emotional support and social presence through a rule-based mechanism rather than monitoring cognitive states. The learning domains addressed range from language learning [16, 19] and mathematics [9] to computer science [21], folk tale [17], mental imagery [14], and neuroscience [13, 20], with outcome measures including questionnaires [9, 13, 16, 17, 20, 21] or classification accuracy [14, 19].

All studies relied on EEG to measure brain activity, with an average of 14 channels, ranging from 1 [13, 17] to 32 [20]. Some studies did not specify the types of electrodes used, but based on the referenced hardware, we inferred that 6 out of 8 studies relied on wet electrodes [9, 13, 14, 19–21], while one study used dry electrodes [17]. Interestingly, [19] explicitly refers to dry channels despite using the Emotive Epoc+ headset, which is a saline-based system. In [16], the authors employed a Unicorn Hybrid Black system (g.tec medical engineering, Austria), which can use either wet or dry electrodes, but they did not specify which configuration was used.

Risk of Bias The methodological quality of the eight included studies is summarised in Table 1. Overall, evidence quality was moderate: three studies showed low risk of bias, two showed some concerns, and only two were rated as high risk. For the randomisation domain (D1), all studies were at most single-blind and generally provided limited information about their randomisation procedures. For example, [17] did not report group homogeneity tests despite visible baseline differences, and in [13], one passive control group was recorded before the treatment group for calibration purposes, preventing proper random allocation and resulting in “high risk” in D1 for both studies. In [14], the control group consisted of previously collected data, which also compromised randomisation. However, because the study aimed to evaluate the social presence of an ITS, reusing earlier data reduced the risk of experimenter and participant awareness to intervention, resulting in an overall assessment of “some concerns”. Carryover effects (DS) were unlikely in [16], as distinct pedagogical materials were

Table 1: Traffic light plot for Risk of Bias (RoB 2.0).

	D1	DS	D2	D3	D4	D5	Overall
[19]	N/A	N/A	N/A	⊕	⊕	⊕	⊕
[20]	N/A	N/A	N/A	⊕	⊕	⊕	⊕
[17]	⊖	N/A	⊕	⊕	⊕	⊕	⊖
[13]	⊖	N/A	⊕	⊕	⊕	⊕	⊖
[14]	⊕	N/A	⊕	⊕	⊕	⊕	⊕
[9]	⊕	N/A	⊕	⊕	⊕	⊕	⊕
[21]	⊕	N/A	⊕	⊕	⊕	⊕	⊕
[16]	⊕	⊕	⊕	⊕	⊕	⊕	⊕

Legend. D1: Randomisation process; DS: Bias arising from period and carryover effects (only in within-subject study); D2: Deviations from the intended interventions; D3: Missing outcome data; D4: Measurement of the outcome; D5: Selection of the reported result.

⊕ : Low risk; ⊕ : Some concern; ⊖ : High risk.

used across sequences. Deviations from intended interventions (D2) and missing outcome data (D3) were minimal, resulting in low risk for both domains. For the outcome measurement (D4), most studies relied on objective assessments of learning outcomes, including multiple choice questionnaires [9, 16, 21], open-ended and essay questions [13, 20], or system performance metrics [14]. In [17], the authors did not specify the type of questions used, other than noting that they assessed participants' ability to recall details of the story. In [19], no learning outcome assessment was provided, instead, the authors reported only the system's accuracy in recognising the engagement index. The absence of control groups in [19, 20] limited between-group comparisons, led to “some concerns”, while being of “low risk” for the other studies. Finally, selective reporting (D5) was generally transparent, but the lack of preregistered protocols resulted in an overall rating of “some concerns”.

Meta-analysis Six of the 8 studies included in the literature review met the inclusion criteria of the meta-analysis [9, 13, 14, 16, 17, 21]. Collectively, these studies provided 8 effect sizes because 2 studies included multiple control groups or treatment groups [13, 17]. Figure 2 presents the forest plot of the effect sizes measuring the learning outcomes from the adaptive ITS. Studies including a passive control group were the largest (treatment $n = 52$ and control $n = 65$), resulting in a pooled effect size SMD of 0.89 [−0.77; 2.54] (95%-CI) and a statistically significant heterogeneity ($I^2 = 77.2\%$ [44.8%; 90.6%], $\tau^2 = 1.2$ [0.2; 17.4], $p = 0.0015$). Three studies included an active control group (treatment $n = 28$ and control $n = 33$), resulting in a pooled effect size SMD of 0.77 [−1.24; 2.78] and a slightly smaller heterogeneity ($I^2 = 66\%$ [0%; 90.2%], $\tau^2 = 0.4$ [0.0; 24.5], $p = 0.053$). Overall, the pooled effect size SMD of 0.79 [−0.08; 1.65] showed a statistical trend in favour of neuroadaptive ITS ($t_7 = 2.15$, $p = 0.0689$) and a large statistically significant heterogeneity ($I^2 = 71.0\%$ [40.1%; 86.0%], $\tau^2 = 0.6$ [0.1; 5.4], $p = 0.0011$) calculated over a large set of participants (treatment $n = 80$

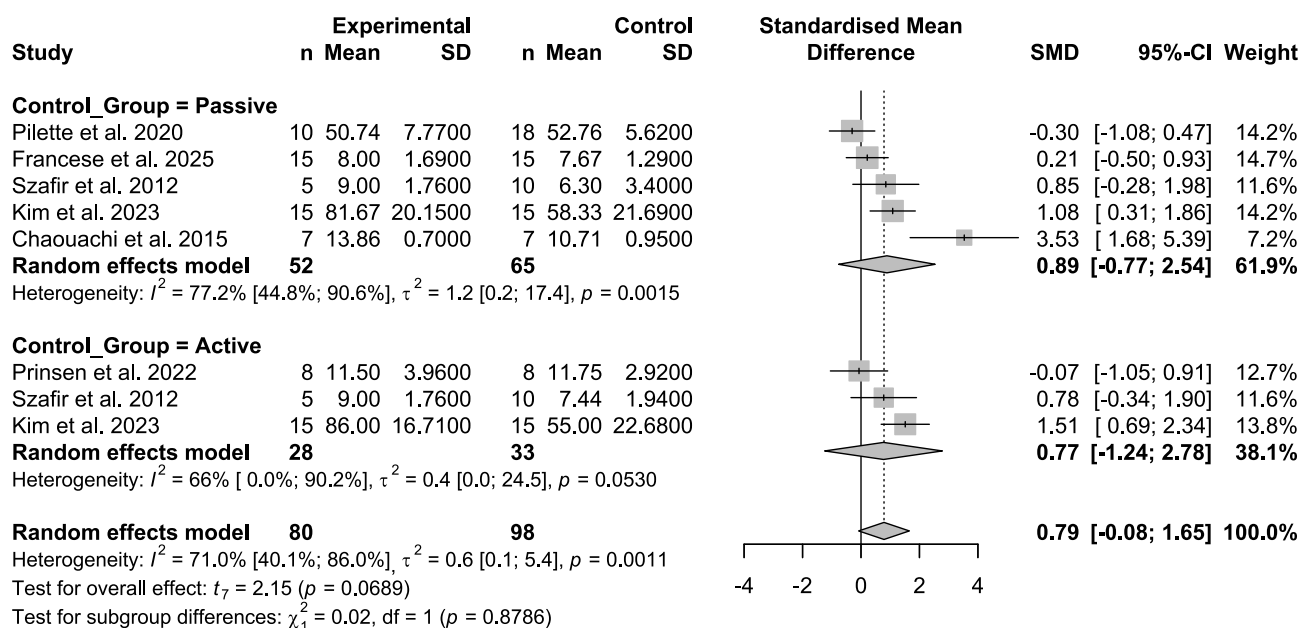


Figure 2: Forest plot of the meta-analysis comparing two subgroups of control groups versus passive BCI-based ITS.

and control $n = 98$). Subgroup differences between the two types of control group were not statistically significant ($\chi^2_1 = 0.02$, $p = 0.88$).

DISCUSSION

Our meta-analysis revealed a statistical trend toward a moderate-to-large positive effect ($SMD = 0.79$, 95% CI $[-0.08; 1.65]$, $p = 0.07$), suggesting a potential benefit of neuroadaptive ITS over standard ITS. However, wide confidence intervals crossing zero, indicate substantial remaining uncertainty. This uncertainty likely arises from several factors, including small sample sizes, heterogeneous study designs, variable outcome measures, and inconsistent signal processing pipelines. As the non-significant subgroup-difference test suggests, the type of control group alone does not account for this variability. Interestingly, studies using active control groups, i.e., where participants received a balanced amount of feedback, isolated more precisely the specific contribution of *adaptivity* within the ITS. These studies exhibited somewhat lower heterogeneity and therefore provide more conclusive empirical evidence in favour of BCI-based adaptive ITS. Greater standardisation in outcome measures, assessment scales, included electrodes, signal processing pipelines, and experimental protocols would help reduce these discrepancies in future work and improve comparability across studies. Finally, this research area remains in its early stages, with all but one study published after 2015. As additional studies accumulate, uncertainty surrounding the true effect size should decrease, allowing for more robust conclusions about the effectiveness of neuroadaptive ITS.

The RoB assessment indicated moderate evidence quality, mainly due to limited use of preregistered protocols and, more importantly, randomisation weaknesses. Unlike clinical research, where preregistration tools are

well-established, this field currently lacks identifiable infrastructure, making preregistration less common. However, as more studies are needed to draw firm conclusions on neuroadaptive ITS effectiveness, adopting more rigorous practices, particularly stronger randomisation procedures and, when feasible, double-blind designs, would produce more robust evidence.

EEG signals are particularly relevant to measure cognitive states. However, EEG-based monitoring of cognitive and affective states should be cautiously included in ITS. In [22], the authors show that EEG-based BCI remain limited for cognitive state monitoring due to low spatial resolution, strong sensitivity to ocular and muscular artifacts, and fluctuations in electrode-skin impedance that degrade signal quality. In turns, degraded signal quality may result in poorer and unstable cognitive state monitoring, like mental workload estimates. These factors should therefore be explicitly controlled, through appropriate preprocessing, artefact rejection, impedance monitoring, or predefined exclusion criteria, as they directly affect the reliability of the cognitive state monitoring. Also, EEG-based BCI suffers from pronounced inter- and intra-individual variabilities, combined with the non-stationarity of EEG, reduce model generalisation across users and sessions [22]. Many cognitive and affective states also share overlapping spectral signatures, constraining classifier specificity and increasing misclassification risks. In this context, the use of proprietary algorithms for which little or no methodological information is available should be avoided, as they prevent any assessment of validity, reproducibility, or bias. By contrast, mental workload is one of the most studied cognitive states in BCI research [23], with robust and reproducible EEG markers, most notably, increased frontal theta and decreased parietal alpha, making it a more reliable and promising target for future ITS applications. Finally, even though dry electrodes are improving, wet

electrodes should still be preferred, as they provide higher signal quality with fewer artefacts [24, 25]. Finally, in this review, we focused on ITS that adapt their pedagogical content or feedback based on brain activity. In future work, it would be relevant to compare the added value of these brain-signal input with other physiological measures, such as heart-rate variability, oculomotor activity, or electrodermal responses [7, 8], which have been used as reliable indicators of arousal, attention, and cognitive effort. Integrating these modalities, alongside EEG, could help disentangle overlapping cognitive states and improve the robustness of adaptive ITS.

CONCLUSION

BCI-based ITS show promising potential to enhance learning by adapting feedback and instructional behaviours to learners' cognitive states. Despite heterogeneous embodiment, learning domains, and signal processing methods, most systems relied on engagement or attention-related indices to guide real-time adaptation. Our meta-analysis suggests a moderate-to-large positive effect, though the wide confidence intervals and limited number of studies highlight substantial uncertainty. As the field is still emerging, with generally small samples and variable methodological reporting, more rigorous and standardised studies are needed to reach definitive conclusions.

Acknowledgements Work supported by the French National Research Agency (ANR-25-CE33-3211).

REFERENCES

- [1] A. Tato, R. N'Kambou, and R. Ghali, "Towards predicting attention and workload during math problem solving," in *Intelligent Tutoring Systems*, 2019.
- [2] W. Li, F. Wang, R. E. Mayer, and T. Liu, "Animated pedagogical agents enhance learning outcomes and brain activity during learning," *JCAL*, 2022.
- [3] E. M. Miletto, M. S. Pimenta, R. M. Vicari, and L. V. Flores, "Codes: A web-based environment for cooperative music prototyping," *Organised Sound*, 2005.
- [4] R. Nkambou, J. Bourdeau, and R. Mizoguchi, Eds., *Advances in Intelligent Tutoring Systems*, 2010.
- [5] R. H. Stevens, T. Galloway, and C. Berka, "EEG-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills," in *User Modeling*, 2007.
- [6] A. Zammouri, A. A. Moussa, and S. Chevallier, "Use of cognitive load measurements to design a new architecture of intelligent learning systems," *Expert Syst. Appl.*, 2024.
- [7] K. F. Van Orden, W. Limbert, S. Makeig, and T. P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human Factors*, 2001.
- [8] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *Int. J. Aerosp. Psychol.*, 2002.
- [9] M. Chaouachi, I. Jraid, and C. Frasson, "MENTOR: A physiologically controlled tutoring system," in *User Modeling, Adaptation and Personalization*, 2015.
- [10] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis," *Journal of Educational Psychology*, 2014.
- [11] T. Xu, Y. Zhou, Z. Wang, and Y. Peng, "Learning emotions EEG-based recognition and brain activity: A survey study on BCI for intelligent tutoring system," *Procedia Computer Science*, 2018.
- [12] A. Doulou, P. Pergantis, A. Drigas, and C. Skianis, "Managing ADHD symptoms in children through the use of various technology-driven serious games: A systematic review," *Multimodal Technol. Interact.*, 2025.
- [13] H. Kim, Y. Chae, S. Kim, and C.-H. Im, "Development of a computer-aided education system inspired by face-to-face learning by incorporating EEG-based neurofeedback into online video lectures," *IEEE Trans. Learn. Technol.*, 2023.
- [14] L. Pillette, C. Jeunet, B. Mansencal, R. N'Kambou, B. N'Kaoua, and F. Lotte, "A physical learning companion for mental-imagery BCI user training," *IJHCS*, 2020.
- [15] C.-Y. Chou, T.-W. Chan, and C.-J. Lin, "Redefining the learning companion: The past, present, and future of educational agents," *Computers & Education*, 2003.
- [16] J. Prinsen, E. Pruss, A. Vrans, C. Ceccato, and M. Alimardani, "A passive brain-computer interface for monitoring engagement during robot-assisted language learning," in *IEEE SMC*, 2022.
- [17] D. Szafir and B. Mutlu, "Pay attention! designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.
- [18] J. Higgins, S. Eldridge, and L. T., "Chapter 23: Including variants on randomized trials," *Cochrane Handbook for Systematic Reviews of Interventions - version 6.5*, 2019.
- [19] T. Xu, X. Wang, J. Wang, and Y. Zhou, "From textbook to teacher: An adaptive intelligent tutoring system based on BCI," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 2021.
- [20] J. Frey, R. Gervais, S. Fleck, F. Lotte, and M. Hachet, "Teegi: Tangible EEG interface," in *27th ACM symposium on User interface software and technology*, 2014.
- [21] R. Francese, M. Buccino, F. Blaso, and L. De Santis, "Adapting educational content to student emotions: A user study with the furhat social robot," *IEEE TCSS*, 2025.
- [22] N. Gudikandula, R. Janapati, R. Sengupta, and S. Chintala, "Brain computer interface based emotion recognition with error analysis and challenges: An interdisciplinary review," *Discover Applied Sciences*, 2025.
- [23] J. Hassan, S. Reza, S. Ahmed, N. H. Anik, and M. Obaydullah, "EEG workload estimation and classification: A systematic review," *JNE*, 2025.
- [24] H. Hinrichs, M. Scholz, A. Baum, J. Kam, R. Knight, and H.-J. Heinze, "Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG system for clinical applications," *Scientific Reports*, 2020.
- [25] N. M. Ehrhardt, C. Niehoff, A.-C. Oßwald, D. Antonenko, G. Lucchese, and R. Fleischmann, "Comparison of dry and wet electroencephalography for the assessment of cognitive evoked potentials and sensor-level connectivity," *Frontiers in Neuroscience*, 2024.

A controlled evaluation of dry textile EEG electrodes using simulated P300 and SSVEP recordings with a gelatin-based skin phantom

Jason Leung¹, Faiz Uddin Qureshi^{1,2}, Tom Chau^{1,2}

¹ Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada

² Institute of Biomedical Engineering, University of Toronto, Toronto, Canada

E-mail: tom.chau@utoronto.ca

ABSTRACT: Dry textile electrodes offer reduced setup time and improved user comfort, making them attractive for everyday electroencephalography (EEG) recordings. However, dry electrodes typically exhibit higher electrode-skin impedance, so their signal quality must be carefully evaluated for brain-computer interface (BCI) applications. In this study, dry textile EEG electrodes were evaluated using a gelatin-based skin phantom. Simulated EEG recordings during P300 and steady-state visual evoked potential (SSVEP) tasks were delivered through electrodes embedded in the phantom model. Signals recorded from textile electrodes placed on the surface of the phantom show strong agreement with those from Ag-AgCl electrodes. P300 components were clearly observable in the event-related potentials, and distinct spectral peaks were detected at the stimulation frequencies in the SSVEP recordings. These results suggest that dry textile electrodes provide signal quality comparable to conventional Ag-AgCl electrodes under controlled conditions, warranting further investigation into their feasibility for BCI applications.

INTRODUCTION

Electroencephalography (EEG) is widely used in brain-computer interface (BCI) systems due to its non-invasive nature, relatively low cost, and high temporal resolution [1-3]. In conventional BCI systems, EEG signals are acquired using silver/silver-chloride (Ag-AgCl) electrodes together with a conductive gel or electrolyte to reduce electrode-skin impedance. While these electrodes provide high-quality recordings, their use requires skin preparation and gel application, which increases setup time and limits their suitability for long-term or continuous BCI use [4-6].

Dry electrodes have emerged as a promising alternative because they do not require an electrolytic medium and thus reduce setup time [7,8]. Textile-based electrodes offer flexibility and integration with wearable garments, making them well suited for everyday continuous recordings [9-11]. However, dry textile electrodes are often reported to exhibit higher electrode-skin impedance and suffer from reduced signal quality compared to gel-based electrodes [12]. As such, careful evaluation of the signal quality of dry textile electrodes is necessary for long term BCI systems.

While previous studies have demonstrated the

feasibility of textile sensors, validation often relied on human trials where physiological noise and contact variability make it difficult to isolate the intrinsic electrical performance of the textile-to-skin interface and [7,13]. Specifically, physiological variability in scalp anatomy, hair density, and user mental states can obscure subtle differences in sensor impedance and frequency response [13-15]. To address this, gelatin-based skin phantoms have been proposed for high-fidelity, repeatable benchtop testing of EEG electrodes since they exhibit electrical properties similar to those of human skin [16]. While various head models exist, gelatin-based phantoms offer an adjustable level of conductivity that closely mimics human skin [16]. In this study, we utilize a gelatin-based skin phantom to provide a controlled environment for quantitative characterization of dry textile electrodes. While this approach does not account for ecological factors like hair or motion artifacts, it provides a necessary baseline validation of signal quality between dry textile electrodes and Ag-AgCl electrodes.

MATERIALS AND METHODS

Textile electrode fabrication: Textile electrodes were fabricated by knitting a 1 cm by 1 cm patch of conductive threads (carbon-impregnated yarns and silver-plated fibers) into an elastic nylon-Lycra substrate (Fig. 1). The electrodes were connected to snap button connectors via conductive traces.

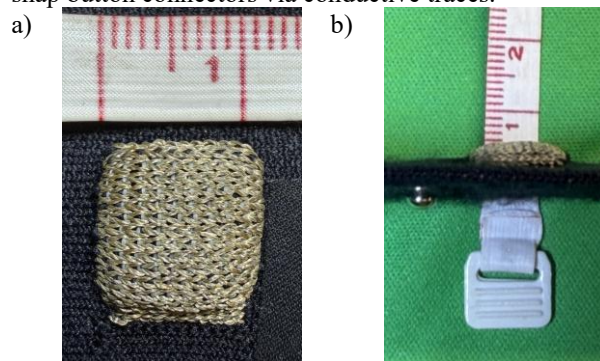


Figure 1: a) Top and b) side view of textile electrode fabricated through Myant proprietary Textile Computing™ (linear mass density of 100-200 Denier, linear resistance <100 Ω /m)

Skin phantom: A skin phantom was prepared according to Owda et al. [16] by mixing 9 g of sodium chloride and 200 g of gelatin powder in 1000 g of deionized water. The mixture was heated in a microwave for 2 minutes and stirred to eliminate clumping. Subsequently, 1 mL of dimethicone and 3 mL of Germall Plus were added as an antifoaming and antimicrobial agent, respectively. The solution was poured into a 1 L plastic food container (21 cm x 15 cm x 7 cm) with two embedded electrodes (3.5 mm gold-plated audio jacks) fixated 10 cm apart at the base of container (Fig. 2). The phantom was cooled at 3°C for 48 hours to allow for complete solidification.



Figure 2: a) Top and b) side view of gelatin-based skin phantom with embedded electrodes

Simulated EEG recordings: To validate the signal quality of the textile electrodes, raw unfiltered EEG recordings were extracted from a public dataset by Lee et al. [17]. These recordings comprise of EEG signals collected during a visual P300 task and a steady-state visually evoked potential (SSVEP) task. In the P300 task, participants selected a target amongst 36 symbols over a total of 1980 trials (330 target, 1650 non-target). In the SSVEP task, participants were instructed to attend one of four target frequencies (5.45 Hz, 6.67 Hz, 8.57 Hz, 12.0 Hz). A total of 100 trials (25 trials per target frequency) were completed. Data from participant 6 in session 2 was selected due to high classification performances reported during this session [17]. The Oz channel was specifically selected for simulation as it is the primary site for recording visual evoked potentials and typically exhibits the highest signal-to-noise ratio (SNR) for these paradigms.

The simulated EEG signals were generated using the 3.3V audio output of a Lenovo Legion Pro 5 laptop. To translate the digital signal into the microvolt range suitable for EEG amplification, a scaling factor was applied. A total gain of 36 was chosen to compensate for signal attenuation through the gelatin medium. This ensured the potential difference at the phantom surface

remained within the typical physiological range (100 μ V), preventing amplifier saturation while maintaining a signal magnitude comparable to real-world recordings. The signals were delivered between two embedded electrodes (output and ground) at the base of the phantom. Signals were recorded simultaneously from textile electrodes and flat snap Ag-AgCl electrodes (OpenBCI) placed on the surface of the gel. Recording electrodes were positioned around the output embedding electrode, while reference and ground electrodes were centered around the ground embedding electrode. A 100 g load was applied to all surface electrodes to ensure uniform pressure application. All signals were collected at a sampling rate of 500 Hz using a wireless EEG amplifier (Liveamp Brain Products, Germany) with an analogue differential DC amplifier (BIP2AUX adapters, Brain Products, Germany). All recordings were made at an ambient temperature of 21°C and a relative humidity of 32%.

Signal analysis: Signals recorded from surface textile and Ag-AgCl electrodes were filtered between 0.3 to 40 Hz. A notch filter was applied at 60 Hz to remove electrical line noise.

For the P300 recording, epochs were extracted from -200 to 1000 ms relative to the visual cue onset. Baseline correction was applied by subtracting the mean of the pre-stimulus interval for each channel. Trial-level agreement was quantified using the root mean-squared error (RMSE) and the Pearson correlation between textile and Ag-AgCl signals.

For the SSVEP task, 4-second epochs were extracted and baseline-corrected using the mean of the entire trial duration for each channel. Power spectral density (PSD) was estimated using the Welch method. The SNR for each target frequency, f_{target} , was computed according to Equation 1. The signal power was defined as the PSD within the 1 Hz band centered around the stimulus frequency, while the noise power was defined as the broadband power between 0.5 Hz and 15 Hz (excluding the signal band).

$$SNR_{f_{target}} = 10 \log \frac{\overline{P_{signal}}}{\overline{P_{noise}}} \quad (1)$$

To facilitate comparison between textile and Ag-AgCl signals, we compute the relative SNR (rSNR) by normalizing the SNR of the textile signals by the SNR of the Ag-AgCl signals (Equation 2).

$$rSNR = \frac{SNR_{textile}}{SNR_{Ag-AgCl}} \quad (2)$$

RESULTS

P300 recordings: Fig. 3 plots the event-related potential (ERP) response from the textile and Ag-AgCl recording during the P300 task. A distinct P300 component can be observed in both ERP signals. The ERP waveform of the textile electrode showed a strong correspondence to that of the Ag-AgCl electrode.

Textile and Ag-AgCl signals had a trial-wise RMSE of $1.10 \pm 0.61 \mu\text{V}$ (mean $\pm$ std dev) and a mean Pearson correlation of 0.9999 ($p < 0.001$), suggesting a high degree of agreement.

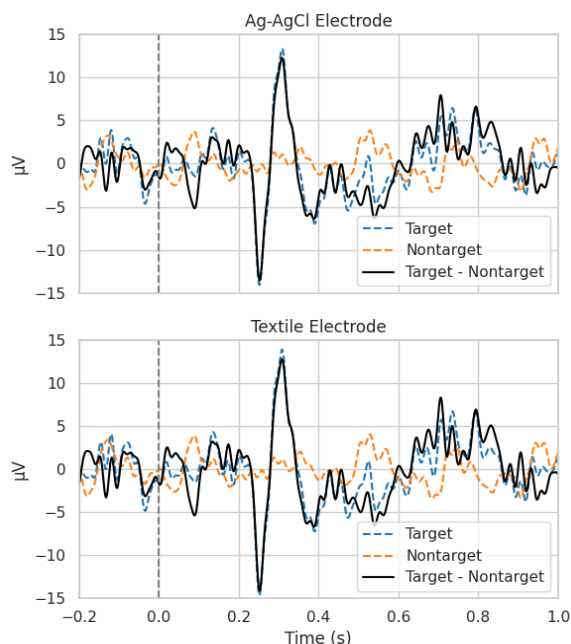


Figure 3: Event related potential response for Ag-AgCl and textile electrodes during simulated P300 recordings

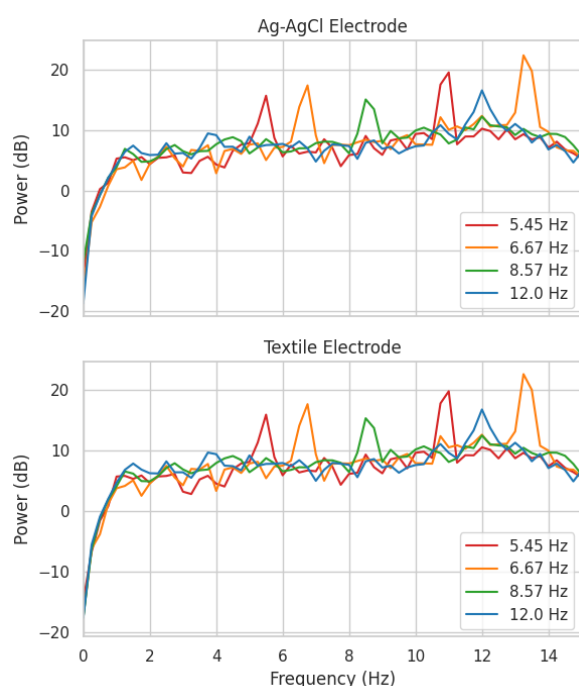


Figure 4: Power spectral density of Ag-AgCl and textile signals during simulated SSVEP recordings

SSVEP recordings: PSDs for each target stimulation frequency during the SSVEP task are plotted in Fig. 4. An increase in spectral energy was observed at the target frequencies in both textile and Ag-AgCl electrodes. Tab. 1 summarizes the SNR from each electrode and the rSNR across all stimulation

frequencies. The mean rSNR remained close to unity across all conditions, indicating strong correspondence between textile and Ag-AgCl recordings during steady-state visual stimulation.

Table 1: Signal-to-noise ratio (mean $\pm$ std dev) of Ag-AgCl and textile electrode during the SSVEP recording

Target Frequency (Hz)	Ag-AgCl Electrode SNR (dB)	Textile Electrode SNR (dB)	rSNR
5.45	7.12 (4.17)	7.08 (4.17)	0.995 (0.014)
6.67	8.79 (4.11)	8.77 (4.10)	0.997 (0.015)
8.57	5.81 (5.28)	5.86 (5.30)	1.021 (0.093)
12.00	9.10 (8.50)	9.05 (8.67)	0.996 (0.022)

DISCUSSION

Signal quality during simulated EEG: Textile signals closely matched Ag-AgCl signals across both simulated P300 and SSVEP recordings. In the P300 task, low RMSE values indicate strong trial-level correspondence between textile and Ag-AgCl recordings, and averaged ERP waveforms revealed clear P300 components in both electrode types, suggesting that dry textile electrodes can capture task-relevant physiological features to facilitate BCI classification.

For the SSVEP task, both textile and Ag-AgCl electrodes exhibited distinct spectral peaks at the target stimulation frequencies and their harmonics. rSNR values remained near unity, suggesting that textile electrodes are comparable in performance to Ag-AgCl electrodes, and can reliably neural responses in the frequency domain. The clear separation between target frequencies further supports the feasibility of using dry textile electrodes for SSVEP-based decoding.

Limitations and future directions: The results demonstrate that dry textile electrodes can replicate both the temporal and spectral characteristics of benchmark Ag-AgCl electrodes under controlled conditions. These findings support the potential of textile electrodes as viable alternatives for EEG recording, particularly in use cases where comfort, wearability, and long-term use are important.

While the present study focused on comparing the signal quality of dry textile electrodes against gold-standard Ag-AgCl electrodes in a controlled phantom environment, further validation is needed to evaluate their performance under realistic recording conditions. Future studies should assess performance in human subjects, where factors such as hair coverage, perspiration, and motion-related artifacts may influence electrode contact and signal stability. Such evaluations will be important for determining the reliability of textile electrodes in practical BCI applications and for guiding the design of wearable EEG systems intended for everyday use.

CONCLUSION

This study presented a systematic characterization of

dry textile EEG electrodes using a gelatin-based skin phantom. By using a controlled phantom setup, we isolated the intrinsic electrical performance of the textile electrode from the biological variability inherent in human trials. Across simulated P300 and SSVEP paradigms, textile signals demonstrated strong correspondence with signals from Ag-AgCl electrodes, supported by high correlation coefficients and consistent spectral power densities at target frequencies. These results suggest that dry textile electrodes are capable of delivering signal quality comparable to conventional sensors in controlled environments, representing a key step toward their integration into wearable, long-term BCI applications.

REFERENCES

- [1] M. Orban, M. Elsamanty, K. Guo, S. Zhang and H. Yang, "A Review of Brain Activity and EEG-Based Brain-Computer Interfaces for Rehabilitation Application," *Bioengineering*, vol. 9, no. 12, p. 768, 12 2022.
- [2] I. Lazarou, S. Nikolopoulos, P. C. Petrantonis, I. Kompatsiaris and M. Tsolaki, "EEG-Based Brain-Computer Interfaces for Communication and Rehabilitation of People with Motor Impairment: A Novel Approach of the 21st Century," *Frontiers in Human Neuroscience*, vol. 12, p. 14, 1 2018.
- [3] W. A. Awuah, A. Ahluwalia, K. Darko, V. Sanker, J. K. Tan, P. O. Tenkorang, A. Ben-Jaafar, S. Ranganathan, N. Aderinto, A. Mehta, M. H. Shah, K. Lee Boon Chun, T. Abdul-Rahman and O. Atallah, "Bridging Minds and Machines: The Recent Advances of Brain-Computer Interfaces in Neurological and Neurosurgical Applications," *World Neurosurgery*, vol. 189, pp. 138-153, 9 2024.
- [4] A. Searle and L. Kirkup, "A direct comparison of wet, dry and insulating bioelectric recording electrodes," *Physiological measurement*, vol. 21, no. 2, pp. 271-283, 2000.
- [5] Y. M. Chi, T. P. Jung and G. Cauwenberghs, "Dry-contact and noncontact biopotential electrodes: Methodological review," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 106-119, 2010.
- [6] M. Zhang, B. Qian, J. Gao, S. Zhao, Y. Cui, Z. Luo, K. Shi, E. Yin, M. Zhang, B. Qian, J. Gao, S. Zhao, Y. Cui, Z. Luo, K. Shi and E. Yin, "Recent Advances in Portable Dry Electrode EEG: Architecture and Applications in Brain-Computer Interfaces," *Sensors* 2025, Vol. 25,, vol. 25, no. 16, 8 2025.
- [7] S. M. Rahman, H. Mattila, M. Janka and J. Virkki, "Impedance evaluation of textile electrodes for EEG measurements," *Textile Research Journal*, vol. 93, no. 7-8, pp. 1878-1888, 4 2023.
- [8] G. L. Li, J. T. Wu, Y. H. Xia, Q. G. He and H. G. Jin, "Review of semi-dry electrodes for EEG recording," *Journal of Neural Engineering*, vol. 17, no. 5, 10 2020.
- [9] C.-W. Kan, Y.-L. Lam, C.-W. Kan and Y.-L. Lam, "Future Trend in Wearable Electronics in the Textile Industry," *Applied Sciences* 2021, Vol. 11,, vol. 11, no. 9, 4 2021.
- [10] G. B. Tseghai, B. Malengier, K. A. Fante and L. Van Langenhove, "The Status of Textile-based Dry EEG Electrodes," *Autex Research Journal*, vol. 21, no. 1, pp. 63-70, 3 2021.
- [11] E. Lam, M. Alizadeh-Meghrizi, A. Schlums, L. Eskandarian, A. Mahnam, B. Moineau and M. R. Popovic, "Exploring textile-based electrode materials for electromyography smart garments," *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 9, p. 20556683211061995, 6 2022.
- [12] C. M. Vidhya, Y. Maithani, J. P. Singh, C. M. Vidhya, Y. Maithani and J. P. Singh, "Recent Advances and Challenges in Textile Electrodes for Wearable Biopotential Signal Monitoring: A Comprehensive Review," *Biosensors* 2023, Vol. 13,, vol. 13, no. 7, 6 2023.
- [13] L. Kalevo, S. Myllymaa, T. Miettinen, A. Leino, S. Kainulainen, H. Korkalainen, K. Myllymaa, J. Toyra, T. Leppanen and T. Laitinen, "Effect of sweating on electrode-skin contact impedances and artifacts in EEG recordings with various screen-printed Ag/AgCl electrodes," *IEEE Access*, vol. 8, pp. 50934-50943, 2020.
- [14] T. Lees, N. Ram, M. M. Swingle and L. M. Gatzke-Kopp, "The effect of hair type and texture on electroencephalography and event-related potential data quality," *Psychophysiology*, vol. 61, no. 3, 3 2024.
- [15] V. Mihajlovic, H. Li, B. Grundlehner, J. Penders and A. C. Schouten, "Investigating the impact of force and movements on impedance magnitude and EEG," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2013, pp. 1466-1469, 2013.
- [16] A. Y. Owda, A. J. Casson, A. Y. Owda and A. J. Casson, "Electrical properties, accuracy, and multi-day performance of gelatine phantoms for electrophysiology," *bioRxiv*, p. 2020.05.30.125070, 5 2020.
- [17] M. H. Lee, O. Y. Kwon, Y. J. Kim, H. K. Kim, Y. E. Lee, J. Williamson, S. Fazli and S. W. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, pp. 1-16, 5 2019.

Movement-Related Interhemispheric Connectivity Dynamics and their Link to Asynchronous EEG-BCI Performance

Kyriaki Kostoglou¹, Markus R. Crell¹, G. R. Müller-Putz^{1,2}

¹ Institute of Neural Engineering, Graz University of Technology, Graz, Austria

² BiotechMed, Graz, Austria

E-mail: kkostoglou@tugraz.at

ABSTRACT: Directed interactions between bilateral sensorimotor cortices are important for voluntary movement, yet their relevance for asynchronous brain-computer interface (BCI) control remains unclear. Using EEG, we quantified time-frequency directional connectivity between sensorimotor regions during cue-based and self-paced hand movements in 25 participants using time-varying multivariate autoregressive models. Connectivity from the contralateral to the ipsilateral cortex was consistently stronger and positively associated with movement detection performance. Connectivity increased in low-frequency bands around movement onset and in the beta band after movement execution. These findings suggest that interhemispheric EEG connectivity reflects motor network coordination and may serve as a neurophysiological marker of successful asynchronous BCI control.

INTRODUCTION

Restoring reliable communication for individuals with severe motor impairments is a key goal of brain-computer interface (BCI) research. Movement-related BCIs provide an intuitive approach by relying on endogenous cortical activity linked to voluntary motor preparation and execution. By translating motor intent into control commands, these systems enable self-paced, asynchronous operation without external stimulation, preserving user autonomy and interface flexibility.

Two complementary neurophysiological signatures underpin motor-based BCIs. Movement-related cortical potentials (MRCPs) are slow negative shifts appearing up to ~2 s before movement onset, reflecting premotor and supplementary motor area activation [1]. Event-related desynchronization/synchronization (ERD/ERS) [2] represents alpha and beta power modulations over sensorimotor cortex during movement preparation and execution. While MRCPs provide robust time-domain markers of impending movement, ERD/ERS reflects oscillatory dynamics of sensorimotor networks. Both have been widely used for movement-intent detection and asynchronous BCI control [3–12].

We recently developed a fully asynchronous MRCP-based EEG speller that enabled reliable online communication in healthy participants, detecting self-initiated hand movements without external cues [3,4]. Performance in a self-paced validation run strongly predicted online spelling performance ($r \approx 0.81$), and high-performing participants showed stronger MRCP

negativity and greater contralateral sensorimotor contributions. These results suggest that individual neurophysiological characteristics influence BCI controllability and motivate investigation of the neural determinants of successful control. Voluntary movement involves coordinated interactions between bilateral sensorimotor cortices, yet most BCI studies focus on local signal features such as MRCP amplitude or ERD magnitude. Although connectivity measures have been used as classifier features [13], particularly in motor imagery paradigms, they are rarely examined as neurophysiological variables explaining differences in asynchronous BCI control. For sensorimotor rhythm couplings, various connectivity measures have been proposed, including phase-based metrics (e.g., phase-locking value) [13,14], coherence-based approaches [13], and other synchronization indices. Among these, multivariate autoregressive (MVAR) modelling [15–19] is widely used to study directed interactions in EEG [20]. Frequency-domain measures such as directed coherence estimate directional information flow [21–24] and can reveal asymmetries between contralateral and ipsilateral motor areas. However, it remains unclear whether interhemispheric directional interactions during movement are consistent across individuals, whether they exhibit systematic time-varying patterns before, during, and after movement, or whether they relate to successful asynchronous BCI control.

This study addresses this gap by quantifying time-frequency directed connectivity between left and right sensorimotor regions using MVAR-based directed coherence (DC). We examine whether interhemispheric directionality is consistent across cue-based and self-paced conditions and whether it relates to local motor signatures (MRCP magnitude, ERD/ERS strength) and the BCI's ability to detect movements. Our goal is to determine whether connectivity provides a stable neurophysiological marker of successful motor-based asynchronous control.

MATERIALS AND METHODS

Experimental Paradigm: In our previous study [3,4] we evaluated a fully asynchronous MRCP-based EEG BCI for online spelling in 25 healthy participants (8 female, mean age 28.1 ± 7.5 years). Participants completed an eye-artifact recording block followed by calibration and online spelling sessions. Calibration included a cue-based movement round with visually

prompted right hand movements (i.e., fist with wrist extension, average inter-trial interval 3.3 s) and a self-paced round in which the same movements were executed without external cues (average inter-trial interval 4.13 s). The cue-based round contained 300 movements, whereas the number of self-paced movements varied across participants, ranging from 50 to 187. A detection model trained during calibration identified MRCPs associated with self-initiated movements during the online spelling task, where detected movements triggered character selections. EEG was recorded at 500 Hz from 60 electrodes (10–10 system) using a BrainAmp amplifier (BrainProducts, GmbH), with four electrooculography (EOG) channels for artifact correction and a motion capture system to determine movement onsets.

EEG Preprocessing: For the present analysis, only the calibration data were used. EEG signals were cleaned of 50 Hz line noise and band-pass filtered (0.5–60 Hz, 4th-order Butterworth), and ocular artifacts were removed using an EOG regression-based approach based on a dedicated eye block. The EEG was resampled to 100 Hz, and movement onset events from motion capture were downsampled accordingly to preserve temporal alignment. In addition, a binary movement trace indicating the presence or absence of movement at each time point was derived from the motion capture data.

Connectivity Analysis: To characterize cortical interactions around movement onset, connectivity was estimated using a time-varying multivariate autoregressive (MVAR) framework with overlapping windows. EEG epochs (channels $\times$ time $\times$ trials) were segmented around movement onset, and for each window a single MVAR model was fitted across all trials using the ARFIT algorithm (window length: 1 s, step: 0.125 s) [25]. Model order was selected using the Akaike Information Criterion, and the median optimal order across windows and participants was used for the final analysis to ensure consistent model complexity (7th order). The estimated MVAR model was transformed into the frequency domain to compute spectral directed coherence using the eMVAR toolbox from Faes et al. [26]. DC values from each window were assigned to all samples within that window; overlapping estimates were averaged to obtain smooth time–frequency connectivity maps across the epoch. Because the self-paced round contained a variable number of movements, the minimum number of valid trials (50) was used. For each participant, 50 self-paced movements were selected, while the 300 cue-based movements were split into six sets of 50 trials, yielding six cue-based and one self-paced connectivity estimate per participant. The cue-based estimates were then averaged, ensuring equal trial numbers across conditions for fair comparison.

The resulting DC estimates form time–frequency maps for all channel pairs, representing directional interactions between EEG channels. In this study, we focused on interhemispheric connectivity between the left and right sensorimotor channel groups (C_L and C_R), analyzing directional interactions from $C_L \rightarrow C_R$ and

$C_R \rightarrow C_L$. For frequency-specific analysis, DC values were averaged within canonical bands (δ : 1–3.5 Hz, θ : 4–7 Hz, α : 8–12 Hz, β : 13–30 Hz, γ : 30–50 Hz) to obtain band-limited time series. These were used to examine interhemispheric connectivity dynamics and compare cue-based and self-paced conditions. Subject-level DC values at each time point were entered into a linear mixed-effects (LME) model [27]:

$$Y \sim \text{Condition} + (1 \mid \text{Subject}) \quad (1)$$

where Y denotes the DC value, *Condition* indicates the experimental condition (cue-based or self-paced) and the term $(1 \mid \text{Subject})$ represents the subject-specific random intercept capturing between-subject variability. The resulting p-values indicated whether DC differed significantly between conditions and were corrected for multiple comparisons across time, frequency bands and connectivity direction using the Benjamini–Hochberg false discovery rate (FDR) procedure [28].

ERD/ERS and MRCP Analysis: In addition to connectivity measures, ERD/ERS maps and MRCP waveforms were estimated for each participant. ERD/ERS maps were computed based on [2], using a baseline interval of –4 to 0 s relative to movement onset. This enabled characterization of time–frequency power modulations associated with movement preparation and execution. MRCP waveforms were obtained by averaging the time-domain EEG signals across trials to capture slow cortical potentials around movement onset.

Linking Directional Connectivity to Self-Paced Decoding Performance: As described previously [3,4], movement detection performance based on MRCP features was quantified using the Matthews Correlation Coefficient (MCC) [29], computed from self-paced movement predictions alongside metrics such as true positive rate and false positives per minute. Here, we focused on MCC as a balanced measure of detection accuracy and false positives. To test whether performance was associated with directional connectivity, linear mixed-effects (LME) models were fitted at each time–frequency bin of the DC maps, separately for each connectivity direction:

$$Y \sim \text{MCC} + (\text{Condition} \mid \text{Subject}) \quad (2)$$

where Y denotes the DC value, *MCC* represents the subject-level classifier performance metric. The term $(\text{Condition} \mid \text{Subject})$ specifies a random intercept for each subject together with a random slope for condition (cue-based vs self-paced), allowing connectivity differences between conditions to vary across participants and accounting for the repeated-measures structure of the data. *Condition* was modeled as a random effect to control for within-subject variability between cue-based and self-paced connectivity estimates while assessing the association between connectivity and decoding performance. Statistical significance of the MCC effect was evaluated across all bins and FDR corrected for multiple comparisons.

Temporal Coupling Between Interhemispheric Connectivity and Sensorimotor Activity: To examine temporal relationships between connectivity and

sensorimotor activity, correlations were computed between DC, ERD/ERS, and MRCP time courses. Band-specific DC trajectories were analyzed separately for cue-based and self-paced conditions and for each connectivity direction. ERD/ERS time series were obtained from left and right sensorimotor channel groups by averaging ERD/ERS maps within the alpha and beta bands, while MRCP time series were derived from the corresponding averaged sensorimotor waveforms. All trajectories were computed from group-average signals to capture dominant temporal dynamics around movement. Pearson correlations between DC, ERD/ERS, and MRCP trajectories produced matrices describing connectivity–activation coupling. P-values were corrected for multiple comparisons, and non-significant correlations were set to zero for visualization. Cue-based and self-paced correlations were compared using Fisher’s r-to-z transformation [30].

RESULTS

Sensorimotor Oscillatory Dynamics: Figure 1a,b shows the grand-average EEG dynamics during cue-based and self-paced movements, including ERD/ERS maps, MRCP time courses, and time–frequency directional connectivity (DC) between the left (C_L) and right (C_R) sensorimotor regions. The average binary movement trace from motion capture is overlaid in all panels, indicating the probability and timing of movement across trials. ERD/ERS maps (Fig. 1a) showed clear sensorimotor rhythm modulations around movement onset in both hemispheres, with alpha and low-beta power (~ 8 – 20 Hz) decreasing before and during movement and a post-movement beta ERS 1–2 s afterward. MRCP waveforms (Fig. 1b) displayed the typical slow negative potential preceding movement onset and returning to baseline within 1–2 s. In the self-paced condition, ERD/ERS and MRCP patterns were more diffuse and temporally variable.

Directional Connectivity Dynamics: The directional connectivity maps (Fig. 1c) revealed structured time–frequency interactions between the two sensorimotor regions. In both directions and conditions, connectivity increased in low frequencies (< 10 Hz) and especially in the delta band around movement onset. After movement execution, connectivity increased in the beta band, temporally coinciding with the beta rebound observed in the ERD/ERS maps. Connectivity from the left (i.e., contralateral) to the right (i.e., ipsilateral) hemisphere ($C_L \rightarrow C_R$) was generally stronger than in the opposite direction (Fig. 1c,d). To quantify this effect, a single average DC value was extracted from the connectivity maps for each participant and direction. In both cue-based and self-paced movements, $C_L \rightarrow C_R$ connectivity was significantly stronger than $C_R \rightarrow C_L$ connectivity ($p < 0.001$). This directional asymmetry was consistent across participants, as illustrated by the violin plots of Fig. 1d showing individual subject values. In contrast, connectivity in the $C_R \rightarrow C_L$ direction was weaker and less structured, with more diffuse low-frequency

fluctuations. Band-specific time courses (Fig. 1e) further showed that the most pronounced differences between cue-based and self-paced conditions occurred in the delta band around movement onset, with a smaller but significant beta-band difference after movement, aligned with the beta rebound. Overall, connectivity profiles across frequency bands were similar between conditions. The main difference arose from movement timing, as the longer and more irregular inter-trial intervals in the self-paced condition produced broader transitions and greater variability in the averaged movement trace.

Relationship Between Connectivity and Decoding Performance: The bottom row of Fig. 1f shows the t-statistics from the LME models relating the DC time–frequency maps to detection performance (MCC) for both connectivity directions. Higher t-statistics indicate stronger positive associations between connectivity and MCC. Overall, higher connectivity in the $C_L \rightarrow C_R$ direction appeared to be associated with higher MCC values, suggesting that stronger interhemispheric coupling from the left to the right sensorimotor cortex may support improved detection performance. Green contours mark clusters where the corresponding p-values remained significant after FDR correction. Significant clusters were observed only in the $C_L \rightarrow C_R$ direction, in the beta band (~ 20 – 25 Hz) both before (~ 2 s) and after movement onset, indicating that stronger connectivity in this direction was associated with higher detection performance. In contrast, no significant relationship between connectivity and MCC was observed for the $C_R \rightarrow C_L$ direction, although the strongest non-significant associations occurred mainly in the 10–20 Hz range.

Temporal Coupling Between Connectivity and Sensorimotor Activity: To investigate the temporal relationships between interhemispheric connectivity and sensorimotor activity, we correlated the time evolution of MRCP signals and ERD/ERS activity in the alpha and beta bands with the band-specific DC time series. The ERD/ERS trajectories were extracted from the ERD/ERS maps by averaging within the alpha and beta bands, while MRCP time courses were obtained from the averaged sensorimotor waveforms of the left (C_L) and right (C_R) regions. These physiological time series were then temporally correlated with DC trajectories computed for the canonical frequency bands and for both connectivity directions ($C_L \rightarrow C_R$ and $C_R \rightarrow C_L$). Figure 2 summarizes the correlations between band-specific DC time series and physiological signals (ERD/ERS and MRCP) for cue-based and self-paced movements. Overall, DC trajectories correlated more strongly with ERD/ERS than with MRCP signals in both conditions, particularly for beta-band ERD/ERS. In the cue-based condition, clear frequency-specific patterns emerged: beta-band ERD/ERS in both hemispheres showed strong positive correlations with alpha and beta DC, while delta connectivity showed negative correlations with ERD/ERS. In contrast, the self-paced condition showed reduced alpha- and beta-band ERD/ERS–connectivity correlations in the $C_R \rightarrow C_L$ direction. For $C_L \rightarrow C_R$, coupling shifted toward correlations with both alpha- and

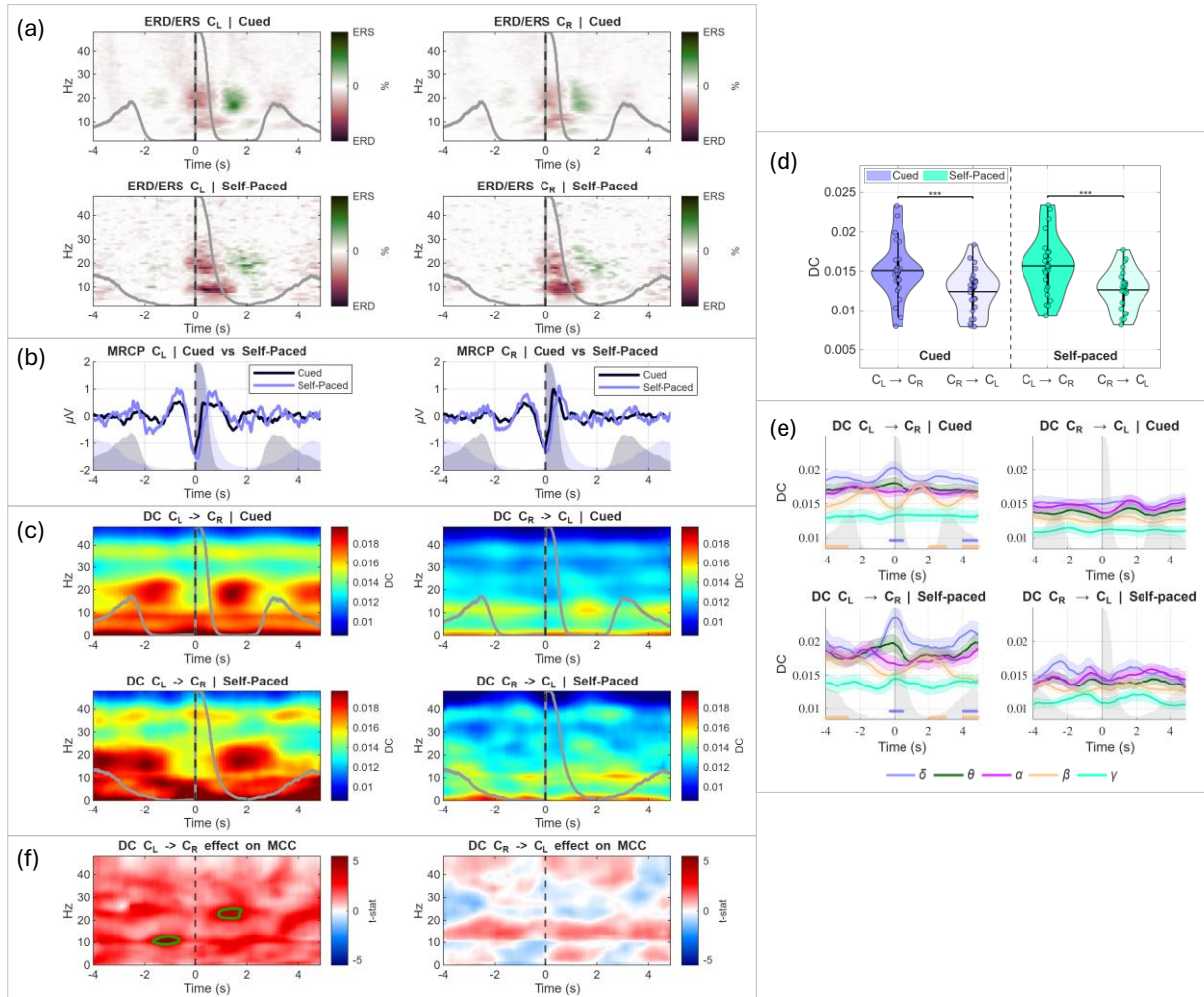


Figure 1. Grand-average (a) ERD/ERS maps and (b) MRCP patterns for the left (CL) and right (CR) sensorimotor regions during cue-based (top panel) and self-paced (bottom panel) movements. (c) Grand-average time–frequency DC maps for both directions (CL → CR left column and CR → CL right column) for cue-based (top row) and self-paced (bottom row) movements, estimated using a sliding-window MVAR framework. (d) Violin plots showing the distribution of average DC values across participants for both connectivity directions and conditions. (e) Band-specific (δ , θ , α , β , γ) DC time courses for the two connectivity directions during cue-based (top row) and self-paced (bottom row) movements. (f) Relationship between directional connectivity and decoding performance for both connectivity directions (left and right columns), expressed as t-statistics from LME models relating DC to the MCC obtained from the MRCP-based decoder in the self-paced round. Red indicates positive associations and blue negative associations between connectivity and decoding performance. Significant regions after FDR correction are outlined in green. The gray curve overlaid in panels (a, c) and the shaded areas in panels (c, e) represent the average binary movement trace derived from motion capture recordings, indicating the probability of movement across trials. Time 0 is the movement onset.

beta-band ERD/ERS, producing a more distributed connectivity–activation pattern across frequency bands.

DISCUSSION

Sensorimotor Oscillatory Dynamics During Movement Preparation: Consistent with previous studies, both cue-based and self-paced movements showed typical sensorimotor rhythm modulation, with alpha and low-beta desynchronization before movement and a post-movement beta rebound. MRCPs also exhibited the characteristic slow negative potential preceding movement onset. While patterns were similar across conditions, self-paced movements showed slightly greater variability, likely reflecting their more irregular timing and longer inter-trial intervals.

Interhemispheric Connectivity Dynamics: Directional connectivity revealed structured time–frequency interactions between the left and right sensorimotor cortices. In both directions, low-frequency connectivity, especially in the delta band, increased around movement onset, likely reflecting slow cortical potentials underlying MRCPs. In contrast, beta-band connectivity (particularly lower beta) increased after movement, coinciding with the beta rebound in ERD/ERS maps and consistent with stronger interhemispheric coupling during post-movement beta synchronization [31]. These results suggest that interhemispheric interactions follow the temporal structure of local oscillatory dynamics, linking connectivity changes to motor-related processes [2,32]. Prior studies comparing internally and externally paced finger movements reported strongest task-related

differences in interhemispheric coupling in the beta band (~20–22 Hz) [33], further supporting our results.

A key finding in our study was the directional asymmetry of interhemispheric connectivity, with connectivity from the left to the right sensorimotor cortex ($C_L \rightarrow C_R$) consistently stronger than in the opposite direction. Because participants performed right-hand movements, the left sensorimotor cortex represents the contralateral motor area responsible for movement generation. The stronger $C_L \rightarrow C_R$ connectivity likely reflects propagation of motor-related activity from the contralateral cortex to ipsilateral motor areas via interhemispheric pathways, consistent with ipsilateral involvement in unilateral movement through transcallosal interactions [34]. This directional asymmetry also argues against a purely power-driven explanation, where connectivity increases arise mainly from higher EEG signal power or signal-to-noise ratio rather than from genuine interactions between cortical regions. In that case, similar connectivity increases would be expected in both directions. Instead, the consistent dominance of $C_L \rightarrow C_R$ connectivity suggests structured interhemispheric communication. Band-specific analyses showed that the main condition differences occurred in the delta band around movement onset and, to a lesser extent, in the beta band after movement. Overall, connectivity profiles were similar between cue-based and self-paced movements, with differences mainly reflecting movement timing due to the longer and more variable inter-trial intervals in the self-paced condition.

Connectivity and Movement Detection Performance:

A key finding was the relationship between directional connectivity and BCI movement detection performance. Higher $C_L \rightarrow C_R$ connectivity was associated with improved MRCP-based detection accuracy (MCC). Significant clusters appeared in the beta band before and after movement onset, suggesting that directional connectivity reflects neural processes relevant for movement prediction. Beta-band activity is known to support sensorimotor network coordination and motor control, indicating that these interactions provide complementary information about movement-related neural activity. Notably, this effect was observed only for the $C_L \rightarrow C_R$ direction, highlighting the functional relevance of this asymmetric pathway.

Coupling Between Connectivity and Sensorimotor Activity:

To further examine how connectivity relates to local brain activity, we analyzed the temporal coupling between DC trajectories and physiological signals derived from ERD/ERS and MRCP measures. Across conditions, connectivity time courses correlated more strongly with ERD/ERS dynamics than with MRCP signals, particularly for beta-band ERD/ERS, suggesting that oscillatory power modulations are more closely linked to large-scale network interactions than slow cortical potentials. In the cue-based condition, beta-band ERD/ERS in both hemispheres showed positive correlations with alpha- and beta-band connectivity. In contrast, the self-paced condition showed reduced

ERD/ERS–connectivity correlations in the $C_R \rightarrow C_L$ direction and a shift toward broader coupling with alpha- and beta-band ERD/ERS in the $C_L \rightarrow C_R$ direction. These findings suggest that self-paced movements may rely on more flexible interhemispheric coordination than externally cued actions. In cue-based tasks, external timing cues impose a consistent temporal structure on motor preparation, which may promote tighter synchronization between hemispheres. In contrast, self-paced movements are internally generated and temporally more variable, likely engaging more distributed network interactions involving higher-order motor planning regions such as premotor and supplementary motor areas [33]. Such differences are consistent with previous findings showing that internally generated and externally triggered movements involve distinct patterns of cortical interaction [33].

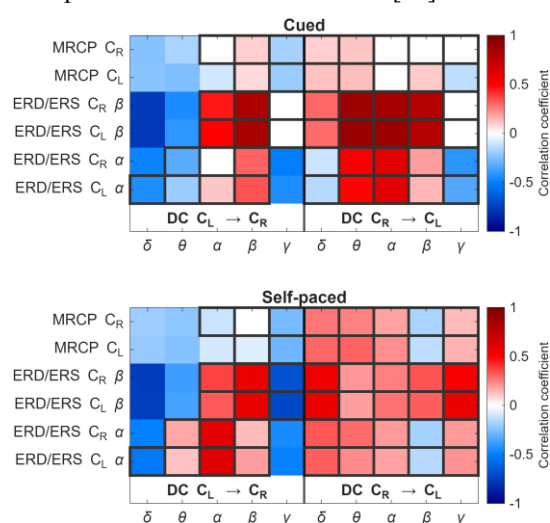


Figure 2. Correlation matrices between band-specific directional connectivity (DC) trajectories and physiological signals during cue-based (top) and self-paced (bottom) movements. DC was computed for canonical bands in both directions ($C_L \rightarrow C_R$, $C_R \rightarrow C_L$). Physiological signals include MRCP time courses from left (C_L) and right (C_R) sensorimotor regions and alpha/beta ERD/ERS trajectories. Each cell shows the Pearson correlation between a physiological signal and a DC trajectory. Red indicates positive and blue negative correlations. Non-significant correlations after multiple-comparison correction are set to zero. Black boxes mark correlations that differed significantly between conditions (Fisher r-to-z test, corrected).

Implications for BCI Systems: These findings highlight the role of interhemispheric connectivity in EEG-based BCI performance. Beta-band connectivity, particularly in the $C_L \rightarrow C_R$ direction, was associated with movement detection despite the classifier relying on MRCP features, suggesting that directional connectivity captures complementary information on motor network coordination. Incorporating connectivity features may improve BCI robustness and detection. Moreover, directional connectivity could serve as a network-level biomarker for predicting and monitoring user-specific BCI performance and explaining inter-individual variability, with relevance for clinical BCI applications where interhemispheric interactions are often altered by

neurological injury.

Limitations and Future Directions: Several limitations should be considered. First, connectivity was computed between aggregated sensorimotor channel groups, capturing large-scale interactions but potentially missing finer spatial patterns. Second, the smaller number of self-paced movements may have increased variability and reduced statistical power despite using equal trial sets. Finally, only unilateral hand movements were analyzed. Future work should integrate directional connectivity into BCI decoding pipelines and further investigate its relationship with MRCPs, ERD/ERS, and motor intention decoding.

CONCLUSION

These results provide a network-level perspective on movement-related brain dynamics and motivate further investigation of interregional interactions in motor control.

ACKNOWLEDGMENTS

This work was supported in part by the INTRECOM project, funded by the European Union's Horizon Europe EIC Pathfinder Challenges programme (Grant Agreement No. 101070939).

REFERENCES

- Shibasaki H, Barrett G, Halliday E, Halliday AM. Components of the movement-related cortical potential and their scalp topography. *Electroencephalogr Clin Neurophysiol*. 1980;49: 213–226.
- Pfurtscheller G, Da Silva FHL. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical neurophysiology*. 1999;110: 1842–1857. doi:10.1016/s1388-2457(99)00141-8
- Crell MR, Kostoglou K, Suwandjeff P, Da Cruz JR, Müller-Putz GR. A non-invasive, MRCP-based BCI for online communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2026.
- Crell MR, Kostoglou K, Sterk K, Müller-Putz GR. A novel paradigm for fast training data generation in asynchronous movement-based BCIs. *Front Hum Neurosci*. 2025. doi:10.3389/fnhum.2025.1540155
- Jiang N, Gizzi L, Mrachacz-Kersting N, Dremstrup K, Farina D. A brain-computer interface for single-trial detection of gait initiation from movement related cortical potentials. *Clinical Neurophysiology*. 2015;126: 154–159.
- Ofner P, Schwarz A, Pereira J, Wyss D, Wildburger R, Müller-Putz GR. Attempted Arm and Hand Movements can be Decoded from Low-Frequency EEG from Persons with Spinal Cord Injury. *Sci Rep*. 2019;9: 1–15. doi:10.1038/s41598-019-43594-9
- Lew E, Chavarriaga R, Silvoni S, Millán J del R. Detection of self-paced reaching movement intention from EEG signals. *Front Neuroeng*. 2012;5: 13.
- Kostoglou K, Müller-Putz GR. Motor-Related EEG Analysis Using a Pole Tracking Approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024. doi:10.1109/TNSRE.2024.3483294
- Scherer R, Schloegl A, Lee F, Bischof H, Janša J, Pfurtscheller G. The self-paced Graz brain-computer interface: Methods and applications. *Comput Intell Neurosci*. 2007;2007: 79826. doi:10.1109/TNSRE.2024.3483294
- He B, Baxter B, Edelman BJ, Cline CC, Ye WW. Noninvasive brain-computer interfaces based on sensorimotor rhythms. *Proceedings of the IEEE*. 2015;103: 907–925.
- Müller-Putz GR, Kaiser V, Solis-Escalante T, Pfurtscheller G. Fast set-up asynchronous brain-switch based on detection of foot motor imagery in 1-channel EEG. *Med Biol Eng Comput*. 2010;48: 229–233.
- Pereira J, Kobler R, Ofner P, Schwarz A, Müller-Putz GR. Online detection of movement during natural and self-initiated reach-and-grasp actions from EEG signals. *J Neural Eng*. 2021;18: 46095.
- Gysels E, Celka P. Phase synchronization for the recognition of mental tasks in a brain-computer interface. *IEEE Transactions on neural systems and rehabilitation engineering*. 2004;12: 406–415.
- Spiegler A, Graimann B, Pfurtscheller G. Phase coupling between different motor areas during tongue-movement imagery. *Neurosci Lett*. 2004;369: 50–54.
- Faes L, Erla S, Nollo G. Measuring connectivity in linear multivariate processes: definitions, interpretation, and practical analysis. *Comput Math Methods Med*. 2012;2012.
- Lütkepohl H. New introduction to multiple time series analysis. Springer Science & Business Media; 2005. doi:10.1007/978-3-540-27752-1
- Pagnotta MF, Plomp G. Time-varying MVAR algorithms for directed connectivity analysis: Critical comparison in simulations and benchmark EEG data. *PLoS One*. 2018;13: e0198846.
- Schlögl A. A comparison of multivariate autoregressive estimators. *Signal Processing*. 2006;86: 2426–2429. doi:10.1016/j.sigpro.2005.11.007
- Kostoglou K, Müller-Putz G. Directed connectivity analysis in people with spinal cord injury during attempted arm and hand movements. *ÖGBMT Jahrestagung 2021*. 2021. pp. 75–78.
- Astolfi L, Cincotti F, Mattia D, Fallani FDV, Tocci A, Colosimo A, et al. Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators. *IEEE Trans Biomed Eng*. 2008;55: 902–913. doi:10.1109/TBME.2007.905419
- Baccalá LA, Sameshima K, Takahashi DY. Generalized partial directed coherence. 2007 15th International Conference on Digital Signal Processing. 2007. pp. 163–166.
- Baccalá L a, Sameshima K. Partial directed coherence: a new concept in neural structure determination. *Biol Cybern*. 2001;84: 463–474. doi:10.1007/PL00007990
- Baccala LA, Sameshima K, Ballester G, Do Valle AC, Timo-laria C. Studying the interaction between brain structures via directed coherence and Granger causality. *Applied Signal Processing*. 1998;5: 40.
- Faes L, Nollo G. Multivariate Frequency Domain Analysis of Causal Interactions in Physiological Time Series. *Biomedical Engineering, Trends in Electronics, Communications and Software*. 2011. doi:10.5772/13065
- Schneider T, Neumaier A. Algorithm 808: ARfit—A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*. 2001;27: 58–65.
- Faes L, Erla S, Porta A, Nollo G. A framework for assessing frequency domain causality in physiological time series with instantaneous effects. *Phil Trans R Soc A*. 2013;371: 20110618.
- Pinheiro JC, Bates DM. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*. 2000; 3–56.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57: 289–300.
- Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21: 1–13. doi:10.1186/s12864-019-6413-7
- Fisher RA. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*. 1921;1: 3–32.
- Zhang X, Zhang S, Zhang H, Wang H, Long J. Postmovement Beta Synchronization Induced by Speed Effects on IHI from the Ipsilateral to Contralateral Motor Cortex. *eNeuro*. 2025;12.
- Leocani L, Toro C, Manganotti P, Zhuang P, Hallett M. Event-related coherence and event-related desynchronization/synchronization in the 10 Hz and 20 Hz EEG during self-paced movements. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*. 1997;104: 199–206.
- Gerloff C, Richard J, Hadley J, Schulman AE, Honda M, Hallett M. Functional coupling and regional activation of human cortical motor areas during simple, internally paced and externally paced finger movements. *Brain*. 1998;121: 1513–1531.
- Beaulé V, Tremblay S, Théoret H. Interhemispheric control of unilateral movement. *Neural Plast*. 2012;2012: 627816.

Tracking Vigilance Fluctuations with Aperiodic Semi-Transparent Visual Flickers

P. Cimarosto¹, F. Dehais¹

¹Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France

E-mail: pietro.cimarosto@isae-supero.fr

ABSTRACT: We investigated whether aperiodic semi-transparent textured rapid visual stimuli (RVS) can provide an unobtrusive neural probe of vigilance fluctuations during a 60-min Mackworth clock vigilance task. Six healthy participants performed the task while EEG was recorded continuously and nearly imperceptible visual stimuli were superimposed on the interface. We analyzed event-related potentials (ERPs) and conventional band-power features at both the group and single-trial levels. Preliminary results showed significant hit-versus-miss differences in the N200 and P300 amplitudes elicited by the stimuli, whereas classical frontal theta and parieto-occipital alpha features did not significantly distinguish the two conditions. Single-trial classification yielded modest performance overall, although ERP-based features appeared more informative than conventional band-power markers, and combining evoked and spectral features slightly improved hit-versus-miss discrimination. These preliminary findings suggest that aperiodic semi-transparent RVS may offer a promising and unobtrusive approach for probing vigilance fluctuations.

INTRODUCTION

Sustained attention and vigilance are critical in many real-world monitoring situations, especially when operators must supervise highly automated systems for prolonged periods with few salient events. In such contexts, attentional lapses or hypovigilance can have serious operational consequences, including in aviation and air-traffic control, where vigilance degradation is linked to performance [1, 2]. Traditional passive BCI approaches for vigilance assessment have therefore mostly relied on spontaneous EEG markers, in particular spectral features such as increased theta/delta activity, alpha modulations, or combinations of frontal theta and parietal alpha, which have long been associated with fatigue, reduced alertness, and vigilance decrement [2–7]. However, while these endogenous frequency-based markers are informative, they often provide indirect estimates of attentional state and can be sensitive to inter-individual variability and contextual factors.

Visual evoked potentials (VEPs) offer a complementary route because they provide an externally time-locked neural probe of moment-to-moment information process-

ing. Several studies have employed the visual oddball paradigm as a secondary task in dual-task settings, reporting reduced ERP amplitudes when the primary task (e.g., driving, walking, or visually demanding activities) became more challenging [8–12]. Yet, conventional visual stimulation paradigms remain difficult to deploy continuously in ecological settings: repetitive flashes and flicker can be visually intrusive, distracting, and fatiguing, especially during prolonged use [13]. This limitation is particularly problematic when the goal is not to solicit an overt response, but rather to unobtrusively probe ongoing cognitive state in parallel with a primary task.

To address this issue, recent work has explored more comfortable forms for rapid visual stimulation (RVS) paradigms, including reduced-luminance and semi-transparent textured flickers that preserve decodable brain responses while limiting visual discomfort. These approaches have shown promise both in steady-state visual evoked potential (SSVEP)-based BCI applications [13] and, more recently, in probing attentional fluctuations [14]. Along this line, StAR (Stimuli for Augmented Response) stimuli are semi-transparent textured visual elements based on a Ricker-wavelet design that can be overlaid onto a task interface with minimal disruption. First introduced in the context of reactive BCIs [15], StAR stimuli have been shown to enable rapid calibration, elicit cortical responses with amplitudes up to twofold greater than those evoked by conventional flickers, and improve user comfort [16–18]. More recently, this family of transparent flickers has also been used in passive settings to tag task-related states such as cognitive workload, while remaining sufficiently unobtrusive to be embedded directly into the visual scene [19]. Such a constant visual probing strategy is attractive because, unlike oddball-style probes, it does not require adding a secondary detection task that could alter the cognitive process under study.

In the present study, we examine whether an aperiodic StAR-based probing approach can track moment-to-moment fluctuations in attention during the Mackworth clock task. We hypothesize that the amplitude of the evoked response elicited by these nearly imperceptible flashes decreases during attentional lapses. To benchmark the sensitivity of this measure, we compare these evoked features with conventional EEG band-power markers commonly used for vigilance assessment.

This comparison is conducted at two complementary levels: first, at the group level, through statistical contrasts between hits and misses; and second, at the single-trial level, within a BCI framework, by evaluating their ability to predict behavioral outcome.

MATERIALS AND METHODS

Participants: We collected experimental data from 6 healthy participants (3 men, 3 women; age: 24.3 ± 4.3 years) with normal or corrected-to-normal vision. All participants provided written informed consent before taking part in the experiment.

Task: The experimental task was based on the Mackworth clock paradigm. Participants observed an interface in which a red dot moved along a circular array of white rings, similarly to the hand of a clock. The dot advanced by one position every 0.8 s. Occasionally, with pseudo-random occurrence, it skipped one position and jumped to the following one. Participants were instructed to respond to these skip events by pressing the spacebar on a standard keyboard. Failure to respond within 3 s after a skip was classified as a miss.

A semi-transparent StAR stimulus based on a Ricker-wavelet textured patch was displayed over the main task interface and flickered throughout the task according to a cyclic aperiodic 2.2 s long code (132 frames on a 60 Hz display). Following the *pseudo-random* method from [18], we use 2-frame bursts (~ 33 ms), resulting in inter-burst intervals ranging from 8 frames (~ 133 ms) to 40 frames (660 ms).

Protocol: Each participant completed a 60-min computer-based vigilance task derived from the Mackworth clock paradigm [14]. Before the experiment, participants completed questionnaires assessing demographic characteristics and current mental state, including the Chalder Fatigue Scale, the Karolinska Sleepiness Scale (KSS), and a Digital Eye Strain questionnaire. At the end of the session, they also completed a questionnaire evaluating the comfort and distractiveness of the StAR visual stimuli. The analysis of these subjective measures falls beyond the scope of the present preliminary report and will be carried out after completion of the full data collection. EEG was recorded continuously throughout the task using a 32-channel semi-dry MBrainTrain system, while participants were simultaneously exposed to aperiodic StAR flickering stimuli superimposed on the task interface in order to probe visual cortical activity.

Group-Level Analysis: We first analyzed behavioral performance using reaction times to skip events and the number of missed responses. At the neurophysiological level, we extracted three classes of EEG features: (1) conventional band-power features, (2) event-related potentials (ERPs), and (3) event-related spectral perturbations (ERSPs).

For band-power analyses, we considered normalized frontal theta activity (4–8 Hz) and normalized parieto-

occipital alpha activity (8–12 Hz). For ERP analyses, we focused on the responses elicited by the StAR stimuli in visual cortical regions and computed the average amplitude in two time windows: an N200 window from 0.175 to 0.225 s and a P300 window from 0.25 to 0.3 s.

These features were analyzed at the group level by comparing their values in the period preceding a successful response (Hit) or a missed response (Miss). To this end, features measured within the 10 s preceding each event were compared using one-sample permutation t-tests on the distribution of subject-level average features. This analysis aimed to evaluate the reliability of the proposed markers over prolonged task duration and their sensitivity to fluctuations in attention.

Single-Trial Level Analyses: We evaluated the practical relevance of these features using single-trial classification to predict *Hit vs Miss*. Each skip event was characterized by four features grouped into two categories: ERP and Band. The ERP category included average N200 and P300 amplitudes; the Band category included normalized frontal theta and parieto-occipital alpha activity computed over the 10 s preceding each skip.

For the classification problem, trials were labeled according to behavioral outcome following the skip event (0: Hit, 1: Miss). Classification was performed using a linear discriminant analysis (LDA) classifier with eigenvalue decomposition and automatic Ledoit–Wolf shrinkage. For each problem, classification was evaluated using each feature category separately and then all features jointly. Performance was estimated using 3-fold cross-validation repeated over 1000 iterations of random undersampling of the majority class. Accuracy values were averaged across folds and iterations. To account for inter-subject variability, the full classification procedure was performed both separately for each participant (subject-specific model) and on the pooled multi-subject dataset (generalized model). In the subject-specific case, reported accuracy corresponds to the average across participants.

RESULTS

Analysis of preliminary data from six participants showed that, behaviorally, the mean number of missed responses per participant was 32.6 ± 19.1 , corresponding to 20.2% of the total number of events. The grand-average response time was 0.595 ± 0.135 s.

Regarding differences in feature states preceding a confirmed response (Hit) versus a missed response (Miss), parieto-occipital alpha and frontal theta showed no significant differences in band activity between conditions. In contrast, significant differences were observed for both the N200 amplitude ($p = 0.03$, Cohen's $d = -0.840$) and the P300 amplitude ($p = 0.03$, Cohen's $d = 1.002$).

Table 1 reports the results of the two classification tasks—Hit vs. Miss and time-on-task (Early vs. Late, corresponding to the beginning and end of the experiment, respectively)—for both subjective and generalized mod-

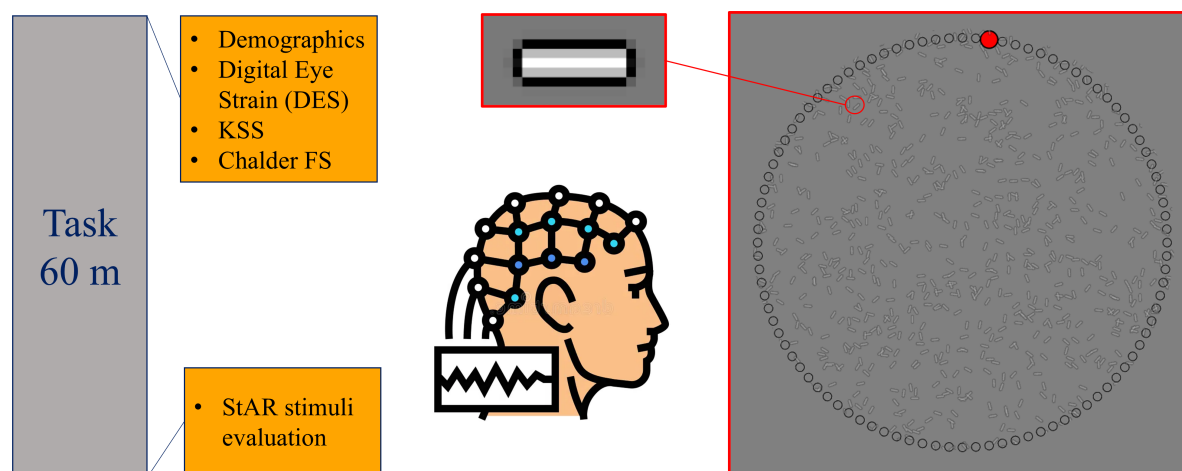


Figure 1: Structure of the experimental procedure (left) and interface (right). Participants were required to answer to questionnaires before and after the task, this included demographic questionnaires, Chalder Fatigue Scale, Karolinska Sleepiness Scale (KSS), Digital Eye Strain (DES) and StAR stimuli evaluation. The task is based on the Mackworth Clock task, in which a red dot moved across a circular path, jumping from one position to the following every 0.8 s. Occasionally, the red dot would skip a position, jumping to the one after (*skip*). Participants were instructed to respond to such event by pressing the spacebar key of a standard keyboard (*Hit*). Failing in doing so within 3 s from the *skip* occurrence would result in a missed response (*Miss*). EEG was collected continuously during the task while a Ricker-wavelet StAR stimuli semitransparent patch was flashed aperiodically over the main task interface following a Burst c-VEP pattern.

Table 1: Results for the hit vs miss using subjective (S) and generalized (G) models, evaluated on two feature categories (ERP and Band), both independently and jointly. For each combination of factors, we report the mean 3-fold cross-validation accuracy across 1000 dataset-shuffling iterations. We also report the dataset size before partitioning for each model and classification task; for the subjective model, this value corresponds to the average sample size across subjects.

Problem	Model	ERP	Band	All	Size
Hit/Miss	S	0.582	0.478	0.554	63.67
Hit/Miss	G	0.603	0.551	0.630	382

els, using ERP, Band, and ERSP features independently and jointly.

DISCUSSION

The objective of this study was to assess whether an aperiodic RVS approach based on semi-transparent textured flickers could be used as an unobtrusive probe of fluctuations in attention during prolonged monitoring. More specifically, we sought to benchmark the sensitivity of flicker-evoked responses against more classical EEG vigilance markers, first at the group level through statistical comparisons between conditions, and then at the single-trial level through classification analyses. In this respect, the Mackworth task provided an appropriate testbed, as it induced a substantial proportion of missed targets (20.2% on average), consistent with the vigilance decrement classically observed in prolonged low-demand monitoring tasks [14, 20].

At the group level, our main finding is that ERP responses elicited by the aperiodic StAR flickers were sen-

sitive to upcoming attentional lapses. Both the N200 and P300 amplitudes, measured in the 10 s preceding a target, were significantly reduced before misses relative to hits. By contrast, conventional oscillatory markers—namely frontal theta and parieto-occipital alpha band activity—did not significantly discriminate between the two conditions within the same analysis window, in contrast with some previous reports [20–22]. These results suggest that evoked responses to textured flickering stimuli may provide a more sensitive index of short-term attentional fluctuations than classical band-power measures in the present paradigm.

This interpretation is in line with the idea that externally evoked responses provide a more direct measure of moment-to-moment processing resources than spontaneous oscillatory activity alone. In this context, the attenuation of the N200 and P300 before misses may reflect a transient reduction in perceptual and attentional engagement preceding the behavioral lapse. Notably, a similar pattern was reported by Ladouce et al. [14], who found that steady-state responses elicited by nearly invisible flickers decreased prior to missed targets in a long-duration vigilance task, while conventional alpha and theta band features were less informative. Taken together, these findings further support the relevance of barely visible rapid visual stimulation paradigms—whether periodic or aperiodic—as visually unobtrusive tools for tracking fluctuations in attention and consolidate previous findings using similar unobtrusive StAR texture stimuli to probe mental workload [19].

The classification analyses provide a more cautious picture. Although ERP features showed statistically signif-

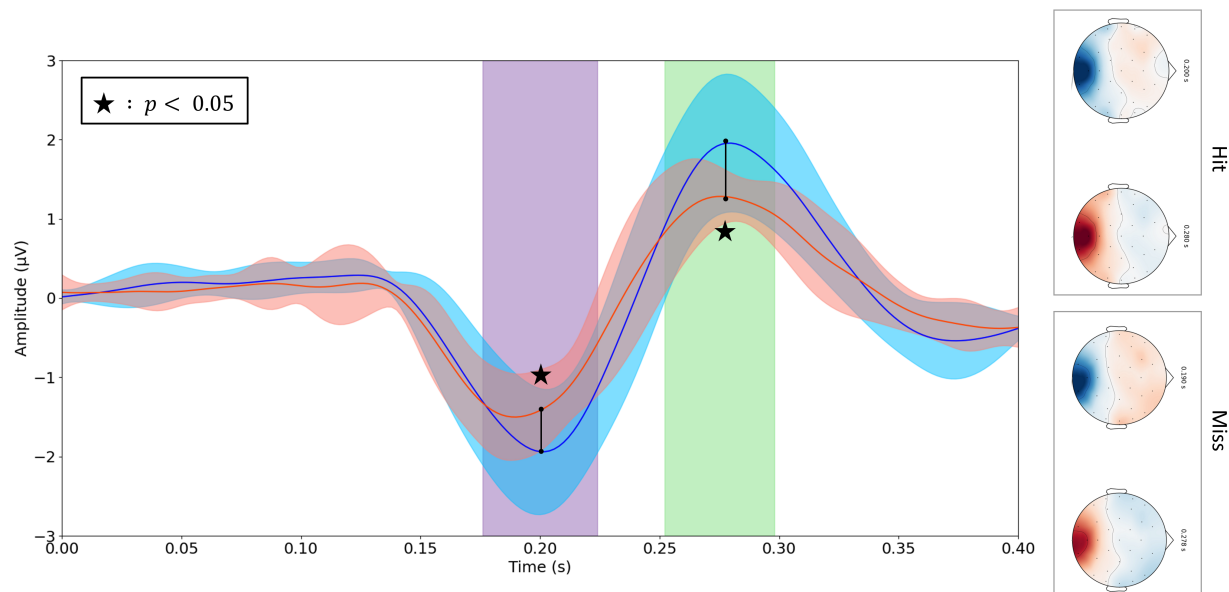


Figure 2: Event-related potentials (ERPs) in the Hit (blue) vs Miss (red) conditions displayed as grand-average wave (line) with subject-wise standard error (shaded region). A statistically significant ($p < 0.05$) difference between the distributions of average amplitude in the N200 (violet) and P300 (green) time windows across subjects between conditions was marked by a black star. On the immediate right, we display topographical plot of the main ERP peaks in both conditions (colder-to-warmer as lower-to-higher).

icant differences between hits and misses at the group level, single-trial decoding performance remained modest overall. This result is not unexpected given the limited size of the present dataset and, more specifically, the relatively small number of miss trials available for training and evaluation. Because misses are by definition less frequent than hits, the *Hit vs Miss* problem suffers from a marked class imbalance, which required repeated undersampling and reduced the effective number of training samples. Under these conditions, moderate classification performance should not be interpreted as a lack of relevance of the proposed features, but rather as a consequence of the difficulty of the problem and the preliminary nature of the dataset.

Nevertheless, some trends are worth noting. For *Hit vs Miss*, ERP features yielded better performance than band-power features in both the subject-specific and generalized models, in line with the statistical results showing greater sensitivity of the evoked markers. Combining ERP and band features further improved performance in the generalized model, suggesting that these feature families may provide partially complementary information. Overall, these findings support the feasibility of the approach, while also highlighting that substantially larger datasets will be required to properly assess its utility for robust single-trial decoding.

Several limitations should be acknowledged. First, the present study is based on a small sample of participants, and the number of miss events per participant remains limited. Second, the classification framework relied on a simple LDA model and a reduced feature set, which was appropriate for a preliminary feasibility study but may not fully exploit the information contained in the

EEG signals. Future work should therefore include additional participants, longer or repeated sessions to increase the number of lapse events, and more advanced feature extraction and decoding strategies. These developments will make it possible to determine more reliably whether the proposed aperiodic StAR paradigm can support real-time vigilance monitoring.

Despite these limitations, the present results provide encouraging evidence that an aperiodic RVS paradigm using semi-transparent textured flickers can probe fluctuations in attention in a sustained monitoring task. More broadly, they reinforce the emerging view that almost invisible flicker-based approaches can offer a useful compromise between signal sensitivity and user comfort. This makes them particularly relevant for ecological passive BCI applications, especially in operational contexts such as aviation, where continuous and unobtrusive monitoring of vigilance could help detect lapses early and support adaptive assistance systems.

CONCLUSION

In this preliminary study, we showed that aperiodic semi-transparent StAR flickers can provide an unobtrusive neural probe of vigilance fluctuations during prolonged monitoring. Flicker-evoked ERP responses, especially the N200 and P300, were sensitive to upcoming attentional lapses at the group level, whereas conventional alpha- and theta-band power features were not. Although single-trial classification performance remained limited, likely because of the small sample size and low number of misses, the present results support the feasibility of the approach. Future work will include collecting more data, integrat-

ing subjective assessments, and extending the analysis to additional neurophysiological markers such as ERSP and inter-trial coherence (ITC). These findings motivate further investigation of aperiodic flicker-based passive BCI strategies for continuous and user-centered vigilance monitoring.

References

- [1] Hamann A et al. The state of the art in assessing mental fatigue in the cockpit using head-worn sensing technology. *Frontiers in Neuroergonomics*. 2026;Volume 6 - 2025.
- [2] Aricò P et al. Adaptive automation triggered by eeg-based mental workload index: A passive brain-computer interface application in realistic air traffic control environment. *Frontiers in Human Neuroscience*. 2016;Volume 10 - 2016.
- [3] Sebastiani M, Di Flumeri G, Aricò P, Sciaraffa N, Babiloni F, Borghini G. Neurophysiological vigilance characterisation and assessment: Laboratory and realistic validations involving professional air traffic controllers. *Brain Sciences*. 2020;10(1):48.
- [4] Dehais F et al. Monitoring pilot's cognitive fatigue with engagement features in simulated and actual flight conditions using an hybrid fnirs-eeg passive bci. In: 2018 IEEE international conference on systems, man, and cybernetics (SMC). 2018, 544–549.
- [5] Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*. 2014;44:58–75.
- [6] Diez PF, Correa AG, Orosco L, Laciari E, Mut V. Attention-level transitory response: A novel hybrid bci approach. *Journal of Neural Engineering*. 2015;12(5):056007.
- [7] Roy RN, Bonnet S, Charbonnier S, Campagne A. Mental fatigue and working memory load estimation: Interaction and implications for eeg-based passive bci. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2013, 6607–6610.
- [8] Yagi Y, Coburn KL, Estes KM, Arruda JE. Effects of aerobic exercise and gender on visual and auditory p300, reaction time, and accuracy. *European Journal of Applied Physiology and Occupational Physiology*. 1999;80(5):402–408.
- [9] Gramann K, Gwin JT, Bigdely-Shamlo N, Ferris DP, Makeig S. Visual evoked responses during standing and walking. *Frontiers in Human Neuroscience*. 2010;4:202.
- [10] Stróżak P, Jakubowska A, Piotrowski K, Krejtz I, Duchowski AT, Krejtz K. Vection reduces the amplitude of visual event-related potentials. *Frontiers in Psychology*. 2016;7:1952.
- [11] Baldwin CL, Coyne JT. Relative validity of the auditory and visual secondary task probes for assessing workload. *Human Factors*. 2005;47(1):157–169.
- [12] Roy RN, Charbonnier S, Campagne A, Bonnet S. Efficient mental workload estimation using task-independent eeg features. *Journal of Neural Engineering*. 2016;13(2):026019.
- [13] Ladouce S, Darmet L, Torre Tresols JJ, Velut S, Ferraro G, Dehais F. Improving user experience of ssvep bci through low amplitude depth and high frequency stimuli design. *Sci. Rep.*. 2022;12(1):8865.
- [14] Ladouce S, Tresols JT, Le Goff K, Dehais F. Eeg-based assessment of long-term vigilance and lapses of attention using a user-centered frequency-tagging approach. *Journal of Neural Engineering*. 2025;22(3):036018.
- [15] Dehais F, Cabrera Castillos K, Ladouce S, Clisson P. Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry eeg code-vep bci. *Journal of Neural Engineering*. 2024;21(6):066023.
- [16] Blanc C, Dehais F, Scannella S. Neuromodulation with transparent textured flicker preserves apha-band entrainment and improves visual comfort: A flanker paradigm. *International Journal of Psychophysiology*. 2025:113307.
- [17] Gomel J, Tresols JT, Cimarosto P, Castillos KC, Dehais F. Pre-decision feedback in code-modulated visual evoked potentials brain-computer interface for an 11-class keypad typing task. In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2025.
- [18] Castillos KC, Ladouce S, Darmet L, Dehais F. Burst c-vep based bci: Optimizing stimulus design for enhanced classification with minimal calibration data and improved user experience. *NeuroImage*. 2023;284:120446.
- [19] Cimarosto P, Velut S, Cabrera Castillos K, Torre-Tresols JJ, Roy RN, Dehais F. Near-invisible c-vep-based passive bci for mental workload monitoring. *Journal of Neural Engineering*. 2026.
- [20] Martel A, Dane S, Blankertz B. EEG predictors of covert vigilant attention. *Journal of Neural Engineering*. 2014.
- [21] Shou G, Dasari D, Ding L. Pre-stimulus alpha and post-stimulus N2 foreshadow imminent errors in a single task. *Neuropsychologia*. 2015;77:346–358.
- [22] O'Connell RG, Dockree PM, Robertson IH, Bellgrove MA, Foxe JJ, Kelly SP. Uncovering the neural signature of lapsing attention: Electrophysiological signals predict errors up to 20 s before they occur. *Journal of Neuroscience*. 2009.

Topological Higher-Order Features for Brain-Computer Interfaces

Wafa Skhiri¹, Fabrizio de Vico Fallani²

Paris Brain Institute - ICM, Sorbonne Universite, CNRS, Inria, Inserm, AP-HP, Pitie Salpetriere
Hospital, Paris, France

E-mail: wafa.skhiri@inria.fr

ABSTRACT: Integrating features from the way different brain signals mutually interact has been recently used to understand and improve brain-computer interfaces (BCIs). However, most existing approaches have focused on pairwise network features, which neglect possibly existing inter-actions involving more than two brain areas. Increasing evidence suggests that higher-order (HO) interactions arising from the simultaneous coordination of multiple groups of regions, bring complementary information for brain states discrimination.

Here, we investigated whether higher-order brain network features extracted from EEG signals can improve the classification of simple motor imagery-based BCI tasks. To this end, we developed a method that combines spectral graph embedding and topological data analysis to extract HO structures from functional connectivity networks estimated using imaginary coherence between EEG signals.

We evaluated the ability of these features to discriminate motor imagery (MI) and resting-state EEG activity in a typical 1D cursor task control. Results showed that MI significantly increases the involvement of sensorimotor electrodes in HO triadic schemes as compared to resting states. These changes were particularly present in *beta* band frequency ranges. Furthermore, we found that in 16 out of 19 participants, such HO features achieved higher accuracy than both local power spectra and connectivity. These findings suggest that topological characterizations of functional brain networks capture behaviorally relevant aspects of neural organization and offer a promising direction for incorporating higher-order interactions into EEG-based BCI applications.

INTRODUCTION

Brain-computer interfaces (BCIs) rely on the extraction of informative features from brain signals in order to discriminate between different mental states. In motor imagery (MI) BCIs in particular, decoding performance strongly depends on the ability to capture neural patterns that reliably distinguish task-related activity from resting-state dynamics. Traditionally, most BCI pipelines rely on local spectral features, such as power spectral density in sensorimotor rhythms, or on pairwise functional connec-

tivity measures that describe interactions between pairs of brain regions[1, 9, 10, 11].

While these approaches have demonstrated their potential, classification accuracy often remain too variable within and between subjects [11, 12, 13]. Furthermore, these state-of-the-art features provide only a partial description of the collective organization of neural activity. Brain processes often involve coordinated activity across multiple neural populations simultaneously, suggesting that relevant information may be contained in interaction patterns that extend beyond pairwise relationships. In recent years, increasing attention has therefore been given to higher-order interactions (HOI), which aim to characterize dependencies involving groups of nodes rather than isolated pairs [2, 14, 15, 16].

Several methodological frameworks have been proposed to investigate HOI in brain networks. Topological approaches model neural systems using hypergraphs or simplicial complexes and apply tools from algebraic topology to describe higher-dimensional structures such as cliques, simplices, and topological cavities [3]. In parallel, statistical approaches grounded in information theory attempt to quantify multivariate dependencies that cannot be reduced to pairwise associations, for instance through measures of redundancy and synergy [4]. Although these approaches provide valuable insights into the collective organization of neural systems, their interpretation in terms of functional brain processes remains challenging, and their direct usefulness for BCI decoding tasks has only been partially explored. [17]

In particular, most BCI feature extraction pipelines remain focused on node-level or pairwise descriptors, such as spectral power or connectivity strength. This raises the question of whether higher-order representations derived from functional connectivity networks can capture additional information relevant for discriminating brain states in BCI settings.

While imaginary coherence provides informative pairwise interactions between EEG channels, it does not directly describe coordinated patterns involving multiple channels simultaneously. A central question of this study is therefore whether higher-order information can be inferred from pairwise connectivity structures and whether such representations provide additional discriminative value for motor imagery decoding. In this work, we investigate whether higher-order brain network features

can improve motor imagery classification in an EEG-based BCI context. The analysis is based on EEG recordings acquired during a motor imagery training protocol in which participants controlled a one-dimensional cursor by imagining right-hand movements versus remaining at rest.

Starting from functional interactions between EEG channels, we derive higher-order descriptors intended to capture coordinated patterns involving multiple channels simultaneously. We then evaluate whether these representations improve the discrimination between motor imagery and resting-state activity compared with traditional features such as power spectral density and node strength.

MATERIALS AND METHODS

Dataset: We used the dataset introduced by Corsi et al. (2020), consisting of recordings from 19 healthy right-handed BCI-naïve participants performing a motor imagery BCI task. Participants were instructed to imagine sustained right-hand grasping movements (MI condition) or remain at rest (REST condition) in order to control a one-dimensional cursor. The experimental protocol comprised four recording sessions per participant. In each session, the BCI testing phase contained six runs of MI/REST trials, yielding 114 available trials per session. The present study focused on one selected session per participant, from which all available MI/REST runs were included for analysis. [5]. which all available MI/REST runs were included for analysis. EEG was recorded with 74 electrodes, from which 72 channels were retained after exclusion of T9 and T10. Signals were acquired at 1000 Hz and subsequently downsampled to 250 Hz. Ocular and cardiac artifacts were reduced using Independent Component Analysis (ICA). Data were segmented into 7-second trials, noisy epochs were rejected through quality control, and signals were average-referenced before further analysis. For the present work, feature extraction was performed on the final 5 seconds of each trial, corresponding to the most stable task-engagement period.

Imaginary coherence networks:

Functional interactions between EEG channels were estimated using imaginary coherence (ImCoh), a frequency-domain connectivity measure based on the imaginary part of coherency between two signals [6]. For two channels $x_i(t)$ and $x_j(t)$, ImCoh at frequency f is defined as:

$$\text{ImCoh}_{ij}(f) = \frac{\text{Im}(S_{ij}(f))}{\sqrt{S_{ii}(f)S_{jj}(f)}}.$$

Here, $S_{ij}(f)$ denotes the cross-spectrum between channels i and j , while $S_{ii}(f)$ and $S_{jj}(f)$ denote their respective auto-spectra. Because the cross-spectrum is a complex-valued quantity, it contains information about the relative phase relationship between the two signals at frequency f . By focusing on its imaginary component, this measure is particularly sensitive to consistent non-zero phase-lag interactions, while attenuating zero-lag couplings that may arise from volume conduction or

common reference effects.

For each trial and frequency, ImCoh matrices were computed between all pairs of EEG channels, resulting in frequency-resolved connectivity networks for both MI and REST conditions.

Spectral embedding: Each connectivity matrix was mapped to a low-dimensional space using Laplacian spectral embedding. More specifically, from the connectivity matrix, a graph Laplacian was constructed, and its leading eigenvectors were used to define the coordinates of the nodes in a reduced-dimensional Euclidean space [7]. In the present study, a two-dimensional embedding was used, as our analysis focused on triangular higher-order motifs (2-simplices).

This transformation converts the connectivity graph into a geometric representation in which each EEG channel is represented as a point in the plane. Channels showing stronger functional connectivity tend to be placed closer to each other, whereas weakly connected channels are positioned farther apart. In this way, the spatial distribution of the points reflects the relational organization of the original network.

The embedding preserves the main structural properties of the connectivity graph while enabling subsequent geometric and topological analysis of the interaction patterns. Higher-dimensional embeddings could naturally be considered in future work to investigate more complex higher-order motifs involving larger groups of nodes.

Topological feature extraction: Topological data analysis was applied to the embedded point clouds. In particular, we constructed α -complexes [8], which are simplicial complexes built from a set of points according to their geometric proximity. Starting from isolated points, nearby points are connected as the scale parameter α increases, progressively forming edges, triangles, and higher-dimensional simplices. More specifically, a simplex is included when its vertices are sufficiently close to share a common ball of radius related to α , which defines the inclusion criterion. This provides a structured representation of the shape and organization of the point cloud while remaining computationally efficient through the associated Delaunay triangulation. A simplicial complex is a geometric structure composed of vertices, edges, triangles, and higher-dimensional elements used to represent relationships among multiple points.

Within this structure, a k -simplex represents a joint interaction among $k + 1$ nodes; for example, a triangle corresponds to a 2-simplex. To characterize these higher-order structures at the node level, we introduced node-wise simplex participation features derived from the α -complexes.

For each node i and frequency f , we define the participation coefficient as

$$T_i(f) = \sum_{\sigma \in \mathcal{S}_k(f)} \mathbf{1}(i \in \sigma),$$

where $\mathcal{S}_k(f)$ denotes the set of k -simplices at frequency f , and $\mathbf{1}(i \in \sigma)$ is an indicator function equal to 1 if

node i belongs to simplex σ , and 0 otherwise. This measure quantifies how strongly each channel participates in higher-order interaction structures.

In this work, we specifically focus on 2-simplices (triangles), which capture simultaneous interactions among triplets, or triads, of EEG channels.

Comparison with node-centered features: To evaluate the contribution of these topological descriptors, we compared them with two standard node-centered representations.

First, power spectral density (PSD) [11, 18] features were computed from selected sensorimotor EEG channels using Welch’s method. For a signal $x_i(t)$ recorded at channel i , the PSD is defined as

$$P_i(f) = |\mathcal{F}\{x_i(t)\}|^2,$$

where $\mathcal{F}$ denotes the Fourier transform.

Second, node strength was extracted from the ImCoh networks [19] as the sum of connectivity weights incident to each node. For node i , the strength is defined as

$$s_i(f) = \sum_j w_{ij}(f),$$

where $w_{ij}(f)$ denotes the imaginary coherence weight between nodes i and j at frequency f .

Classification: All feature representations were evaluated for the classification of MI versus REST trials using a Linear Discriminant Analysis (LDA) classifier. The model was trained and evaluated using a stratified train-test split in order to preserve the class distribution across the two conditions. Classification performance was assessed using accuracy as the primary evaluation metric. To ensure a fair comparison between feature representations, we applied a feature selection step in which the five most discriminative node-level features were retained for each method. This selection was performed independently for each participant using all available sessions, with the objective of identifying subject-specific features that remained stable across sessions rather than optimizing session-specific classification performance. Features were ranked according to the average magnitude of the difference between MI and REST conditions across sessions, together with their variability, allowing us to favor channels whose activity patterns showed both strong discrimination and low variability over time. Notably, the selected nodes were predominantly located over sensorimotor regions, which are known to play a central role in motor imagery tasks.

RESULTS

Figure 2(a) illustrates examples of higher-order network activity derived from the triangle participation features for two representative sensorimotor channels (C1 and FC5) for Subject 01 during Session 04. These electrodes lie over the sensorimotor cortex. The triangle participation spectrum reveals systematic differences between MI and REST conditions across several frequency ranges. In

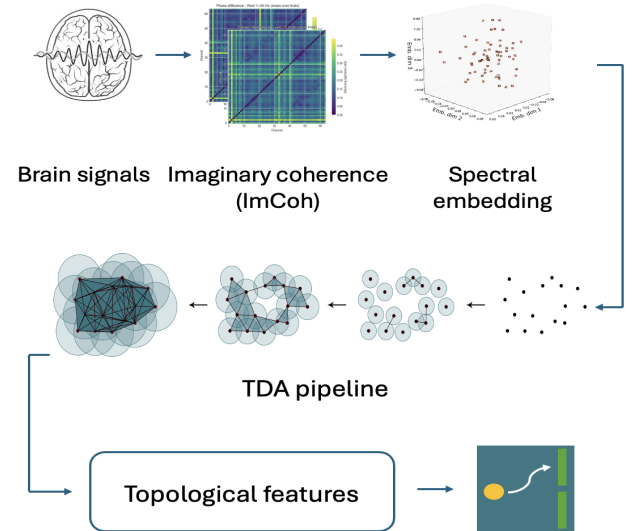


Figure 1: Pipeline for extracting higher-order topological features from EEG signals for motor imagery BCI decoding.

particular, MI trials tend to exhibit higher triangle participation values compared to REST in the beta frequency band, suggesting that motor imagery is associated with stronger higher-order interaction patterns within sensorimotor regions. This observation is coherent with the findings reported in the original dataset study by Corsi et al. (2020) [5], where beta-band connectivity patterns were also shown to play an important role during motor imagery tasks and BCI learning.

To evaluate whether these higher-order features contribute to decoding performance, we compared classification accuracy obtained using three feature representations: power spectral density (PSD), node strength derived from imaginary coherence (ImCoh), and the proposed higher-order interaction (HOI) features. Figure 2(b) presents the per-subject classification accuracy obtained with a Linear Discriminant Analysis (LDA) classifier for all participants in Session 04. Overall, HOI-based features achieve higher or equal classification performance compared to both PSD and node strength for most subjects.

Specifically, HOI features outperform the baseline representations in the majority of participants (16 out of 19 subjects), while also exhibiting relatively stable performance across individuals. In contrast, PSD and node strength features show larger inter-subject variability and lower accuracy for several participants. These results indicate that higher-order topological descriptors capture discriminative information that is not fully represented by local spectral activity or pairwise connectivity measures, and that they provide complementary features for motor imagery decoding in EEG-based BCI systems.

DISCUSSION

These preliminary results provide initial evidence that

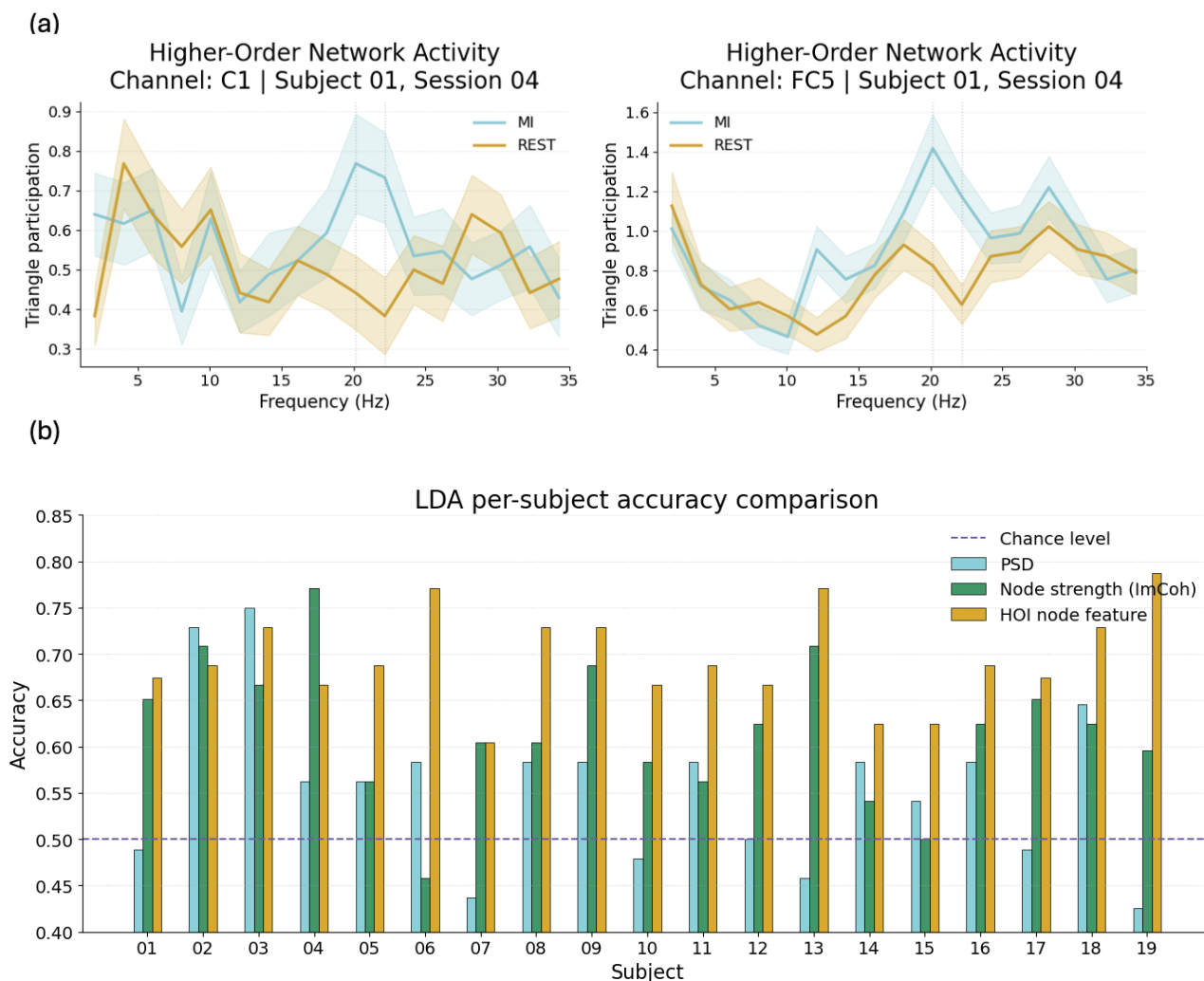


Figure 2: Higher-order network activity and classification performance. (a) Mean triangle participation as a function of frequency for MI and REST conditions in two sensorimotor channels (C1 and FC5). (b) Per-subject classification accuracy obtained with Linear Discriminant Analysis (LDA) using PSD, node strength (ImCoh), and higher-order interaction (HOI) features.

higher-order features derived from phase-based representations, specifically imaginary coherence (ImCoh) in this study, can discriminate motor imagery (MI) from resting-state (REST) activity. The simplex profiles reveal systematic differences between the two conditions, with consistently higher levels of higher-order activity during MI than during REST across several frequencies. At first glance, this pattern may appear counterintuitive, as resting-state activity is often associated with large-scale distributed coordination across brain networks, whereas MI is typically considered a more localized task-driven process centered on sensorimotor regions.

The observed increase in simplex counts during MI likely reflects the specific type of higher-order interactions captured by the proposed framework. Rather than describing slowly varying functional integration across distant regions, these features emphasize simultaneous phase coordination across multiple channels. In contrast to classical functional connectivity measures, which may capture delayed or indirect interactions, the present approach high-

lights time-aligned coordination patterns among groups of electrodes. Under this interpretation, motor imagery may induce stronger multi-channel coordination within sensorimotor networks than unconstrained resting activity.

This interpretation is supported by the classification results. When HOI features are used, classification accuracy generally improves compared to traditional node-centered metrics such as power spectral density and node strength. Although the performance gain remains moderate, HOI features outperform the baseline representations for most participants and show relatively stable performance across subjects. This suggests that higher-order structures capture complementary information that is not fully represented by local spectral activity or pairwise connectivity measures, and that they may provide useful features for motor imagery decoding in EEG-based BCI systems.

Further work is required to validate these findings and refine their interpretation. Future studies should evalu-

ate the proposed approach across larger datasets, additional recording sessions, and other cognitive or motor tasks. Moreover, comparisons with alternative connectivity and information-theoretic measures may help clarify the neural mechanisms underlying the extracted higher-order structures. Such investigations will be important to assess the generality and practical relevance of higher-order representations for EEG-based brain-computer interface applications. From an applied perspective, the moderate computational cost of the proposed pipeline and the stability of the extracted subject-specific features suggest that this framework could also be relevant for future online BCI implementations. In particular, after an initial calibration phase, such higher-order descriptors could potentially be integrated into real-time decoding pipelines for adaptive motor imagery control.

CONCLUSION

In this work, we investigated the use of higher-order interaction (HOI) features derived from phase-based functional connectivity for motor imagery decoding in an EEG-based BCI context. By combining imaginary coherence (ImCoh) networks, spectral graph embedding, and topological data analysis, we extracted simplex-based descriptors that capture coordinated activity patterns across multiple EEG channels.

Our results demonstrate that these higher-order features reveal clear differences between motor imagery and resting-state conditions and improve classification performance compared to traditional representations such as power spectral density and node strength. In particular, HOI features achieve higher accuracy for the majority of subjects while maintaining stable performance across participants. These findings indicate that higher-order structures capture discriminative information that is not fully represented by local spectral activity or pairwise connectivity measures.

Overall, this study shows that topological higher-order descriptors provide an effective representation of functional brain networks and constitute a valuable feature space for motor imagery decoding in EEG-based brain-computer interfaces.

ACKNOWLEDGEMENT

The authors acknowledge Marie-Constance Corsi for the preparation of the EEG dataset. This study was supported by the Agence Nationale de la Recherche under the France 2030 program (Grant Agreement ANR-23-IACL-0008). F.d.V.F. was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement N° 864729) and by the Agence Nationale de la Recherche under the France 2030 program (Grant Agreement ANR-23-IACL-0008). The authors declare that they have no competing interests.

REFERENCES

- [1] J. Gonzalez-Astudillo, T. Cattai, G. Bassignana, M.-C. Corsi, and F. De Vico Fallani, "Network-based brain-computer interfaces: principles and applications," *Journal of Neural Engineering*, vol. 21, no. 1, 2024.
- [2] A. Santoro, F. Battiston, M. Lucas, G. Petri, and E. Amico, "Higher-order connectomics of human brain function," *Nat. Commun.*, vol. 15, Art. no. 10244, 2024.
- [3] Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P.J., Vaccarino, F., 2014. Homological scaffolds of brain functional networks. *J. R. Soc. Interface* 11, 20140873.
- [4] Rosas, F.E., Mediano, P.A., Gastpar, M., Jensen, H.J., 2019. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys. Rev. E* 100, 032305.
- [5] M. C. Corsi, M. Chavez, D. Schwartz, N. George, L. Hugueville, A. Kahn, S. Dupont, D. Bassett, and F. De Vico Fallani, "Functional disconnection of associative cortical areas predicts performance during BCI training."
- [6] K. Sekihara, J. P. Owen, S. Trisno, and S. S. Nagarajan, "Removal of spurious coherence in MEG source-space coherence analysis," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 11, pp. 3121–3129, 2011.
- [7] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [8] Edelsbrunner, H., Harer, J., *Computational Topology: An Introduction*, American Mathematical Society, 2010.
- [9] Pfurtscheller, G., Neuper, C., Motor imagery and direct brain-computer communication, *Proceedings of the IEEE*, 2001.
- [10] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Muller, K.-R., Optimizing spatial filters for robust EEG single-trial analysis, *IEEE Signal Processing Magazine*, 2008.
- [11] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., Yger, F., A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update, *Journal of Neural Engineering*, 2018.
- [12] Vidaurre, C., Blankertz, B., Towards a cure for BCI illiteracy, *Brain Topography*, 2010.
- [13] Ahn, M., Cho, H., Ahn, S., Jun, S. C., High theta and low alpha powers may be indicative of BCI-illiteracy in motor imagery, *PLoS ONE*, 2013.
- [14] Battiston, F., Petri, G., Andaloro, C., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., Musciotto, F., Rosvall, M., Vaccarino, F., Higher-order interactions in complex systems, *Physics Reports*, 2021.
- [15] Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P. J., Vaccarino, F., Homological scaffolds of brain functional networks, *Journal of The Royal Society Interface*, 2014.
- [16] Giusti, C., Ghrist, R., Bassett, D. S., Two's company,

three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data, *Journal of Computational Neuroscience*, 2016.

[17] Peixoto, T. P., Peel, L., Gross, T., De Domenico, M., Graphs are maximally expressive for higher-order interactions, *Physical Review Letters*, 2025.

[18] Welch, P. D., The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Transactions on Audio and Electroacoustics*, 1967.

[19] Cattai, T., Colonnese, S., Corsi, M.-C., Bassett, D. S., Scarano, G., De Vico Fallani, F., Phase/amplitude synchronization of brain signals during motor imagery BCI tasks, 2019.

ERP-XTTN: Interpretable Cross-Subject Error-Related Potential Classification via Cross-Attention to Data-Driven ERP Prototypes

Charlotte Genevieve Wyman¹, Leanne Hirshfield¹

¹University of Colorado Boulder, Boulder, CO, United States

E-mail: charlotte.wyman@colorado.edu

ABSTRACT: Error-related potential (ErrP) classification methods for brain-computer interface deployment have separately addressed causal filtering, calibration-free generalization, and interpretability, but no existing approach combines all three. We propose ERP-XTTN, a cross-attention architecture using fixed difference-wave prototypes with temporal windows derived from training data. A peak-detection algorithm identifies ErrP components from the grand-average difference wave and sets window boundaries at smoothed zero-crossings. The prototypes serve as keys in a cross-attention mechanism with no value projection, so classification depends entirely on interpretable attention routing. We evaluate via leave-one-subject-out (LOSO) cross-validation on two datasets (BNCI Horizon 2020, 6 subjects; HRI Cursor, 11 subjects) with causal filtering and no hyperparameter tuning. ERP-XTTN achieves 0.782/0.836 AUROC (BNCI/HRI) using three midline channels and remains competitive with two. Attention maps reveal class-discriminative routing aligned with event-related potential (ERP) component timing on both datasets. These results show that data-driven prototypes can approach black-box performance while preserving interpretability grounded in ERP component timing, without sacrificing causal filtering or calibration-free generalization.

INTRODUCTION

Error-related potentials (ErrPs) are neural signals measurable via electroencephalography (EEG) that arise when a person observes or commits an erroneous action [1]. These potentials can provide brain-computer interfaces (BCIs) with implicit error-detection capability, enabling system correction without explicit user input.

Deployment-ready ErrP classification requires meeting several constraints at once. For online compatibility, filtering must be causal (forward-only), which can reduce signal quality relative to acausal filtering. We emphasize causal filtering because it is a necessary prerequisite that most offline evaluations omit: acausal filtering incorporates future samples, producing inflated accuracy estimates that cannot be reproduced in any forward-processing pipeline. Practical deployment also limits channel count (many commercial devices lack research-grade coverage [2]) and cannot depend on subject-specific calibration, which delays setup and limits scal-

ability. Finally, clinical and safety-critical applications demand that classification decisions be interpretable, not merely accurate.

Prior works have addressed each of these issues separately. Cross-subject causal ErrP classification has been demonstrated online using PCA-based dimensionality reduction and shrinkage linear discriminant analysis (LDA) with personalized decision thresholds under a full 61-channel montage [3]. This pipeline requires dedicated electrooculographic (EOG) channels with a calibration recording for artifact removal, and the PCA feature space is defined over the full electrode set, preventing direct application to reduced montages. Calibration-free cross-subject ErrP decoders have also been demonstrated using ensemble classifiers in a simulated online setting [4] and using generalized LDA models on one of the datasets used in this work [5], though without inherent interpretability or minimal-channel constraints. At reduced channel counts, Ancau et al. [6] demonstrated online single-channel ErrP decoding from a consumer-grade headset using deep learning, though without cross-subject evaluation or interpretability. Wimmer et al. [7] showed that combining minimal EEG montages with pupillometric signals can partially recover accuracy lost from channel reduction, though under acausal preprocessing and without interpretability. LOSO evaluation on BNCI has been demonstrated using convolutional neural network (CNN)-transformer architectures [8, 9], reaching 0.755 area under the receiver operating characteristic curve (AUROC) with two channels (FCz, Cz), but without inherent interpretability. Gradient-weighted Class Activation Mapping (Grad-CAM) has been applied to a neuroscience-informed CNN architecture for interpretable cross-subject error-related negativity (ERN) classification [10], demonstrating alignment between learned features and known ERP component timing, though under offline preprocessing and with post-hoc rather than inherent interpretability. Chan et al. [11] used post-hoc explainability methods to systematically reduce the EEG setup for asynchronous ErrP decoding under LOSO evaluation, retaining performance with as few as 10 channels, though the interpretability informs setup design rather than individual classification decisions. Prototype-based ERP classification has been explored for P300 in rapid serial visual presentation (RSVP) tasks [12], using learned prototypes matched via

Euclidean distance in a CNN feature space, but without attention-based routing or application to ErrPs. To the best of our knowledge, no prior work has met all these deployment constraints at once: causal filtering, cross-subject generalization without calibration, minimal channels, and inherent interpretability.

We propose a cross-attention architecture designed to satisfy all these constraints simultaneously. Fixed difference-wave prototypes are extracted automatically from training data via a constrained peak-detection procedure applied to the grand-average difference wave of the training subjects. Classification is based directly on attention weights, ensuring that all discriminative information flows through interpretable prototype-to-signal routing. We evaluate the model under LOSO cross-validation with causal infinite impulse response (IIR) filtering on two public benchmarks and show competitive performance on 2- and 3-channel montages.

To our knowledge, this is the first inherently interpretable ErrP classifier evaluated under LOSO with causal filtering on standard benchmarks. We show that this interpretability comes at an average cost of only 0.026 AU-ROC relative to EEGNet [13] in the 3-channel configuration, and that attention routing patterns suggest subject-specific differences in component-level routing across two datasets.

METHODS

Datasets: Two publicly available datasets (BNCI Horizon 2020 and HRI Cursor) were used to evaluate the proposed model.

BNCI Horizon 2020 is a 64-channel EEG dataset collected from 6 subjects in an observation-error paradigm [14]. Subjects were asked to monitor the performance of an agent moving a cursor toward a target; errors occurred on roughly 20% of trials. The dataset is publicly available at <https://bnci-horizon-2020.eu/database/data-sets>.

HRI Cursor is a 27-channel EEG dataset collected from 11 subjects during a cursor-control choice-reaction-time task [15]. Participants responded to one of three target stimuli via keypress, and feedback was presented as a cursor moving toward (congruent) or away from (incongruent) the target; errors occurred on roughly 35% of trials. Only the cursor scenario was used in this work; a companion robot-head-turn scenario from the same study was excluded. The dataset is publicly available at <https://github.com/stefan-ehrich/dataset-ErrP-HRI/>.

Preprocessing: A causal IIR 4th-order Butterworth 1-10Hz bandpass filter was applied to both datasets before epoching to 0-800ms after event onset. BNCI was downsampled from 512Hz to 256Hz after filtering. HRI was sampled at 256Hz and required no downsampling. Baseline correction, re-referencing, and spatial filtering were omitted to evaluate performance under constraints compatible with online, zero-calibration deployment.

Three channel sets were used: a 2-channel midline set (Fz, Cz), a 3-channel midline set (Fz, Cz, Pz), and the

full montage (64 for BNCI and 27 for HRI). The 2- and 3-channel configurations were selected based on the known frontocentral topography of ErrP components [1]. While FCz is often reported as the highest-amplitude ErrP channel [16], it is absent from many commercial and research-grade montages, including HRI. Selecting a channel set available across both datasets tests robustness to realistic montage variability and supports deployment across heterogeneous recording setups.

Architecture (ERP-XTTN): The proposed model, ERP-XTTN, classifies trial epochs by routing input EEG patches to a set of fixed ERP prototypes via query-key (QK)-only cross-attention, then classifying directly from the resulting attention-weight distribution (Figure 1).

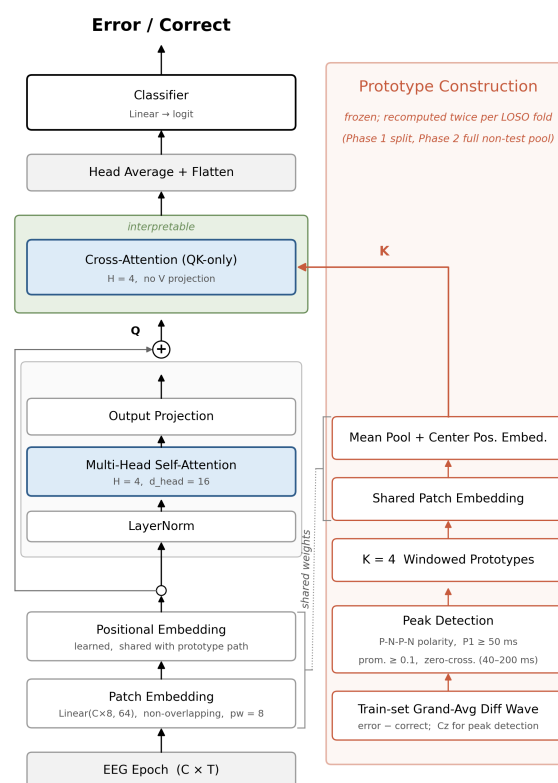


Figure 1: ERP-XTTN architecture. Input patches serve as queries (Q) into a QK-only cross-attention module against $K=4$ frozen ERP prototypes derived from the training set's grand-average difference wave via P-N-P-N peak detection. The attention-weight distribution over prototypes (the interpretable representation) is directly classified. Patch and positional embeddings (dotted line) are shared between the input and prototype paths. Prototypes are recomputed per LOSO fold.

Prototypes are derived from the grand-average training-set difference wave (mean error - mean correct) and, within each LOSO fold, are recomputed twice: first on the phase-1 training split used for validation-based epoch selection, and then on the full non-test training pool for phase-2 retraining at the selected epoch count. On the Cz channel of the smoothed difference wave (Gaussian kernel, $\sigma = 2$ samples), peaks are detected subject to a P-N-P-N polarity constraint reflecting the canonical ErrP morphology: an initial positivity (PI), the error negativ-

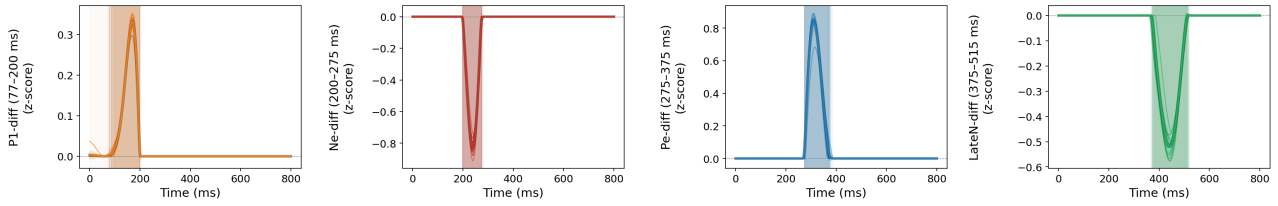


Figure 2: Difference-wave prototypes (Cz shown; HRI dataset; 3-channel configuration) derived from the grand-average error-minus-correct waveform. Each column shows one prototype: *P1* (77–200ms), *Ne* (200–275ms), *Pe* (275–375ms), and *LateN* (375–515ms). Shaded regions indicate prototype windows. Thin lines show individual LOSO fold prototypes ($n = 11$); thick lines show their mean.

ity (*Ne*), the error positivity (*Pe*), and a late negativity (*LateN*) [16]. All local extrema exceeding a prominence threshold of 0.1 are identified; *P1* is anchored as the earliest positive peak at least 50ms post-stimulus, and the remaining three components are selected as the N-P-N chain maximizing total prominence. Each peak is expanded to a window bounded by zero-crossings of the smoothed signal and clamped to 40–200ms width. The multichannel difference wave within each window forms one prototype; values outside the window are zeroed, yielding a prototype tensor of shape ($K=4$, C , T). Prototypes are frozen during training. An example set of resulting prototypes for the HRI 3-channel configuration is shown in Figure 2.

Input epochs (shape $C \text{ channels} \times T = 206$ samples) are divided into non-overlapping patches of width 8 samples, linearly projected to $d_{\text{model}} = 64$, and summed with a learned positional encoding ($N = 25$ patches). Dropout of 0.3 is applied after embedding. Prototypes pass through the same shared patch embedding, are mean-pooled across patch positions, and receive a positional encoding indexed at each prototype’s temporal center. Input patches are first passed through a single transformer-style self-attention layer. This layer applies pre-norm LayerNorm, multi-head self-attention ($H = 4$ heads, $d_{\text{head}} = 16$), output projection, and a residual connection, with dropout of 0.3 on the attention weights. Cross-attention is then computed between input patch queries and prototype keys using separate LayerNorms, linear projections to d_{model} , and multi-head scaled dot-product attention ($H = 4$), with logits clamped to $[-10, 10]$. Crucially, no value projection is used; instead, the softmax attention weights over the $K = 4$ prototypes are themselves the learned representation. This design constrains the model’s discriminative representation to prototype routing patterns, providing inherent interpretability without post-hoc analysis.

Finally, attention weights are averaged across heads, yielding a (B, N, K) tensor, flattened to $(B, 100)$, and passed through a single linear layer to produce a scalar logit. The full model has approximately 28,645 trainable parameters in the 3-channel configuration. Single-trial inference time is 0.94 ms (median) on CPU (AMD Ryzen 9 8945HS).

The model was trained with AdamW ($lr = 1 \times 10^{-3}$, weight decay $= 1 \times 10^{-4}$) using linear warmup for 5 epochs followed by cosine annealing over 100 epochs;

beyond 100 epochs, the learning rate was held at 10^{-5} . Gradient norms were clipped at 1.0. The loss was binary cross-entropy with a positive class weight equal to the inverse class frequency ratio ($n_{\text{correct}}/n_{\text{error}}$). Batch size was 32. Data augmentation consisted of temporal jitter (uniform ± 10 samples) and additive Gaussian noise ($x \leftarrow x + 0.1 \cdot N(0, 1)$ in normalized units), applied during training only.

Baseline (EEGNet): As a deep learning baseline for comparison, we used EEGNet with $F1=8$, $D=2$, and a kernel length of 128 samples, yielding 2,017 trainable parameters for the 3-channel configuration. We used the same preprocessing, training pipeline, hyperparameters, and evaluation protocol as for ERP-XTTN, including matched augmentation (temporal jitter and amplitude noise).

Evaluation: We evaluated both models using leave-one-subject-out (LOSO) cross-validation. Within each dataset, each subject served as the test set exactly once, with all remaining subjects forming the training set. To simulate zero-calibration deployment, no subject data from the held-out test subject was used during training. Within each LOSO fold, early stopping on a stratified validation split determined the optimal epoch count (patience 15, max 250 epochs, AUROC criterion). The final model was then retrained from scratch on all training and validation data for that epoch count.

The primary evaluation metric was AUROC, which is threshold-independent and robust to class imbalance. Balanced accuracy was computed at a fixed decision threshold of 0.5 on the sigmoid output, applied uniformly across all subjects and configurations. All reported results were obtained without hyperparameter optimization. Identical architectures, hyperparameters, and training procedures were used across both datasets and all channel configurations.

RESULTS

Table 1 summarizes AUROC for both architectures across all channel configurations, while Table 2 reports balanced accuracy (BA). At the 3-channel midline configuration, ERP-XTTN achieves 0.782 ± 0.058 AUROC on BNCI and 0.836 ± 0.107 on HRI, compared to 0.806 ± 0.077 and 0.864 ± 0.091 for EEGNet. At the 2-channel midline configuration, ERP-XTTN scores 0.775 ± 0.076 and 0.817 ± 0.115 , versus 0.777 ± 0.082 and

Table 1: AUROC (mean $\pm$ std, LOSO)

Dataset	Channels	ERP-XTTN	EEGNet
BNCI	2	.775 $\pm$.076	.777 $\pm$.082
BNCI	3	.782 $\pm$.058	.806 $\pm$.077
BNCI	64	.805 $\pm$.060	.842 $\pm$.071
HRI	2	.817 $\pm$.115	.839 $\pm$.105
HRI	3	.836 $\pm$.107	.864 $\pm$.091
HRI	27	.834 $\pm$.100	.884 $\pm$.073

Table 2: Balanced accuracy (mean $\pm$ std, LOSO)

Dataset	Channels	ERP-XTTN	EEGNet
BNCI	2	.698 $\pm$.062	.703 $\pm$.065
BNCI	3	.687 $\pm$.058	.722 $\pm$.064
BNCI	64	.719 $\pm$.056	.756 $\pm$.084
HRI	2	.739 $\pm$.102	.762 $\pm$.104
HRI	3	.759 $\pm$.100	.779 $\pm$.091
HRI	27	.747 $\pm$.089	.784 $\pm$.082

0.839 $\pm$ 0.105 for EEGNet. At full montage (64-channel BNCI, 27-channel HRI), ERP-XTTN reaches 0.805 $\pm$ 0.060 and 0.834 $\pm$ 0.100, compared to 0.842 $\pm$ 0.071 and 0.884 $\pm$ 0.073. BA patterns are consistent with AUROC: EEGNet leads across all configurations, and the gap between models follows the same channel-scaling trend.

The 3-channel midline configuration retained nearly all of the full-montage AUROC for ERP-XTTN. On HRI, 3-channel AUROC slightly exceeds full-channel performance (0.836 vs 0.834).

For the 3-channel midline configuration, ERP-XTTN outperformed EEGNet on 2 of 6 BNCI subjects, including subject 6 (0.702 vs 0.672 AUROC). On HRI, subject 4 showed near-ceiling performance for both models (0.980 and 0.986), while subject 10 had the lowest AUROC for both (0.546 and 0.636). Cross-subject AUROC standard deviation on BNCI was lower for ERP-XTTN than EEGNet (0.058 vs 0.077). Per-subject AUROC rank-order agreement between models was high (Spearman ρ = 0.94 on BNCI and ρ = 0.93 on HRI).

At the 3-channel midline configuration, attention routing differed systematically between error and correct trials. The most consistent pattern was reduced attention to the *Ne* prototype on error trials: mean *Ne* attention dropped from 0.466 to 0.340 on BNCI and from 0.503 to 0.337 on HRI, with compensatory increases distributed across the remaining prototypes. The strongest temporal contrast on BNCI occurred at $\sim$ 359ms, where *Ne* routing decreased (-0.190) and *Pe* routing increased ($+0.221$). On HRI, positive routing effects were distributed primarily between *PI* ($+0.234$ at 266ms) and *LateN* ($+0.234$ at 422ms), with a smaller *Pe* contribution.

Per-subject routing was heterogeneous but structured. On BNCI, *Pe* carried the dominant positive routing effect in 5/6 subjects, with subject 6 dominated by *PI* instead. On HRI, subjects split between *PI*-dominant (5/11) and *LateN*-dominant (5/11) routing, with one *Pe*-dominant subject. *Ne* routing remained negative (stronger for correct than error) in 5/6 BNCI and 10/11 HRI subjects. Routing contrast magnitude tracked with perfor-

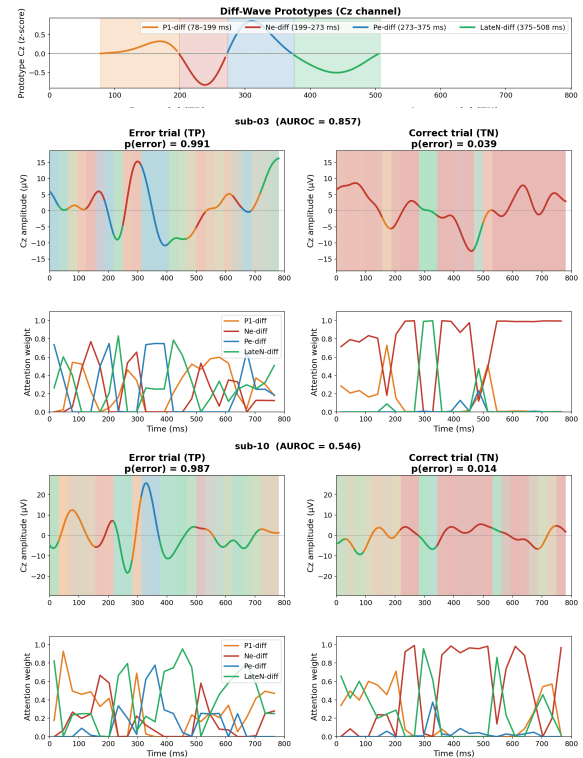


Figure 3: Difference-wave prototypes (top) and attention routing for high-confidence, true positive (error) and true negative (correct) trials from two HRI subjects in the 3-channel midline configuration. Both subjects show class-dependent routing, with *Ne*-dominant attention on correct trials and increased *Pe* and *LateN* engagement on error trials. Per-subject differences are visible in which prototypes are attended: for example, subject 3 shows *Ne*-dominant routing in the 250–300ms range on error trials, while subject 10 routes more strongly through *LateN* and *PI* in the same window. These routing profiles are read directly from the model’s attention weights without post-hoc analysis.

mance: the lowest-performing subjects showed the weakest error-vs-correct separation (BNCI subject 6: AUROC 0.702, max $|\Delta|$ =0.208; HRI subject 10: AUROC 0.546, max $|\Delta|$ =0.129), while the highest-performing subject showed the largest (HRI subject 4: AUROC 0.980, max $|\Delta|$ =0.722). The same patterns appeared at 2 channels and full montage, though *Ne*-dominant contrast was less stable. Figure 3 shows representative high-confidence error and correct trials from two HRI subjects at the 3-channel midline configuration, illustrating the class-dependent and subject-specific routing patterns described above.

DISCUSSION

The proposed ERP-XTTN classifier provides an inherently interpretable approach to deployment-compatible, cross-subject ErrP classification. With sub-millisecond CPU inference, it is compatible with deployment on modest hardware. Classification is based entirely on cross-attention routing to data-driven prototypes derived from canonical ErrP morphology, making the decision process

directly inspectable.

Figure 3 shows trial-level routing for two HRI subjects spanning a wide AUROC range. The *Ne*-dominant routing on correct trials, present at the group level on both datasets, may reflect the absence of the distinctive ErrP deflection sequence: without the *Pe* and *LateN* components that characterize error processing, the signal in the *Ne* window becomes the most informative template for identifying correct trials. This is consistent with the model exploiting known ERP component timing, though the alignment is not guaranteed by the architecture; the model could in principle learn to ignore the prototypes or route arbitrarily. Regardless of the underlying neurophysiological explanation, the routing patterns provide a concrete, trial-level account of each classification decision. This degree of transparency is difficult to obtain with black-box alternatives.

As Table 1 shows, the cost of this interpretability is modest: the mean gap between ERP-XTTN and EEGNet is 0.026 AUROC at the 3-channel midline configuration, similar across both datasets (0.024 BNCI, 0.028 HRI). The gap widens with channel count, from 0.002–0.022 at 2 channels to 0.024–0.027 at 3 channels and 0.037–0.050 at full montage. This pattern is expected: EEGNet’s depthwise-separable convolution learns spatial filters that benefit from additional channels, while ERP-XTTN’s fixed prototypes constrain spatial learning by design. The interpretability cost is therefore smallest precisely where deployment constraints are tightest. At 3 midline channels, ERP-XTTN largely matches its own full-channel AUROC, suggesting that additional channels may introduce noise without improving discrimination for this architecture.

BNCI subject 6 is the sole case where ERP-XTTN outperforms EEGNet (0.702 vs 0.672). This exception is notable given the otherwise high rank-order agreement between models on BNCI (Spearman $\rho = 0.94$). Subject 6 has the fewest error trials and the highest noise floor in the dataset, and Ren et al. [9] independently noted that this subject’s ERP morphology differed substantially from that of the remaining five participants. For such atypical subjects, fixed prototypes may provide a stronger inductive prior: when error trials are scarce, they regularize against overfitting, and when individual ERP morphology deviates from the population, they anchor classification to canonical component timing rather than to idiosyncratic patterns. ERP-XTTN also shows lower cross-subject variance on BNCI (0.058 vs 0.077), though $N=6$ limits the strength of this observation. If this pattern holds in larger samples, ERP-XTTN may offer an advantage precisely where it matters most: atypical or low-data subjects who are hardest to serve in zero-calibration settings.

The correlation between routing contrast magnitude and per-subject AUROC suggests that attention weights carry information beyond classification. In an online BCI context, routing profiles could help flag subjects whose ERP morphology deviates from the training population, in-

forming whether to trust the classifier output or initiate recalibration. This dual-use potential (classification and diagnostics from the same forward pass) is unavailable from black-box classifiers, even when augmented with post-hoc analysis. Jalilpour and Müller-Putz [10] achieve interpretability only through post-hoc Grad-CAM applied to a learned feature space, and Wei et al. [12] learn prototypes for P300 classification in a learned feature space using Euclidean distance with no attention transparency over the prototypes. In ERP-XTTN, the attention weights are themselves the explanation.

On the classification pipeline side, several prior works have addressed subsets of the constraints targeted here. Lopes-Dias et al. [3] demonstrated online cross-subject causal ErrP classification with causal filtering, PCA-based dimensionality reduction, and shrinkage LDA over a full 61-channel montage, with subspace subtraction requiring dedicated EOG channels and a pre-experiment calibration recording. Personalized decision thresholds were additionally updated during the first three experimental blocks. These montage and calibration requirements limit direct transfer to reduced or heterogeneous channel configurations. Ren et al. [8] evaluated a CNN-transformer architecture under LOSO on BNCI but used common average reference (CAR) re-referencing and non-causal finite impulse response (FIR) filtering across all 64 channels and included no interpretability mechanism. A follow-up study [9] reduced the input to 2 channels (FCz, Cz) with Butterworth filtering, achieving 0.755 and 0.737 AUROC on BNCI sessions 1 and 2 respectively under LOSO. ERP-XTTN achieves 0.775 AUROC at 2 channels (Fz, Cz), exceeding both per-session results despite stricter, deployment-oriented preprocessing constraints and the substitution of Fz for the typically higher-amplitude FCz. On HRI, Schönleitner et al. [5] reported 72.7% balanced accuracy with a generalized LDA using all 27 channels; ERP-XTTN achieves 75.9% balanced accuracy with three midline channels and inherent interpretability. Taken together with the interpretability distinctions from Jalilpour and Müller-Putz [10] and Wei et al. [12] discussed above, ERP-XTTN is, to our knowledge, the first ErrP classifier to combine LOSO cross-validation, causal filtering, no spatial filtering, and inherent interpretability on standard benchmarks, with competitive AUROC scores.

Limitations: Several limitations should be noted. The evaluation is limited to two datasets totaling 17 subjects, which limits the strength of the generalization claims. All results are based on offline replay; no online or real-time validation was performed. Causal filtering addresses only one prerequisite of online deployment; asynchronous detection without time-locking markers remains an open challenge. No hyperparameter tuning was conducted for either model, so reported performance may not reflect the full capacity of either architecture. Fixed prototypes are derived from grand-average difference waves and may not optimally capture individual variation in ERP morphology. No direct comparison with classical pipelines

(xDAWN+LDA) was included. Finally, balanced accuracy lags AUROC more for ERP-XTTN than for EEG-Net, suggesting that the fixed 0.5 decision threshold is suboptimal in the LOSO setting. In deployment, where a fixed threshold must be chosen without per-subject calibration, this gap warrants further investigation.

CONCLUSION

This work introduced ERP-XTTN, a cross-attention ErrP classifier that routes input EEG through fixed, data-derived prototypes grounded in canonical ErrP morphology. We evaluated it under strict deployment-relevant constraints across two public datasets and three channel configurations, demonstrating competitive performance with inherent per-trial interpretability.

Several directions remain for future work. Expanded evaluation across additional datasets (including different ERPs such as P300, N200, and feedback-related negativity), FIR filter comparison, and classical pipeline benchmarking (xDAWN+LDA) are planned for a journal extension. Online real-time validation is the most important next step for assessing deployment viability. Minimal subject-specific fine-tuning with a brief calibration session could further close the performance gap while retaining the interpretability benefits of the fixed-prototype architecture.

ERP-XTTN demonstrates that inherent interpretability in ErrP classification need not come at a substantial performance cost. With 3 midline channels, it achieves AUROCs of 0.782 on BNCI and 0.836 on HRI, within 0.026 of EEGNet on average, while providing direct, per-trial transparency through interpretable routing over data-driven ERP prototypes. These results hold under causal filtering, LOSO cross-validation, and no spatial filtering, satisfying the constraints that online, calibration-free BCI deployment demands.

Code is publicly available at <https://github.com/cgenevier/ERP-XTTN>; the version used to produce these results is archived on Zenodo [17].

REFERENCES

- [1] Chavarriaga R, Sobolewski A, Millán JDR. Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*. 2014;8:208.
- [2] Sabio J, Williams NS, McArthur GM, Badcock NA. A scoping review on the use of consumer-grade EEG devices for research. *PLOS ONE*. 2024;19(3):e0291186.
- [3] Lopes-Dias C et al. Online asynchronous detection of error-related potentials in participants with a spinal cord injury using a generic classifier. *Journal of Neural Engineering*. 2021;18(4):046022.
- [4] Bhattacharyya S, Konar A, Tibarewala DN, Hayashibe M. A Generic Transferable EEG Decoder for Online Detection of Error Potential in Target Selection. *Frontiers in Neuroscience*. 2017;11:226.
- [5] Schönleitner FM, Otter L, Ehrlich SK, Cheng G. Calibration-Free Error-Related Potential Decoding With Adaptive Subject-Independent Models: A Comparative Study. *IEEE Transactions on Medical Robotics and Bionics*. 2020;2(3):399–409.
- [6] Ancau DM, Ancau M, Ancau M. Deep-learning online EEG decoding brain-computer interface using error-related potentials recorded with a consumer-grade headset. *Biomedical Physics & Engineering Express*. 2022;8(2):025006.
- [7] Wimmer M, Weidinger N, Veas E, Müller-Putz GR. Multimodal decoding of error processing in a virtual reality flight simulation. *Scientific Reports*. 2024;14(1):9221.
- [8] Ren G, Mahmoud SS, Kumar A, Fang Q, Yu B. A Transformer Encoder and Convolutional Neural Network Combined Method for Classification of Error-related Potentials. In: 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE: Toronto, ON, Canada, Oct. 2023, 1–5.
- [9] Ren G, Kumar A, Mahmoud SS, Fang Q. A deep neural network and transfer learning combined method for cross-task classification of error-related potentials. *Frontiers in Human Neuroscience*. 2024;18:1394107.
- [10] Jalilpour S, Müller-Putz G. A framework for Interpretable deep learning in cross-subject detection of event-related potentials. *Engineering Applications of Artificial Intelligence*. 2025;139:109642.
- [11] Chan HTJ, Wimmer M, Šimić I, Müller-Putz GR, Veas E. Informing EEG-Based Error Decoding With Explainable AI. In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE: Vienna, Austria, Oct. 5, 2025, 4550–4557.
- [12] Wei W, Qiu S, Zhang Y, Mao J, He H. ERP prototypical matching net: A meta-learning method for zero-calibration RSVP-based image retrieval. *Journal of Neural Engineering*. 2022;19(2):026028.
- [13] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.
- [14] Chavarriaga R, Millán JDR. Learning From EEG Error-Related Potentials in Noninvasive Brain-Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2010;18(4):381–388.
- [15] Ehrlich SK, Cheng G. A Feasibility Study for Validating Robot Actions Using EEG-Based Error-Related Potentials. *International Journal of Social Robotics*. 2019;11(2):271–283.
- [16] Ferrez PW, Millán JDR. Error-Related EEG Potentials Generated During Simulated Brain-Computer Interaction. *IEEE Transactions on Biomedical Engineering*. 2008;55(3):923–929.
- [17] Wyman CG. *ERP-XTTN: Interpretable cross-attention ERP classifier*. Version v1.0.0. 2026. [Online]. Available: <https://doi.org/10.5281/zenodo.20087551>.

TREAT: a toolbox for trial selection and alignment of unlabeled neural data

S.H. Geukes¹, A. Bekius¹, F. Fabris^{1,2}, Z.V. Freudenburg¹, N.F. Ramsey^{1,2}, M.P. Branco^{1,2}

¹ Department of Neurology and Neurosurgery, University Medical Center Utrecht Brain Center, Utrecht University, Utrecht, The Netherlands

² Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

E-mail: m.pedrosobranco@umcutrecht.nl

ABSTRACT: Brain-computer interfaces (BCIs) enable paralyzed individuals to control external devices using brain signals. For training decoders, trial-to-trial variability in response time may obscure meaningful task-related neural activity. We introduce the TRial sElection and AlignmenT (TREAT) toolbox to provide an automated neural event onset detection method that supports both trial selection and event alignment, based on the slope of the neural response (neural slope marker, NSM). We validated TREAT using high-density electrocorticography recordings from three able-bodied individuals performing a speech task. We extracted high-frequency and alpha band power and compared the NSM to the voice onset times. For two of the three participants, the neural slope markers preceded voice onset by 0.4 seconds. Trial selection improved the precision of neural onset estimates across trials as the spread of NSM timing compared to voice onset time decreased. TREAT offers event-detection in the absence of behavioral markers and may thereby improve interpretation of neural activity and training of BCI decoding models.

INTRODUCTION

Brain-computer interfaces (BCIs) enable individuals with paralysis to control external devices for communication or environmental control, without requiring any muscle output [1]. Variability in task responses due to missed trials and variable reaction times can obscure meaningful task-related neural activity, particularly when signals are averaged across trials [2]. Time-locking neural activity to behavioral markers (obtained for example with a data glove or microphone) typically mitigates this variability [3], [4], [5]. However, behavioral markers are unavailable for individuals who cannot move or speak; individuals who may particularly benefit from a BCI. On top of that, the absence of movement-related sensory feedback in this population can itself introduce variability in task responses, which may be further increased by disease-related factors such as reduced attention, as observed for example in amyotrophic lateral sclerosis (ALS) [6]. These limitations underscore the need for signal-curation methods for trial selection/rejection and response alignment that do not depend on behavioral markers but rely solely on neural data. Others previously presented

trial selection based on both the trial's contribution to classification accuracy and templates of independent components reflecting expected motor-related power changes in the neural signal [7]. Here we propose to accomplish trial selection by thresholding the change in signal amplitude after the cue relative to a baseline period. In addition, previously proposed alignment methods relied either on the peak [2] or the shape [8] of the neural response. To most accurately capture movement onset and maximize the contrast between movement and rest, we previously proposed alignment based on the increase in neural activity, more specifically the slope of the response [3]. We extend this approach by developing a toolbox that integrates both trial selection and alignment without the need for behavioral markers. Here, we introduce the TRial sElection and AlignmenT (TREAT) toolbox: a user-friendly automated neural event onset detection method for signals acquired with electrocorticography (ECoG) and electroencephalography (EEG). This toolbox facilitates offline analysis of unlabeled datasets, supporting both trial selection and trial alignment based on neural

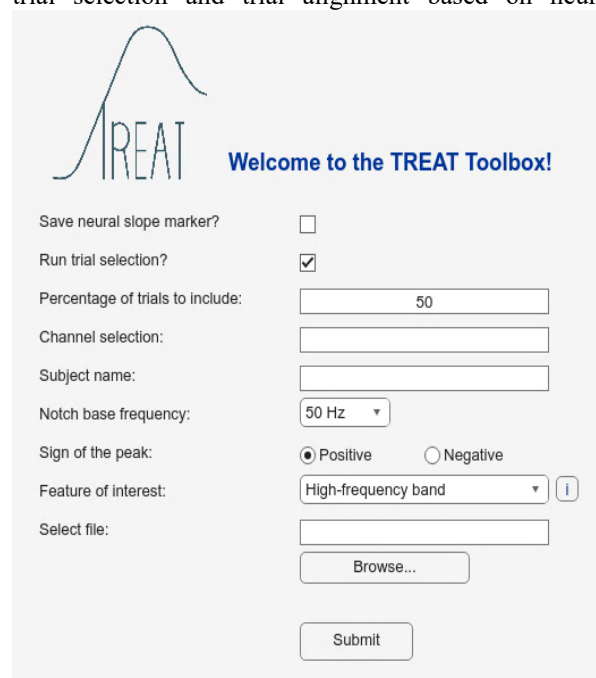


Figure 1: Graphical user interface of the TREAT toolbox.

features. We validated this approach in a high-density ECoG dataset from three able-bodied individuals performing a speech task, by comparing alignment with TREAT to alignment with a behavioral marker: the voice onset time (VOT).

MATERIALS AND METHODS

The TREAT toolbox: The toolbox consists of a graphical user interface developed in MATLAB 2025b (The MathWorks, Natick, USA) (Fig. 1). The interface allows the user to input raw data with .dat (BCI2000 [9]), .nev/.nsx (Blackrock Neurotech, Salt Lake City, USA) or .trc files (Micromed, Treviso, Italy) formats. The toolbox preprocesses the data, segments it into trials and performs trial selection (optional) and trial alignment (see Fig. 2 for an overview of this process).

Preprocessing: The user provides a selection of channels to be analyzed. This selection should exclude flat or noisy channels. Subsequently, the user can specify 1) the center frequency of the line noise (50 or 60 Hz), 2) the feature of interest, high-frequency band (HFB; 65-95 Hz), alpha band (8-12 Hz), beta band (13-30 Hz) or local motor potentials, as well as 3) the expected direction of the feature's effect (e.g., an increase in HFB power as a positive effect, or a decrease in alpha band power as a negative effect). The toolbox then applies a common average reference filter to remove noise shared by all signals, a band-pass filter (0.15-120 Hz; 3rd order Butterworth), a notch filter at the indicated notch base frequency (4th order Butterworth), and down samples the data to 250 Hz. Spectral features are extracted using a Gabor wavelet dictionary, and the local-motor potentials are extracted by applying a 375ms causal moving average on the time-domain data [10].

Trial segmentation and channel selection: The user determines the event code that corresponds to the relevant stimulus being presented to the participant (active cue, Fig. 2), based on a figure that displays all event codes in the data. The user subsequently determines the trial length (TL). The toolbox then extracts active (0s to TL) and rest trials (-TL to 0s) centered around the active cues. To determine which channels are included in subsequent analysis, the signed R^2 value is calculated for each channel by averaging the signal across time for all active and rest trials and regressing those values to a task-regressor (0 for rest and 1 for active trials). Channels are included if the response is significant ($p < 0.05$) and the sign of the R^2 corresponds to the direction of the effect selected by the user (see Preprocessing). At least two channels should be included for method to proceed. The signal is then averaged over the significant channels and smoothed with a 0.2 second window.

Trial selection: This is an optional step in the toolbox. The user enters in the graphical user interface the percentage of trials (T%) to be included in further analysis. The selected trials correspond to the T% of trials with the largest ratio (R_t , where t is a trial) of change in the feature of interest following the active cue ($\Delta y_t =$

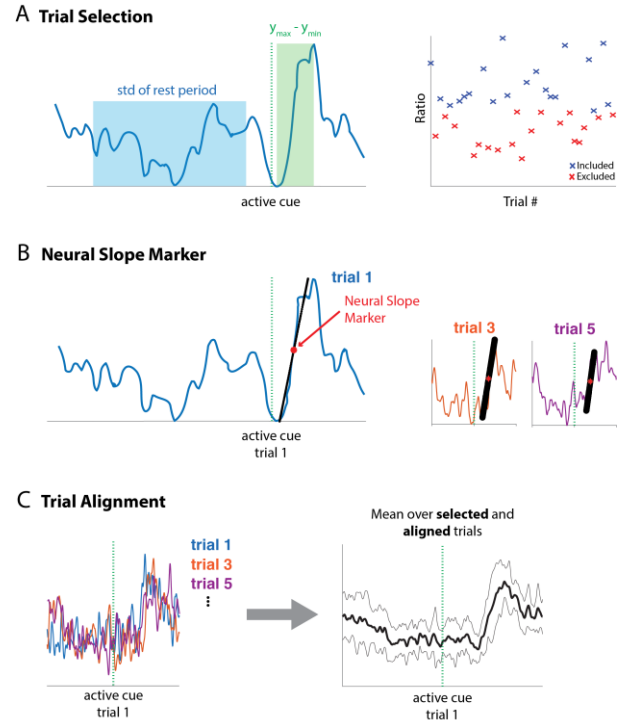


Figure 2: Example output of the TREAT toolbox for one participant. A) Trial selection computes a ratio of the range of active period versus the standard deviation of ‘rest’ period data per trial. A chosen percentage of trials with highest values is selected. B) Trial alignment slides a line segment across trials to identify the minimum distance between the line and the signal. C) Trials are aligned to a reference trial (trial 1).

$y_{max} - y_{min}$ in the time window 0.3 to 1.5 seconds) relative to the standard deviation of the rest period prior to the cue ($\sigma_{rest,t}$; -1.5 to -0.3 seconds) (Fig. 2A):

$$R_t = \frac{\Delta y_t}{\sigma_{rest,t}}. \quad (1)$$

Trial alignment: The alignment procedure builds on the method previously reported in [3] that uses the midpoint of the HFB increase as a marker for movement onset (Fig. 2B). In the current toolbox, this method is extended to other neural features as well. In short, per trial the method slides a line segment over the neural feature (averaged across significant channels) and determines the best fit by minimizing the distance between line and signal (Fig. 2B). It then uses the midpoint of the sliding line segment, hereafter referred to as the ‘neural slope marker’ (NSM), to align all trials (Fig. 2C). In the graphical user interface, the user chooses a time window after the active cue within which the NSM should be detected (Fig. 2B). To compute the slope of the line segment, TREAT first calculates the slope per trial (m_t) by dividing the signal amplitude change in the feature of interest (Δy_t) by the difference in the (time) samples corresponding to the minimum and maximum amplitude values (Δn_t), which is then averaged across trials.

$$m_t = \frac{\Delta y_t}{\Delta n_t} \quad (2)$$

When sliding the line segment over the time window, the method computes for each time point the distance between all the unique y-values of the signal and the line segment that are smaller than a threshold (ε : 0.5 seconds). This threshold ensures that distant time points are not included (for details see [3]). Finally, the midpoint of the line segment corresponding to the minimum distance is used as the NSM (Figure 2B, red point), and each trial is shifted such that its NSM is aligned with the NSM of the first selected trial (reference trial; Fig. 2C). The toolbox displays several steps in the process, including the trial-averaged signal, the best fit of the line segment in each trial and the distance between the two lines for each time point.

Validation data set: We validated the toolbox with a high-density ECoG dataset, of which the task was previously described in detail [4]. In short, three participants (P1: female, 36 years old; P2: female, 30 years old; P3: male, 19 years old) implanted with clinical ECoG grids for epilepsy monitoring, had an additional high-density ECoG grid (3-4 mm inter-electrode distance, 1-1.17 mm exposed electrode diameter) implanted over the sensorimotor cortex for research purposes. Only the high-density data is considered for the current analysis. Four channels with flat or noisy signals were excluded from analysis for P3. Each participant performed 39 trials, in which they produced the /i/ vowel 3-5 times. Voice onsets of the first vowel were determined with Praat software [11]. With the TREAT toolbox, we extracted the HFB and alpha band and compared the timing of the NSM to the timing of the voice onsets. We ran trial selection, including both 100% and 50% of the trials. To investigate potential temporal bias in the distribution of neural slope markers, we also investigated the distribution of NSM over time on the HFB trace of a baseline recording of participant P2 with randomly generated event codes (repeated 1000 times). To run the TREAT toolbox, we chose a trial length of TL = 3 seconds for all participants and frequency bands. The period to search for the NSM was selected per participant based on the timing of their neural responses (Tab. 1).

Table 1: Time periods for sliding the line segment relative to the cue (in seconds) per participant and neural feature.

<i>Participant</i>	<i>HFB</i>	<i>Alpha band</i>
P1	0.3 - 1.8	0.1 - 2.0
P2	0.1 - 1.8	0.1 - 2.5
P3	0.1 - 1.8	0.1 - 2.2

RESULTS

Overall, neural responses to the task were strong for P1 and P2, and less pronounced for P3, as can be seen by the high number of significant responding channels for HFB

and alpha bands (Tab. 2). With TREAT we compared no trial rejection with 50% trial rejection. We noted that a 50% trial rejection increased the signal amplitude change in the HFB by 12.00% for P1, 26.68% for P2 and 17.37% for P3, and in the LFB by 14.11% for P1, 42.51% for P2 and 67.96% for P3.

Table 2: Number of channels with significant positive (in case of the HFB) or negative (in case of the alpha band) R2 values ($p < 0.05$) together with the number of channels included in the analysis.

<i>Participant</i>	<i>HFB</i>	<i>Alpha band</i>
P1	39/128	66/128
P2	68/128	20/128
P3	15/60	12/60

In addition, when using 50% trial selection, the cut-off ratio (see Fig. 2A for an illustration) was larger than 1 for all participants and signal features, apart from the alpha band of P3 (Tab. 3). This indicates that for at least half of the trials, the change in HFB or alpha band power following the cue was larger than the standard deviation over the rest period. Selecting the 50% trials with ratios equal to or larger than the cut-off value resulted in more precise estimation of the NSM relative to VOT compared to including all trials, as reflected by the reduced spread around the means (Fig. 3).

Table 3: Cut-off ratio for 50% trial selection.

<i>Participant</i>	<i>HFB</i>	<i>Alpha band</i>
P1	7.50	5.27
P2	8.07	1.78
P3	4.36	0.98

Our results also show that the NSM occurred before voice onset time (VOT, Fig. 3) for all three participants. For the HFB, the average NSM across trials occurred 0.450 seconds (± 0.229) prior to VOT for P1, 0.407 seconds (± 0.407) for P2 and 0.017 seconds (± 0.711) for P3. For the alpha band, the average NSM across trials occurred 0.420 seconds (± 0.312) for P1, 0.528 seconds (± 0.576) for P2 and 0.008 seconds (± 0.801) for P3 prior to VOT. Repeatedly calculating the NSMs with randomly generated events over a baseline recording resulted in a bimodal distribution, with peaks at the beginning (0.3-0.4 seconds) and end (1.7-1.8 seconds) of the sliding window (Fig. 4A). In contrast, the NSMs acquired with the true events followed a single-peaked distribution (Fig. 4B).

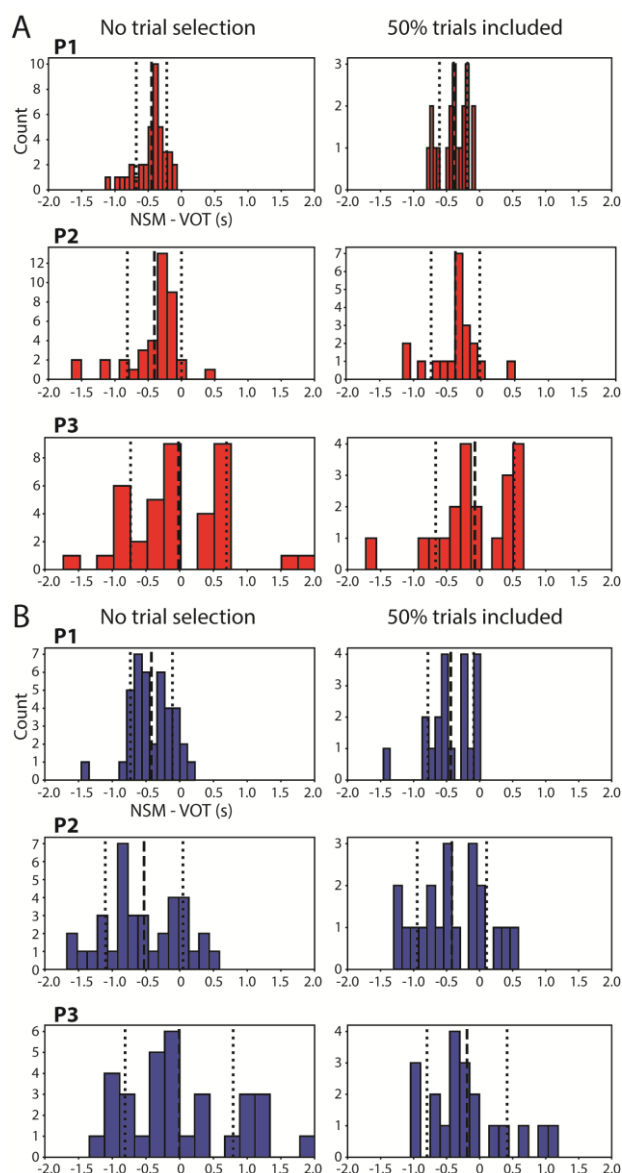


Figure 3: Neural slope marker (NSM) output relative to voice onset time (VOT). We ran the TREAT toolbox without trial selection (left column) and with 50% trial selection for both the high frequency band (A) and the alpha band (B). The dashed line indicates the mean value across trials; the dotted line indicates the standard deviation.

DISCUSSION

Here we introduced the TREAT toolbox, which selects trials based on the magnitude of the neural response to a task, and aligns these responses using the slope of the neural response. We found that the neural onset preceded VOT. This is expected, as the neural response over the sensorimotor cortex initiates the muscle contractions necessary for voice onset to occur [12]. Our results show that NSM centered approximately 400 milliseconds prior to VOT for both the HFB and alpha band for two out of three participants. This value is in line with previous reports that show larynx movement 250 milliseconds before VOT when repeatedly producing the /a/ vowel,

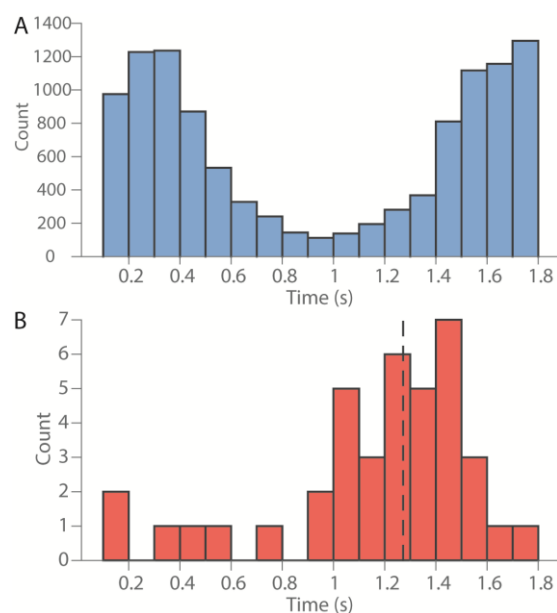


Figure 4: A) TREAT toolbox output with randomly generated event times for the HFB of a baseline recording for P2. 1000 iterations resulted in 293 runs with at least 2 electrodes significantly responding to the task. B) Neural slope marker output for the HFB of P2. The dashed line indicates the median value.

and identified magnetoencephalography activity at least 125 milliseconds prior to movement onset [13], and others measuring EEG activity over the sensorimotor cortex 300 milliseconds prior to voice onset in picture naming tasks [12], [14]. Moreover, spoken phonemes have been shown possible to be decoded from a window 500 milliseconds prior to VOT [5].

Although our sample size is limited, our results indicate that the HFB may serve as a more precise movement-onset marker than the alpha band, which agrees with the hypothesis that HFB power reflects neural population firing underneath the electrodes [15], [16], [17], [18]. The spread of NSM values relative to VOT was larger for P3 than P1 and P2, which indicates that identified NSM were less reliable for P3. This may be related to the relatively poor neural response of P3 compared to the other two participants, as reflected by the lower number of significantly responding channels and cut-off ratios for trial selection. Selection of 50% trials with the strongest response to the active cue reduced the spread of NSM values relative to VOT.

Comparing with other alignment methods, aligning the slopes of the neural responses should increase both the signal amplitude and sharpen the onset in the average across trials, in contrast to aligning to the peak [2] or shape [8] of the responses, that increase the amplitude, but do not sharpen the signal onset per se.

Limitations and future directions: Running the toolbox through data with randomly generated event labels indicates that the random detection of a NSM is more likely at the beginning and the end of the sliding window. This result is expected as at the edges of the sliding window the number of points to compute the difference between the line segment and the signal that falls under

the threshold of 0.5 seconds is smaller than in the middle of the sliding window. Secondly, the participants pronounced the /i/ vowel multiple times within each trial. If they returned their tongues to resting position in between repetitions, each articulation could constitute a distinct movement onset, each accompanied by its own peak in HFB power [4]. As a result, the TREAT toolbox may have detected movement onset not at the first vowel articulation, but later in the trial, if that articulation elicited a stronger increase in HFB power. More extensive comparisons to behavioral markers are required to further validate our approach, including different movements.

Further optimization of the method will involve more stringent electrode selection, fixed line segment size across trials, and optimizing the various settings in the toolbox, such as the percentage of trials to include and signal smoothing, which may be particularly relevant for the HFB as that is an inherently noisy signal. In conclusion, the TREAT toolbox may serve as a versatile event-detection method when behavioral markers are absent, which may improve the interpretation of neural activity as well as the training of decoding models for BCIs.

ACKNOWLEDGEMENTS

This research was supported by the National Institute of Neurological Disorders and Stroke (CortiCom project, grant number: UH3NS114439), the Dutch Technology Foundation STW (PANDA project, grant number 19072), and Health~Holland Top Sector Life Sciences and Health (PPP Allowance made available to the Dutch Brain Foundation; NOTION project, grant number: PPS-2021-02).

REFERENCES

- [1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002, doi: 10.1016/S1388-2457(02)00057-3.
- [2] W. G. Coon and G. Schalk, "A method to establish the spatiotemporal evolution of task-related cortical activity from electrocorticographic signals in single trials," *J. Neurosci. Methods*, vol. 271, pp. 76–85, 2016, doi: 10.1016/j.jneumeth.2016.06.024.
- [3] M. P. Branco, Z. V. Freudenburger, E. J. Aarnoutse, M. G. Bleichner, M. J. Vansteensel, and N. F. Ramsey, "Decoding hand gestures from primary somatosensory cortex using high-density ECoG," *Neuroimage*, vol. 147, pp. 130–142, 2017, doi: 10.1016/j.neuroimage.2016.12.004.
- [4] E. Salari, Z. V. Freudenburger, M. J. Vansteensel, and N. F. Ramsey, "Repeated Vowel Production Affects Features of Neural Activity in Sensorimotor Cortex," *Brain Topogr.*, vol. 32, no. 1, pp. 97–110, Jan. 2019, doi: 10.1007/S10548-018-0673-4.
- [5] N. F. Ramsey, E. Salari, E. J. Aarnoutse, M. J. Vansteensel, M. G. Bleichner, and Z. V. Freudenburger, "Decoding spoken phonemes from sensorimotor cortex with High-Density ECoG grids," *Neuroimage*, vol. 180, no. Pt A, p. 301, Oct. 2017, doi: 10.1016/J.NEUROIMAGE.2017.10.011.
- [6] S. Abrahams, "Neuropsychological impairment in amyotrophic lateral sclerosis–frontotemporal spectrum disorder," *Nature Reviews Neurology* 2023 19:11, vol. 19, no. 11, pp. 655–667, Oct. 2023, doi: 10.1038/s41582-023-00878-z.
- [7] B. Zhou, X. Wu, Z. Lv, L. Zhang, and X. Guo, "A Fully Automated Trial Selection Method for Optimization of Motor Imagery Based Brain-Computer Interface," *PLoS One*, vol. 11, no. 9, p. e0162657, Sep. 2016, doi: 10.1371/JOURNAL.PONE.0162657.
- [8] A. H. Williams *et al.*, "Discovering Precise Temporal Patterns in Large-Scale Neural Recordings through Robust and Interpretable Time Warping," *Neuron*, vol. 105, no. 2, pp. 246–259.e8, 2020, doi: 10.1016/j.neuron.2019.10.020.
- [9] G. Schalk and J. Mellinger, "A practical guide to brain-computer interfacing with BCI2000: General-purpose software for brain-computer interface research, data acquisition, stimulus presentation, and brain monitoring," *A Practical Guide to Brain-Computer Interfacing with BCI2000: General-Purpose Software for Brain-Computer Interface Research, Data Acquisition, Stimulus Presentation, and Brain Monitoring*, pp. 1–260, 2010, doi: 10.1007/978-1-84996-092-2/COVER.
- [10] G. Schalk *et al.*, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 4, no. 3, p. 264, Jun. 2007, doi: 10.1088/1741-2560/4/3/012.
- [11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 1992. [Online]. Available: <https://www.praat.org>
- [12] M. Laganaro, "Time-course of phonetic (motor speech) encoding in utterance production," *Cogn. Neuropsychol.*, vol. 40, no. 5–6, pp. 287–297, Aug. 2023, doi: 10.1080/02643294.2023.2279739.
- [13] H. Kothare *et al.*, "Temporal specificity of abnormal neural oscillations during phonatory events in laryngeal dystonia," *Brain Commun.*, vol. 4, no. 2, Mar. 2022, doi: 10.1093/BRAINCOMMS/FCAC031.
- [14] M. Hassan, P. Benquet, A. Biraben, C. Berrou, O. Dufor, and F. Wendling, "Dynamic reorganization of functional brain networks during picture naming," *Cortex*, vol. 73, pp.

- 276–288, Dec. 2015, doi:
10.1016/J.CORTEX.2015.08.019.
- [15] K. J. Miller *et al.*, “Spectral Changes in Cortical Surface Potentials during Motor Movement,” *Journal of Neuroscience*, vol. 27, no. 9, pp. 2424–2432, 2007, doi: 10.1523/JNEUROSCI.3886-06.2007.
- [16] K. J. Miller, L. B. Sorensen, J. G. Ojemann, and M. den Nijs, “Power-law scaling in the brain surface electric potential,” *PLoS Comput. Biol.*, vol. 5, no. 12, p. e1000609, Dec. 2009, doi: 10.1371/journal.pcbi.1000609.
- [17] N. E. Crone, D. L. Miglioretti, B. Gordon, and R. P. Lesser, “Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band,” *Brain*, vol. 121, no. 12, pp. 2301–2315, 1998, doi: 10.1093/brain/121.12.2301.
- [18] A. Dubey and S. Ray, “Cortical Electroencephalogram (EEG) Is a Local Signal,” *The Journal of neuroscience*, vol. 39, no. 22, pp. 4299–4311, May 2019, doi: 10.1523/JNEUROSCI.2917-18.2019.

Selection and Validation of Spectral EEG Markers for Real-Time Monitoring of Attention and Cognitive Engagement in Learning Contexts

P.-B. Mathieu de Carvalho^{1,2}, M.-C. Corsi², L. Bougrain^{1,2}

¹Université de Lorraine, CNRS, LORIA, Nancy, France;

² Sorbonne Université, ICM, CNRS, Inria, Inserm, Paris, France

E-mail: pierre-baptiste.mathieu-de-carvalho@inria.fr

ABSTRACT: Electroencephalography (EEG) provides precise neural measures for brain-computer interface (BCI) applications monitoring cognitive states. This empirical study evaluates the sensitivity of established EEG spectral markers to distinguish sustained attention from cognitive engagement. Fifteen participants completed a sustained attention task (D2-R), a N-Back working memory task (2-Back), and an ecological video learning task. Results indicate that the sustained attention task induced significant modulations in the $\beta/(\alpha + \theta)$ ratio, posterior theta, and occipital beta power. Conversely, the cognitive engagement task yielded no significant spectral deviations, highlighting potential overlaps in cognitive networks and the critical impact of task-order effects on post-effort baselines. Furthermore, linear mixed models revealed distinct, non-linear temporal dynamics during the ecological task, demonstrating that aggregating EEG metrics over the entire task duration obscures critical real-time state fluctuations. This study provides a rigorous framework for selecting appropriate EEG markers, emphasizing the necessity of continuous monitoring and controlled experimental designs in educational BCI applications.

INTRODUCTION

The capacity to maintain attention over extended periods and to actively engage cognitive resources constitutes fundamental aspects of human performance across educational, professional, and clinical contexts. Brain-computer interfaces (BCIs) increasingly require robust, real-time estimation of these cognitive states to enable adaptive systems that respond dynamically to user mental states [1].

Electroencephalography (EEG) offers unique advantages for monitoring cognitive states: high temporal resolution, non-invasiveness, portability, and relative cost-effectiveness compared to neuroimaging techniques. However, the multiplicity of available EEG markers and their overlapping sensitivities necessitate a clear conceptual framework distinguishing sustained attention from cognitive engagement.

Sustained attention refers to the ability to maintain vigilance and receptivity to stimuli over time, characterized by resistance to fatigue and mind-wandering. Its decline

manifests as increased response times, errors, and subjective reports of mental fatigue [2].

Cognitive engagement, conversely, reflects the active investment of mental resources in task processing, modulated by task difficulty, motivation, and strategic control. It encompasses the depth of information processing rather than mere wakefulness [3].

This study examines EEG markers across two principal analytical domains: (1) spectral power measures and (2) brain connectivity metrics.

Spectral Power Signatures: Specific frequency bands across cortical regions provide distinct signatures for sustained attention and cognitive engagement (Tab. 1).

- **Sustained Attention Markers:** Maintaining vigilance modulates multiple bands. Prolonged tasks induce progressive δ and θ power increases over occipital regions, reflecting mental fatigue and decreased alertness [2], whereas frontal θ increases indicate compensatory effort [4]. Parieto-occipital α acts as an attentional filter: external focus induces desynchronization [5, 6], while synchronization signals disengagement [7, 8]. Posterior β exhibits duality: occipito-temporal decreases precede lapses [2], while occipital increases predict active processing [9]. The centro-parietal $\beta/(\alpha + \theta)$ ratio aggregates these dynamics into a continuous vigilance index [1, 10].
- **Cognitive Engagement Markers:** Active mental investment relies on distinct spectral modulations. Frontal Midline Theta (FMT) transiently increases with strategic mental investment and cognitive load [3, 11]. Engagement intensity is indexed by strong parietal α suppression (11–13 Hz), correlating with motivation and learning efficacy [12]. Fronto-central β modulations reflect efficient top-down control [11, 13], and left-temporal γ (>30 Hz) marks high-quality information integration [14].

Brain Connectivity Measures: While spectral power analysis examines activity within isolated brain regions, functional connectivity approaches assess statistical dependencies between regional activities, thereby capturing

Table 1: Synthesis of Spectral EEG Estimators and Corresponding Experimental Variables

Domain	Estimator (Feature Name)	Modulation Interpretation	Target Regions
Sustained Attention	Occipital Theta & Delta Power (Attention_Theta)	Increase: Reflects mental fatigue installation and alertness decline [2, 4].	Occipital (O1, O2, Oz)
	Posterior Alpha Power (Attention_Alpha)	Increase: Disengagement, internal attention, mind-wandering [7, 8]. Decrease: External stimulus focus [5, 15].	Parieto-occipital (Pz, O1, O2)
	Occipital Beta Power (Attention_Beta)	Decrease: Associated with vigilance decline. High baseline required for alertness maintenance [2, 9].	Occipital (O1, O2, Oz)
	Ratio $\beta/(\alpha + \theta)$ (Attention_Ratio)	Increase: High vigilance index [1]. Decrease: Fatigue/drowsiness index [10].	Centro-parietal (Cz, Pz, P3, P4)
Cognitive Engagement	Frontal Midline Theta (Engagement_Theta)	Increase: Strategic mental investment, controlled motivated effort or compensatory effort [3, 4, 11].	Frontal midline (Fz, FC1, FC2)
	Parietal Alpha Suppr. (Engagement_Alpha)	Stronger decrease: Intense, efficient attentional allocation. Sensitive to motivation, predicts learning [3, 11, 12].	Parietal α_2 (11–13 Hz) (Pz, PO3, PO4)
	Beta Power	Increase: Reflects efficient top-down control, especially in experts [11, 13].	Fronto-central

the inherently interconnected nature of brain functioning [16].

- **Sustained Attention:** In the context of sustained attention, graph-theoretical approaches have revealed that attention decline is associated with network disorganization, characterized by decreased local segregation and reduced global integration. Multiple studies demonstrate reduced clustering coefficient (CC) and increased characteristic path length (CPL) during fatigue, particularly in the theta band [17]. Similarly, dyslexic children performing attention tasks exhibit lower CC, reduced global efficiency, and longer CPL compared to controls, predominantly in θ and α bands [18]. However, the heterogeneity across studies and occasional inverse findings suggest the potential involvement of compensatory mechanisms [19].

Beyond local network topology, large-scale network interactions, particularly between internally-oriented networks (e.g., the Default Mode Network, DMN) and task-oriented networks, provide robust markers of attentional state. Sustained attention is characterized by stronger anti-correlations between these networks, which are significantly enhanced during high-performance periods [20]. The fronto-parietal control network (FPCN) functions as a dynamic switch: it couples with the dorsal attention network (DAN) during attentive states and shifts toward the DMN during attention lapses, with connection strength predicting individual attentional performance even at rest [21]. Complementing these functional connectivity findings, effective connectivity analyses reveal that voluntary attentional control manifests as top-down information flow from prefrontal to parietal cortex, particularly dominant in the θ band, which is considered a neural signature of the cognitive effort required to maintain at-

tentional focus [22, 23].

- **Cognitive Engagement:** Turning to cognitive engagement, similar connectivity patterns emerge but with distinct characteristics. Increased engagement promotes a more integrated and efficient network architecture, with graph theoretical analyses demonstrating increased mean node degree [24] and enhanced betweenness centrality of central-frontal regions [25]. Notably, dynamic functional connectivity (dFC) patterns have proven particularly valuable as neural signatures for predictive modeling of engagement. fMRI-based dFC models successfully predict momentary fluctuations in engagement [25, 26], revealing that high engagement is associated with increased within-DMN connectivity and enhanced DMN-FPCN connectivity, while within-FPCN connectivity decreases [26]. These findings are corroborated by EEG-based dFC models, which emphasize the importance of central region connections [25], while leading eigenvector dynamics analysis (LEiDA) identifies recurrent connectivity states with remarkable discriminative power, achieving nearly 99% accuracy in classifying social engagement contexts [24].

Collectively, these findings highlight the potential of functional connectivity measures as biomarkers of both sustained attention and engagement. A promising avenue for future research would be to systematically evaluate various connectivity estimators using independent datasets, establishing their reliability and generalizability across different experimental paradigms and populations.

However, while connectivity measures offer rich insights into network dynamics, spectral power features remain the most established and computationally efficient candidates for real-time BCI implementation in educational settings. Therefore, to test

the replicability and relevance of these spectral findings, we developed an experimental protocol combining standardized cognitive tasks with an ecological learning task (viewing a scientific lecture). The empirical contribution of this study focuses on evaluating the sensitivity of selected spectral EEG markers to distinguish sustained attention from cognitive engagement in a cohort of university students.

MATERIALS AND METHODS

Participants: Fifteen healthy, native French-speaking university students (2 females; age 23.1 ± 2.2) with no neurological history participated. To ensure the validity of the learning task, all were non-experts in immunology. The protocol was approved by the INRIA ethics committee (COERLE).

Experimental protocol: The session lasted approximately 90 minutes in a controlled environment. The protocol consisted of four sequential blocks:

1. Sustained Attention Task (D2-R): A paper-based cancelation task assessing attention speed and precision under time pressure [27]. Behavioral performance was measured using the Concentration Performance (CP) score.
2. Cognitive Engagement Proxy (N-Back): A computerized 2-back working memory task (PsyToolkit [28]) soliciting executive functions and working memory updates. Performance was evaluated using the discriminability index (d').
3. Baseline Recording: Two 30-second resting-state EEG baselines (eyes open and closed).
4. Ecological Learning Task (Video): A 27-minute scientific video lecture on immunology (Collège de France). Following the video, participants completed a 20-item comprehension questionnaire and 10-point Likert scales evaluating subjective fatigue and engagement.

Data Acquisition and Feature Extraction: EEG was recorded using a 32-channel BioSemi ActiveTwo system (10-20 montage, 2048 Hz sampling, impedances < 20 k Ω). Preprocessing in MNE-Python included band-pass (1-100 Hz) and notch (50 Hz) filtering, followed by Infomax ICA for artifact removal. Power spectral density (PSD) was estimated via Welch's method on 4-second epochs with a 50% overlap [29, 30]. Median power across epochs was calculated to mitigate transient noise. Targeted spectral biomarkers (Delta: 1-4 Hz, Theta: 4-8 Hz, Alpha: 8-13 Hz, Beta: 13-30 Hz) were computed by aggregating standard frequency bands over predefined electrode clusters, strictly corresponding to the theoretical markers synthesized in Tab. 1 (e.g., Attention_Theta).

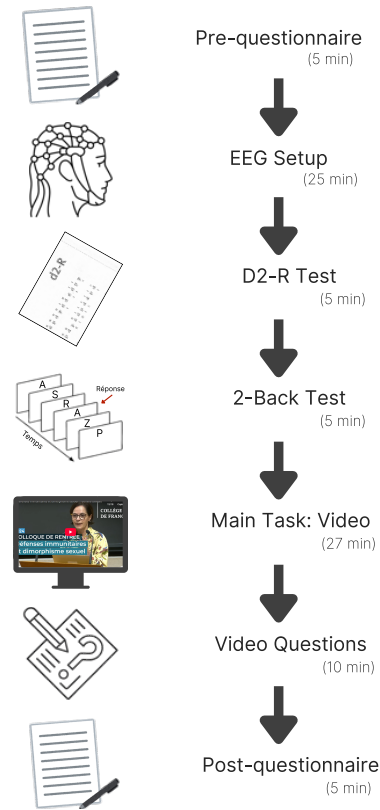


Figure 1: Timeline of the experimental protocol.

RESULTS

Predefined EEG markers during tasks (D2-R and 2-Back): To evaluate whether the biomarkers identified in previous literature reliably reflect sustained attention and cognitive engagement within our specific experimental paradigm, we conducted direct evaluations and comparisons between the two tasks (D2-R and 2-Back). Predefined EEG marker (Tab. 1) differences between the cognitive tests were assessed using Wilcoxon signed-rank tests. As detailed in Tab. 2, significant differences were observed for the Attention_Ratio, Attention_Theta, and Attention_Beta markers ($p < 0.01$). Conversely, no significant differences were found for the predefined engagement markers (Engagement_Theta and Engagement_Alpha) between the two tasks.

Table 2: FDR-corrected statistical comparison of predefined EEG markers between Sustained Attention (D2-R) and Cognitive Engagement (2-Back) tasks.

Marker	Stat (W)	p_{FDR}	η^2
Attention_Alpha	54.00	0.7615	0.00
Attention_Ratio	15.00	0.0167	0.54
Attention_Theta	9.00	0.0060	0.75
Attention_Beta	6.00	0.0051	0.54
Engagement_Theta	33.00	0.2031	0.00
Engagement_Alpha2	44.00	0.4673	0.00

Predefined EEG Marker Contrast between tasks and Baseline: To contextualize these task-specific differ-

ences, absolute spectral deviations were evaluated against the resting baseline (eyes open) using Wilcoxon signed-rank tests. A striking asymmetry was observed between the two experimental tasks. The D2-R task induced massive and highly significant spectral modulations compared to the baseline. Notably, the Attention_Beta marker exhibited a 206% median power increase ($p < 0.001$), alongside significant increases in the Attention_Ratio (+46.1%, $p < 0.001$) and Attention_Theta (+34.0%, $p < 0.001$).

Conversely, the 2-Back working memory task did not elicit significant spectral differences relative to the baseline across any of the predefined markers (all $p > 0.05$). The Engagement_Theta and Engagement_Alpha markers, for instance, exhibited negligible median variations (+0.76% and +6.72% respectively, n.s.).

Predefined EEG Marker Correlation with Behavioral Performance: To investigate the relationship between neural markers and task performance, we computed Spearman's rank correlations (ρ) between the EEG features and the behavioral metrics: the CP score for the D2-R and the d' index for the 2-Back. Following False Discovery Rate (FDR) correction, no significant correlations were found across any task (all $p_{FDR} > 0.5$). For instance, the expected correlation between the Attention_Ratio and the D2-R CP score was not observed ($\rho = -0.26$, $p_{FDR} = 0.53$). Similarly, frontal theta modulations did not significantly correlate with the 2-Back d' score ($\rho = 0.25$, $p_{FDR} = 0.53$).

Regional Band Power during tasks (D2-R and 2-Back): To expand our contrast between sustained attention and cognitive engagement to a whole-brain approach, regional band power comparisons were conducted. Given the experimental design involving three tasks, differences were first assessed using a global non-parametric Friedman test, followed by Wilcoxon signed-rank post-hoc comparisons. To control for the false discovery rate across multiple regional and spectral comparisons, all p -values were adjusted using the Benjamini-Hochberg FDR correction.

Tab. 3 presents the FDR-corrected post-hoc p -values specifically isolating the contrast between the D2-R sustained attention task and the 2-Back cognitive engagement task.

Table 3: FDR-corrected p -values for regional EEG band power comparisons (D2-R vs. 2-Back post-hocs).

	Frontal	Central	Parietal	Occipital	Temporal	Global
δ	0.638	0.978	0.001	0.020	0.421	0.160
θ	0.934	0.488	0.012	0.007	0.025	0.106
α	0.302	0.207	0.761	0.978	0.890	0.524
α_1	0.599	0.276	1.000	0.846	1.000	0.846
α_2	0.311	0.359	0.421	0.846	0.678	0.584
β	0.561	0.180	0.095	0.003	0.207	0.032

The analysis reveals distinct spatial modulation patterns characterizing the two cognitive states. No significant

differences were observed in the frontal and central areas, nor across any of the α sub-bands (all $p_{FDR} > 0.05$). Conversely, modulations in the lower frequencies were highly significant and posterior-dominant. The θ frequency band exhibited the most spatially distributed differences, significantly discriminating the two tasks across the parietal ($p_{FDR} = 0.0128$), occipital ($p_{FDR} = 0.0079$), and temporal lobes ($p_{FDR} = 0.0251$). The δ band showed highly significant localized differences in the parietal ($p_{FDR} = 0.0013$) and occipital lobes ($p_{FDR} = 0.0201$). Finally, the β band power was significantly modulated in the occipital lobe ($p_{FDR} = 0.0030$) as well as globally across all regions combined ($p_{FDR} = 0.0323$).

Temporal Dynamics across Tasks: To assess the continuous temporal evolution of the neural markers, Linear Mixed Models (LMM) were fitted to the time-series data of each experimental task. For each marker and task, three LMM were evaluated leading to five possible temporal trajectories: Null (flat), Linear (increasing/decreasing), and Quadratic (U-shape or inverted U-shape). P -values were corrected for multiple comparisons using the False Discovery Rate (FDR) method (Tab 4).

The analysis revealed that temporal dynamics were highly task-dependent. During the ecological Video task, the Attention_Ratio demonstrated a significant inverted U-shape trajectory ($p_{FDR} = 0.0197$), whereas it followed a U-shape dynamic during the D2-R task ($p_{FDR} = 0.0283$) and showed no significant temporal variation during the 2-Back task. A similar task-specific divergence was observed for the Attention_Theta marker, exhibiting a U-shape temporal dynamic during the Video task ($p_{FDR} = 0.0211$) but an inverted U-shape during the D2-R task ($p_{FDR} = 0.0006$).

Significant linear trends were also identified. The power of the Engagement_Alpha marker increased linearly over time exclusively during the Video task ($p_{FDR} = 0.0217$). Conversely, a significant linear increase in Attention_Beta power was strictly isolated to the 2-Back task ($p_{FDR} = 0.0053$).

DISCUSSION

Spectral Differences Between Tasks: Direct comparison between the D2-R and 2-Back tasks confirms the sensitivity of specific spectral markers to sustained attention. Significant modulations in the Attention_Ratio, Attention_Theta, and Attention_Beta markers align with literature linking sustained vigilance to posterior low-frequency dynamics and occipital beta power [2, 9].

Conversely, the predefined engagement markers (Engagement_Theta and Engagement_Alpha) failed to discriminate the tasks. Two hypotheses explain this: the 2-Back task may not have elicited a significantly higher cognitive load than the speed-constrained D2-R for university students, and median PSD aggregation

Table 4: LMM selection for temporal dynamics across tasks (PFDR).

Marker	Task	Best Model	PFDR
Attention_Alpha	Video	Quad. (U)	0.1125
	D2-R	Quad. (Inv-U)	0.0651
	2-Back	Quad. (Inv-U)	0.0196
Attention_Ratio	Video	Quad. (Inv-U)	0.0197
	D2-R	Quad. (U)	0.0283
	2-Back	Null (Flat)	1.0000
Attention_Theta	Video	Quad. (U)	0.0211
	D2-R	Quad. (Inv-U)	0.0006
	2-Back	Null (Flat)	1.0000
Attention_Beta	Video	Quad. (Inv-U)	0.0584
	D2-R	Quad. (Inv-U)	0.0393
	2-Back	Linear (Inc.)	0.0053
Engagement_Theta	Video	Quad. (U)	0.0594
	D2-R	Quad. (Inv-U)	0.0138
	2-Back	Quad. (Inv-U)	0.0263
Engagement_Alpha	Video	Linear (Inc.)	0.0217
	D2-R	Quad. (Inv-U)	0.1125
	2-Back	Quad. (Inv-U)	0.0263

likely smoothed out the transient, phase-locked bursts characteristic of Frontal Midline Theta (FMT) [3].

Spatio-Spectral Signatures of Cognitive States: Regional band power analysis (Tab. 3) provides a granular view of the networks distinguishing these tasks, revealing a strictly posterior distribution of low-frequency modulations. Significant differences in δ and θ bands across parietal and occipital lobes align with established neural signatures of visual vigilance decline [2]. Crucially, the absence of α band differences across all regions is informative. Literature reports parieto-occipital α increases during disengagement for both states [7, 11]. Our results corroborate this shared mechanism: α networks were likely equally suppressed by the external focus required in both tasks, rendering α ineffective for discriminating between them.

Baseline and Order Effects: The fixed sequential design (D2-R $\rightarrow$ 2-Back $\rightarrow$ Baseline $\rightarrow$ Video) introduced a critical confounding factor. The D2-R task was performed by “fresh” participants, yielding massive spectral deviations (e.g., +206% in Attention_Beta) compared to a post-effort baseline. Conversely, the 2-Back spectral signature was compared to an immediately subsequent baseline. This “resting” state was actually a post-effort recovery phase, likely saturated with residual fatigue and engagement markers, thereby masking statistical differences for the 2-Back task. Future protocols must employ true pre-task baselines or randomized orders to isolate state-specific deviations.

Temporal Dynamics in Ecological Settings: Despite baseline limitations, Linear Mixed Models (LMM) applied to continuous data revealed robust, task-dependent temporal dynamics (Tab. 4), highlighting the critical loss of information when temporal data is globally averaged. During the 27-minute Video task, Attention_Ratio followed an inverted U-shape trajectory, while Attention_Theta exhibited a U-shape.

This suggests a tri-phasic response: initial high-vigilance, mid-task decline (rising θ and dropping ratio), and a final attention regain. Although participants were aware of the 27-minute total duration, the absence of a visible progress bar suggests this late-stage rebound reflects an internal time estimation or spontaneous resource re-allocation, rather than a cue-driven anticipation. Interestingly, Engagement_Alpha2 showed a linear increase exclusively during the Video, indicating gradual disengagement of parietal attentional filters. These trajectories emphasize the necessity of continuous, non-linear tracking for educational BCIs.

Behavioral Correlations and Statistical Power: The lack of significant correlations between predefined EEG markers and behavioral performance underscores a common limitation in BCI studies. Following FDR correction, no robust brain-behavior links remained. This is primarily attributable to the limited sample size ($N = 15$), which drastically reduces statistical power for detecting medium-to-small effect sizes. Furthermore, due to the loss of temporal information, discrete behavioral scores may inadequately estimate online cognitive effort fluctuation (e.g., maintaining performance despite fatigue) during task execution in BCI learning contexts.

CONCLUSION

This study establishes a framework for distinguishing EEG markers of sustained attention from cognitive engagement for real-time BCI monitoring. Sustained attention decline manifested robustly through increased posterior θ and occipital β power, alongside modulations of the $\beta/(\alpha + \theta)$ ratio. Conversely, the absence of significant spectral divergence for the engagement task (2-Back) underscores overlapping neural networks and the susceptibility of markers like Frontal Midline Theta (FMT) to cumulative fatigue. Furthermore, linear mixed models revealed that cognitive states exhibit distinct, non-linear temporal trajectories during continuous ecological learning, demonstrating that aggregating EEG metrics over the entire task duration obscures critical real-time state fluctuations.

Methodologically, marker selection must align with both the task order and the baseline position. Future protocols must enforce strictly randomized task orders and true pre-task baselines to prevent post-effort confounds. As BCI technology transitions from laboratory to real-world applications, research should investigate subject-specific frequency calibration and explore functional connectivity measures (e.g., DMN-DAN anti-correlations) within hybrid machine-learning architectures to enhance adaptive systems. Differentiating these states provides a rigorous foundation for nuanced, theory-driven BCI design in attention-aware educational technologies.

REFERENCES

- [1] Pope AT, Bogart EH, Bartolome DS. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*. 1995;40(1):187–195.
- [2] Ko LW, Komarov O, Hairston WD, Jung TP, Lin CT. Sustained Attention in Real Classroom Settings: An EEG Study. *Front. Hum. Neurosci.*. 2017;11.
- [3] Arnau S, Liegel N, Wascher E. Frontal midline theta power during the cue-target-interval reflects increased cognitive effort in rewarded task-switching. *Cortex*. 2024;180:94–110.
- [4] Wascher E et al. Frontal theta activity reflects distinct aspects of mental fatigue. *Biological Psychology*. 2014;96:57–65.
- [5] Klimesch W. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*. 2012;16(12):606–617.
- [6] Bacigalupo F, Luck SJ. Lateralized Suppression of Alpha-Band EEG Activity As a Mechanism of Target Processing. *J Neurosci*. 2019;39(5):900–917.
- [7] Compton RJ, Gearinger D, Wild H. The wandering mind oscillates: EEG alpha power is enhanced during moments of mind-wandering. *Cogn Affect Behav Neurosci*. 2019;19(5):1184–1191.
- [8] Conrad C, Newman A. Measuring Mind Wandering During Online Lectures Assessed With EEG. *Front. Hum. Neurosci.*. 2021;15.
- [9] Gola M, Magnuski M, Szumska I, Wróbel A. EEG beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects. *International Journal of Psychophysiology*. 2013;89(3):334–341.
- [10] Berka C et al. EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. 2007;78(5).
- [11] Chikhi S, Matton N, Blanchet S. EEG power spectral measures of cognitive workload: A meta-analysis. *Psychophysiology*. 2022;59(6):e14009.
- [12] Jaquess KJ et al. Self-Controlled Practice to Achieve Neuro-Cognitive Engagement: Underlying Brain Processes to Enhance Cognitive-Motor Learning and Performance. *Journal of Motor Behavior*. 2020;52(5):544–557.
- [13] Bellos C, Stefanou K, Tzallas A, Stergios G, Tsipouras M. Methods and Approaches for User Engagement and User Experience Analysis Based on Electroencephalography Recordings: A Systematic Review. *Electronics*. 2025;14(2):251.
- [14] Halderman LK, Finn B, Lockwood J, Long NM, Kahana MJ. EEG Correlates of Engagement During Assessment. *ETS Research Report Series*. 2021;2021(1):1–17.
- [15] Bacigalupo F, Luck SJ. Alpha-band EEG suppression as a neural marker of sustained attentional engagement to conditioned threat stimuli. *Social Cognitive and Affective Neuroscience*. 2022;17(12):1101–1117.
- [16] De Vico Fallani F, Richiardi J, Chavez M, Achard S. Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philos Trans R Soc Lond B Biol Sci*. 2014;369(1653):20130521.
- [17] Liu J, Malesevic N, Antfolk C. Long term sustained attention alters dynamic functional connectivity patterns. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Jul. 2023, 1–4.
- [18] Seshadri NPG, Singh BK, Pachori RB. EEG Based Functional Brain Network Analysis and Classification of Dyslexic Children During Sustained Attention Task. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2023;31:4672–4682.
- [19] Ismail LE, Karwowski W. A Graph Theory-Based Modeling of Functional Brain Connectivity Based on EEG: A Systematic Review in the Context of Neuroergonomics. *IEEE Access*. 2020;8:155103–155135.
- [20] Seeburger DT et al. Time-varying Functional Connectivity Predicts Fluctuations in Sustained Attention in a Serial Tapping Task. *Cogn Affect Behav Neurosci*. 2024;24(1):111–125.
- [21] Rosenberg MD et al. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat Neurosci*. 2016;19(1):165–171.
- [22] Francisco-Vicencio MA, Góngora-Rivera F, Ortiz-Jiménez X, Martínez-Peón D. Sustained attention variation monitoring through EEG effective connectivity. *Biomedical Signal Processing and Control*. 2022;76:103650.
- [23] Huang H, Li R, Zhang J. A review of visual sustained attention: Neural mechanisms and computational models. *PeerJ*. 2023;11:e15351.
- [24] Ozel P. Dynamic Neural Network States During Social and Non-Social Cueing in Virtual Reality Working Memory Tasks: A Leading Eigenvector Dynamics Analysis Approach. *Brain Sciences*. 2025;15(1):4.
- [25] Dini H, Simonetti A, Bruni LE. Exploring the Neural Processes behind Narrative Engagement: An EEG Study. *eNeuro*. 2023;10(7).
- [26] Song H, Finn ES, Rosenberg MD. Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proceedings of the National Academy of Sciences*. 2021;118(33):e2021905118.
- [27] Brickenkamp R, Schmidt-Atzert L, Liepmann D. D2-R : Test d'attention concentrée révisé. Hogrefe France (2014).
- [28] Stoet G. PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*. 2017;44(1):24–31.
- [29] Khosla A, Khandnor P, Chand T. A comparative analysis of signal processing and classification methods for different applications based on EEG signals. *Biocybernetics and Biomedical Engineering*. 2020;40(2):649–690.
- [30] Zhao L, He Y. Power Spectrum Estimation of the Welch Method Based on Imagery EEG. *AMM*. 2013;278-280:1260–1264.

Beyond Visual Feedback: Isolating Vestibular Error-Related Potentials During Wheelchair Operations

F. Virgolini¹, P. Simonetto¹, S. Tortora^{1, 2}, E. Menegatti^{1, 2}, L. Tonin^{1, 2}

¹Department of Information Engineering, University of Padova, Padova, Italy

²Padova Neuroscience Center, University of Padova, Padova, Italy

E-mail: francesca.virgolini@phd.unipd.it

ABSTRACT: The Error-Related Potential (ErrP) is a well-established neurophysiological marker of error perception, historically elicited through visual feedback in Brain-Machine Interfaces (BMIs). However, to our knowledge, no study has yet investigated whether such a response can be elicited in the complete absence of visual input, relying solely on vestibular feedback generated by physical movement. This study addresses this question by recording electroencephalographic (EEG) signals from participants during a powered wheelchair motion task, under two conditions: one with visual feedback available, and one with visual occlusion imposed through blindfolding. Our findings demonstrate that a recognizable ErrP waveform is elicited under full visual occlusion, exhibiting a morphology consistent with that observed in the sighted condition. Observed differences in latency and amplitude between conditions are discussed in relation to their experimental contexts. These results provide preliminary evidence that vestibular feedback alone constitutes a sufficient substrate for ErrP generation, opening a promising avenue for BMI systems not relying on visual input for implicit error monitoring.

INTRODUCTION

Error-related Potentials (ErrPs) are a class of event-related potentials (ERPs) that arise as an endogenous neurophysiological response to the perception of erroneous outcomes [1]. First characterized in the context of response monitoring, these signals are generated whenever a mismatch is detected between an expected and an observed outcome, independently of whether the error was self-generated or externally observed [1].

In the specific context of assistive devices, where the error originates from a misrecognition of the user's intent by the interface, the elicited signal is referred to as interaction ErrP [1]. At the waveform level, this subtype is characterized by a fronto-central waveform comprising a negative peak (ERN/Ne) at around 250 ms post-feedback, followed by a positive peak (Pe) at around 320 ms, and a subsequent negative peak at around 450 ms [1][2]. The consistency and robustness of this morphology across subjects and paradigms have positioned the ErrP as a particularly tractable target for single-trial decoding in brain-machine interface (BMI) research [3].

Early investigations into the functional properties of ErrPs were conducted almost exclusively within visual paradigms, in which participants monitored screen-rendered interfaces and evaluated the correctness of cursor trajectories, agent movements, or graphical decision outputs [4][5][6]. While these studies established the decodability of the signal under controlled laboratory conditions, they were structurally incapable of dissociating the role of visual processing from other potential contributors to error representation. Subsequent research extended the applicability of ErrPs to more naturalistic and physically grounded settings: reliable ErrP generation was demonstrated during the observation of robotic arm trajectories, and the signal was shown to be sufficiently stable for integration into closed-loop reinforcement learning architectures, enabling autonomous agents to update their control policies in response to implicitly decoded human evaluative feedback [7][8]. Across all of these paradigms, however, visual feedback remained the primary—and often exclusive—sensory channel through which error information was delivered to the participant. A distinct but related body of work has examined the utility of ErrPs in assistive technology (AT) applications, with a particular focus on physical human-robot interaction. Notably, it was demonstrated that ErrP responses are preserved even when the interface between user and robot involves tactile rather than purely visual coupling, as in the case of exoskeleton-based paradigms [9], providing evidence that non-visual afferent feedback can independently contribute to error-related cortical activity. Other findings enabled the construction of multi-component BMI systems in which ErrP-based error monitoring operates in concert with motor imagery-based command generation, forming a complete supervisory loop for rehabilitation robotics [10][11]. Nevertheless, a fundamental limitation persisted across this literature: in the overwhelming majority of studies, participants retained full visual access to the task environment, rendering it impossible to determine the extent to which the observed ErrP reflected visual error processing as opposed to proprioceptive, somatosensory or vestibular feedback. The application of ErrP-based BMIs to powered wheelchair navigation has received comparatively limited attention. An early investigation in this domain employed a teleoperation paradigm, in which the user re-

mained physically decoupled from the robot and evaluated its behavior through a visual display [12]—a configuration that, despite its applied motivation, did not substantively depart from the visual interface paradigms that preceded it. A more recent study proposed a complete ErrP-driven navigation pipeline; however, its validation was restricted to simulated environments and provided no evidence regarding the integrity of the signal during real wheelchair locomotion [13]. In neither case was it possible to assess whether ErrP generation would be sustained in the absence of visual feedback, or whether the physical dynamics of wheelchair movement could themselves serve as a sufficient substrate for error elicitation.

The present study was designed to address these open questions directly. Participants were asked to perform a wheelchair navigation task under two conditions: one in which visual feedback was available, and one in which full visual occlusion was imposed through blindfolding. This design allowed, for the first time, the recording of ErrPs from participants physically situated within a moving powered wheelchair, while simultaneously enabling a systematic dissociation of visual and vestibular contributions to the error-related response. Any error-related neural activity recorded under the blindfolded condition could thus be attributed unambiguously to non-visual, movement-derived afferent feedback, providing the first characterization of an ErrP elicited exclusively through vestibular channels during active self-motion in a real navigation context.

Beyond resolving this methodological ambiguity, the present investigation carries broader implications for the design of BMI systems intended for naturalistic use. If error-related cortical activity can be elicited independently of visual input — on the sole basis of movement-derived afferent feedback — then error-correction paradigms need not be restricted to interfaces providing explicit visual feedback. Such a finding would suggest that vestibular signals may constitute a sufficient, or at minimum complementary, substrate for implicit error monitoring, substantially expanding the range of contexts in which ErrP-based BMIs could be deployed.

MATERIALS AND METHODS

Participants: Five healthy male volunteers (mean age 27.8 years, range 26-31) took part in the experiment, one of whom had prior experience with ErrP-based paradigms. Informed consent was obtained from all participants prior to the study, which was conducted in compliance with the ethics committee "Comitato etico della ricerca psicologica", University of Padova (1575-b).

Wheelchair setup: A commercial powered wheelchair (18.68E HI-LO Electric, Vassilli s.r.l., Padua, Italy) was modified to meet the requirements of this study. Wheelchair kinematics were tracked by two Kübler KIS40 encoders attached to the front wheels via a belt, and odometry was derived using the method in [14]. User interaction was mediated by a Logitech F710 joystick, while a screen mounted in front of the participant pro-

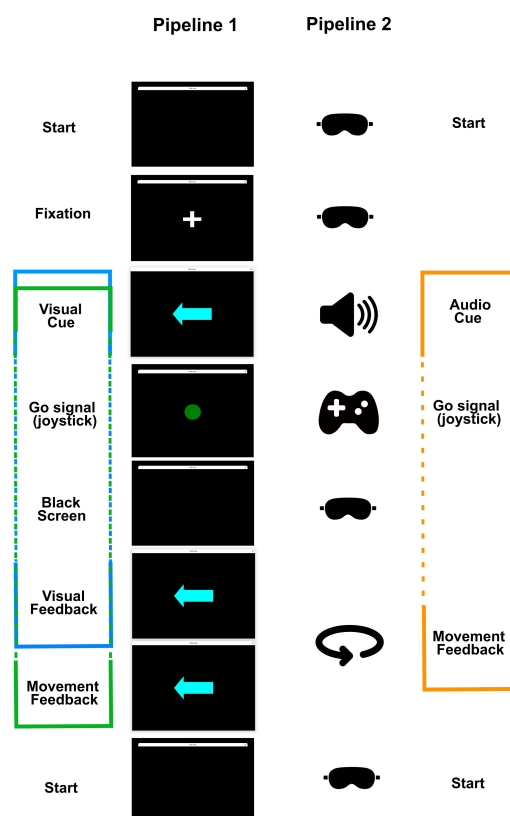


Figure 1: Schematic representation of the two experimental pipelines employed for ErrP elicitation. Pipeline 1 (left) follows a fully visual paradigm, in which each depicted symbol on the display marks a specific event. Pipeline 2 (right) preserves the same temporal structure but replaces visual stimuli with audio cues, with participants remaining blindfolded throughout. Colored boxes highlight the three experimental contrasts of interest: the blue dashed box isolates the visual ErrP epoch, when the mismatch is between the visual cue and the visual feedback; the green solid box captures vestibular ErrPs in the presence of vision, when the mismatch is between the visual cue originally given and the rotational feedback; and the orange dashed box identifies the purely vestibular ErrP under blindfolded conditions, when the mismatch is between the audio cue and the rotational feedback.

vided the visual feedback integrated into the experimental pipeline. The central computing unit was a Lenovo ThinkPad T15g Gen 2 laptop (11th Gen Intel Core i9-11950H CPU, NVIDIA GeForce RTX 3080 GPU, 32 GB RAM), to which all components were connected. An R-Net chipset (Curtiss-Wright) bridged the wheelchair's CAN-bus to the laptop via a USB isolator, minimizing electrical interference and ground-loop artifacts induced by the motor system. The software architecture relied on ROS Noetic (v1.17.3) as the communication backbone, with ROS-Neuro middleware [15] ensuring synchronization between the BMI, GUI, and robotic components. Post-hoc data analysis was conducted on the same platform using MATLAB 2024a (MathWorks, Natick, MA, USA).

Signal acquisition: The EEG acquisition system consisted of a 64-channel amplifier (AntNeuro eego 64, The Netherlands), with the following 32 channels retained for

analysis: AF3, AF4, F3, F1, FZ, F2, F4, FC3, FC1, FCZ, FC2, FC4, C3, C1, CZ, C2, C4, CP3, CP1, CPZ, CP2, CP4, P3, P1, PZ, P2, P4, PO3, POZ, PO4, O1, O2. The reference electrode was placed at Iz and the ground electrode at AFz. Ocular artifacts were monitored via electrooculography (EOG), recorded through an external bipolar box connected by adapter cable to adhesive pads positioned around the right eye. Both EEG and EOG signals were sampled at 512 Hz and filtered using a 4th-order zero-phase Butterworth bandpass filter with a passband of 2–10 Hz. Spatial resolution of the EEG was further enhanced using a small Laplacian spatial filter computed from the four nearest neighboring electrodes for each channel.

Experimental setup: To isolate the vestibular error potential and establish a robust basis for comparison, data were collected across two distinct experimental pipelines (Fig. 1). In both conditions, each participant was seated in the powered wheelchair and instructed to follow a pre-defined sequence of cues.

In the first pipeline, each trial consisted of two successive phases: a visual-feedback phase and a movement-feedback phase. Trials began with a black screen (1.5 s), followed by a fixation cross (1.5 s) and a directional cue (1.5 s). Upon the appearance of a green dot, the participant provided input via the joystick. Following an inter-stimulus interval (2.0 s), a visual feedback arrow (2.0 s) indicated the direction of the upcoming movement. The wheelchair remained stationary throughout these steps. After a delay of 1.5 s from the onset of the visual cue, the wheelchair performed an in-place rotation at a constant angular velocity for 2.0 s, constituting the movement feedback consistent with the previously displayed arrow. Each trial concluded with a final black screen (1.5 s). ErrPs were elicited by introducing unexpected events with a probability of 30%: in these trials, both the visual feedback arrow and the subsequent physical rotation were presented in the direction opposite to the participant's joystick input. In the remaining 70% of trials, the system's response was congruent with the participant's intended command.

In the second pipeline, designed to isolate the purely vestibular error potential, participants were blindfolded to eliminate any visual information. The temporal structure of the first pipeline was preserved to allow direct comparison; however, directional instructions were delivered via audio cues reproducing the words 'Left' or 'Right'. In the absence of visual feedback, participants were able to respond with the corresponding joystick command immediately after the audio cue. After a delay of 1.5 s from joystick input, the wheelchair executed an in-place rotation at a constant angular velocity for 2.0 s, at a lower speed than that employed in the first pipeline, as a precautionary measure to minimize vestibular discomfort in participants during the blindfolded condition. The direction of the rotational feedback was incongruent with the audio cue with a probability between 30% and 40%.

To ensure consistency across participants, a full writ-

ten description of the task was provided before each session for both pipelines. Each participant completed ten runs per pipeline, each consisting of ten consecutive trials. Prior to the recording phase, a brief familiarization period with the wheelchair was provided to ensure comfort and procedural understanding. All sessions were conducted with motors continuously active, a design choice motivated by the need to evaluate EEG signal quality under realistic electromagnetic conditions and to avoid artifacts associated with motor switching transients. To further ensure reproducibility, participants were instructed to limit head movements and maintain a consistent posture throughout each trial. Trial rejection was performed by visual inspection: only trials exhibiting the expected EEG frequency spectrum were retained, while those showing evident noise in the EOG channel or EEG amplitudes exceeding 35 μ V were discarded (12.9% of total trials).

RESULTS

The grand-average ErrPs recorded at FCz are shown in Figure 2 for three experimental conditions: a) Visual ErrP, b) Vestibular ErrP in the presence of vision and c) Vestibular ErrP in the absence of vision. Each subplot displays the difference waveform obtained by subtracting the mean correct-trial signal from the mean error-trial signal, averaged across participants; the shaded area represents ± 1 standard error of the mean (SEM). In all three conditions, the ErrP morphology appears clearly identifiable: a prominent negative deflection (Ne/ERN-like component) followed by a positive peak (Pe-like component), consistent with the canonical error-related potential described in the literature. The topographic maps computed within a ± 30 ms window centered on each relevant peak suggest a fronto-central scalp distribution focused on FCz and Cz, in agreement with the expected neural generators of the ErrP [1].

Condition 1 – Visual ErrP: The first row shows the potential evoked by a visual mismatch, in which the visual cue and the visual feedback were incongruent. The negative peak is sharp and well-defined and occurs at a relatively short latency (mean latency = 0.302s) with respect to feedback onset ($t = 0$). The topographic maps at the peaks reveal a highly focal activation restricted to the FCz–Cz region.

Condition 2 – Vestibular ErrP in the presence of vision: The second row depicts the response to a mismatch between the visual cue and the wheelchair rotation feedback, delivered while participants retained full visual access. The characteristic negative-positive complex is preserved. A slight increase in peak latency (mean negative peak latency = 0.446s) relative to Condition 1 is observed, which may reflect the additional processing time required to perceive the direction of rotation. The topographic distribution at the peaks shows both some focal activity over FCz–Cz and also a broader activation. The dotted and dashed traces overlaid on the subplot represent, respectively, the delta in angular velocity with

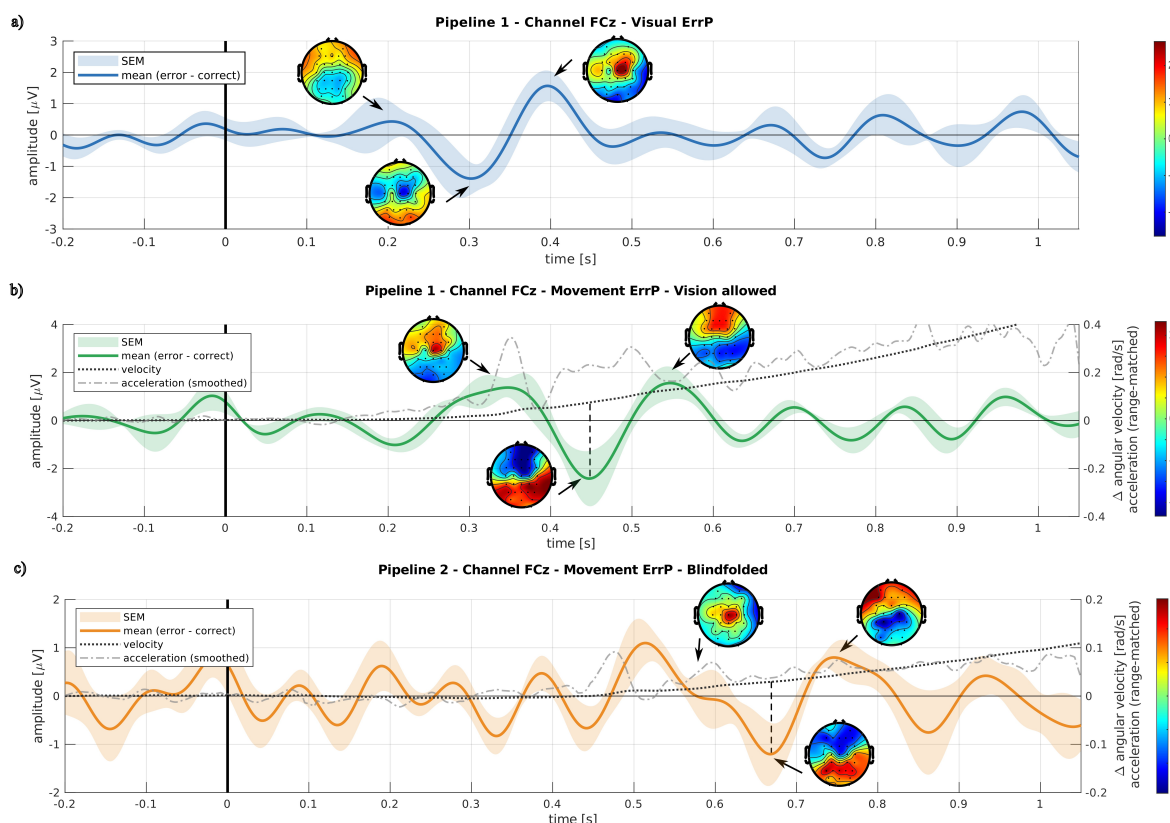


Figure 2: Grand-average ErrP difference waveforms (error minus correct) recorded at FCz across the three experimental conditions, averaged across participants (shaded area: ± 1 SEM). **a)** *Condition 1*: visually-evoked ErrP, time-locked to the onset of the visual feedback stimulus. **b)** *Condition 2*: vestibular ErrP recorded in the presence of vision, time-locked to wheelchair rotation onset; **c)** *Condition 3*: vestibular ErrP recorded under full visual occlusion, time-locked to wheelchair rotation onset. Note that amplitude scales differ across panels to enhance morphological comparability. For each condition, topographic maps computed within a ± 30 ms window centered on the most prominent peaks illustrate the fronto-central scalp distribution of the response. For the second and third condition, the dotted trace represents the delta in angular velocity of the wheelchair with respect to its rest state, while the dashed trace represents the corresponding angular acceleration, scaled to match the velocity axis range to facilitate visual comparison of peak timing. The vertical line at instant 0 represents the feedback onset, visual or rotational. Vertical dashed lines show the velocity value corresponding to the negative peak of ErrP.

respect to its rest state and the angular acceleration of the wheelchair, scaled to match the velocity axis range to facilitate visual comparison of peak timing. A vertical dashed line connects the ErrP negative peak to the corresponding velocity value. Inspection of these kinematic signals suggests that the positive peak preceding the negative ErrP peak occurs in temporal proximity to, or slightly after, the peak in angular acceleration, raising the possibility that the transient acceleration event may constitute the primary vestibular cue through which the direction of the erroneous movement is first detected.

Condition 3 – Vestibular ErrP in the absence of vision: The third row presents the ErrP elicited under purely vestibular conditions: participants were blindfolded and the mismatch occurred between an auditory cue and the wheelchair movement. The negative deflection is present but exhibits reduced amplitude and a longer latency compared to the two preceding conditions (mean negative peak latency = 0.667s). The SEM band is slightly wider, suggesting greater inter-subject variability. Neverthe-

less, the first topographic map still displays a fronto-central focus qualitatively comparable to that of Condition 1, which may support the presence of a genuine error-related signal. The second and third topoplots show a more diffuse distribution, broadly similar to that observed in Condition 2, possibly suggesting a shared neural substrate across the two motor conditions regardless of visual availability. A qualitatively similar temporal relationship between the positive peak preceding ErrP negativity and the angular acceleration peak is observed under blindfolded conditions, suggesting that this kinematic feature may retain its role as a salient afferent trigger for error detection even in the absence of visual input.

DISCUSSION

A central motivation for investigating vestibular ErrPs in assistive navigation contexts stems from the practical limitations of display-based feedback paradigms in such scenarios. In real-world wheelchair use, sustained visual attention to a monitor is neither practical nor desirable,

as dynamic environments demand that the user's gaze be directed toward the surroundings. The possibility of eliciting an ErrP on the basis of the physical movement itself, with feedback delivered through vestibular channels rather than through a visual display, would therefore eliminate this gaze-allocation constraint and constitute a meaningful step toward more ecologically valid and natural BMI-driven navigation.

The present results show that an error-related potential can be elicited across three distinct sensory conditions — visual, visuo-vestibular, and purely vestibular — in the context of wheelchair-based BCI control. The consistent ERN–Pe morphology and the fronto-central topography observed across all conditions are in line with the established ErrP literature, and are compatible with the interpretation that the recorded signals reflect error monitoring activity involving the anterior cingulate cortex and supplementary motor area [1].

However, some considerations need to be made regarding the amplitude differences between the three conditions, since Conditions 1 and 2 yielded larger ErrP amplitudes than Condition 3. This pattern may be attributable to at least two concurrent factors. First, the error stimulus probability in Conditions 1 and 2 was fixed at 30%, whereas in Condition 3 it ranged between 30% and 40% depending on the run. Given the well-established inverse relationship between error probability and ErrP amplitude, the slightly elevated error rate in Condition 3 may have contributed to the observed amplitude reduction. Second, it has been reported that ErrP amplitude may co-vary with the perceived salience or urgency associated with an erroneous movement, which may be approximated by the velocity of the unintended motion [16]. The lower wheelchair rotation speed employed in Condition 3 relative to Condition 2 would predict a smaller error signal under this account. These two explanations are not mutually exclusive and may have acted in concert to attenuate the vestibular ErrP; future work could focus on disentangling their respective contributions.

Moreover, blindfolding in Condition 3 may have introduced alpha-band contamination, as eye closure is known to increase posterior alpha power [17], potentially contributing to amplitude differences between Conditions 2 and 3 beyond those attributable to the sensory manipulation itself. However, the available evidence argues against alpha contamination as the primary explanation for the observed ErrP. In fact, the 2–10 Hz bandpass filter partially attenuates the alpha band, and more importantly, the fronto-central topography and time-locked ERN–Pe morphology of the signal are inconsistent with the posterior, sustained fluctuations characteristic of alpha activity [17]. Taken together, these observations support the interpretation that the signal in Condition 3 reflects genuine error-related activity, even if alpha modulation may have contributed to the increased inter-subject variability observed in that condition.

The latency differences across conditions also deserve consideration. The visual ErrP in Condition 1, being

time-locked to the discrete onset of a visual stimulus, would be expected to exhibit the shortest and most consistent latency. The vestibular ErrPs in Conditions 2 and 3, by contrast, are contingent on the participant's ability to perceive the direction of wheelchair rotation — a process that does not begin instantaneously at command delivery but rather unfolds over a temporal window shaped by the dynamics of the rotation itself. The longer latency observed in the motor conditions is therefore consistent with the additional sensorimotor processing required to infer movement direction from vestibular or combined visuo-vestibular signals. The further latency prolongation in Condition 3 may suggest that the absence of visual information and the reduced rotational speed jointly delay the detection of the erroneous movement direction, thereby shifting the ErrP peak and increasing inter-subject variability — the latter also reflected in the wider SEM band.

A further kinematic observation is consistent with this latency interpretation. Visual inspection of the grand-average waveforms suggests that, in both motor conditions, the positive peak preceding the negative ErrP peak coincides with, or follows slightly after, the peak in wheelchair angular acceleration. This pattern is compatible with the hypothesis that the transient acceleration event, rather than sustained rotational velocity *per se*, may represent the critical afferent trigger for error detection in this paradigm. This interpretation aligns with known properties of vestibular and somatosensory encoding, in which acceleration-sensitive afferents respond preferentially to changes in motion state rather than to sustained movement [18]. Whether acceleration magnitude modulates ErrP amplitude in a manner analogous to the velocity effects discussed above remains an open question that warrants systematic investigation in future work.

The topographic similarity between the second and third topoplots of Conditions 2 and 3 is particularly noteworthy. Despite the fundamentally different sensory context, the spatial pattern of cortical activation appears strikingly comparable across the two conditions, which may imply that a common neural mechanism is engaged when monitoring errors in wheelchair rotation, irrespective of whether this monitoring is supported by concurrent visual input. This observation is compatible with the hypothesis that the signal recorded in Condition 3 belongs to the same functional class as the ErrP in Condition 2, rather than reflecting an unrelated cognitive process; nonetheless, further evidence from single-trial analyses and larger samples would be needed to confirm this interpretation.

Taken together, these preliminary findings suggest the feasibility of eliciting a vestibular ErrP in a real-world assistive device scenario, and point to movement velocity and acceleration and sensory modality as potential factors shaping its amplitude and latency. Notably, the present results provide evidence that a recognizable ErrP waveform can be elicited under purely vestibular condi-

tions, indicating that error-related signal generation is not contingent upon the availability of visual feedback, but instead reflects a more general error-monitoring mechanism capable of operating across different sensory modalities. As a consequence, visual and vestibular error signals may represent independent contributions to a common monitoring process, which could in principle be engaged either in isolation or in combination depending on the sensory context. Future work should systematically manipulate rotational speed in controlled designs to establish a more quantitative characterization of the velocity-amplitude and acceleration-amplitude relationships, and investigate whether the signal parameters observed here are compatible with reliable single-trial classification for closed-loop BCI error correction.

CONCLUSION

The present study, by identifying an ErrP under purely vestibular conditions, provides evidence that error-related neural activity is not exclusive to visually-driven paradigms, but may instead reflect an error-monitoring mechanism that operates across sensory modalities. From an applied perspective, these findings suggest that BMI error-correction paradigms need not be restricted to interfaces providing explicit visual feedback, opening the way toward more naturalistic, attention-preserving systems in which movement-contingent vestibular signals serve as a sufficient, or complementary, substrate for implicit error monitoring in real-world wheelchair navigation.

ACKNOWLEDGMENTS

This work was carried out within the Piano Nazionale di Ripresa e Resilienza (PNRR) research activities, funded by the European Union – NextGenerationEU. Support was provided through: (i) Missione 4, Componente 2, Investimento 1.5 (D.D. 1058, 23/06/2022, ECS_00000043), within the iNEST (Interconnected North-East Innovation Ecosystem) consortium; and (ii) PE8 – Project Age-It: "Aging Well in an Aging Society" (DM 1557, 11/10/2022). The views expressed are solely those of the authors; the European Union and the European Commission bear no responsibility for them.

REFERENCES

- [1] Ferrez PW, Millan JDR. Error-related eeg potentials generated during simulated brain-computer interaction. *IEEE Transactions on Biomedical Engineering*. 2008;55(3):923–929.
- [2] Falkenstein M, Hohnsbein J, Hoormann J, Blanke L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*. 1991;78(6):447–455.
- [3] Chavarriaga R, Sobolewski A, Millán JdR. Error-related potentials: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*. 2014;8.
- [4] Lopes Dias C, Sburlea AI, Müller-Putz GR. Masked and unmasked error-related potentials during continuous control and feedback. *Journal of Neural Engineering*. 2018;15(3):036031.
- [5] Milekovic T, Ball T, Schulze-Bonhage A, Aertsen A, Mehring C. Detection of error related neuronal responses recorded by electrocorticography in humans during continuous movements. *PLOS ONE*. 2013;8(2):e55235.
- [6] Spüler M, Niethammer C. Error-related potentials during continuous feedback: Using eeg to detect errors of different type and severity. *Frontiers in Human Neuroscience*. 2015;9.
- [7] Kim SK, Kirchner EA, Stefes A, Kirchner F. Intrinsic interactive reinforcement learning - using error-related potentials for real world human-robot interaction. *Scientific Reports*. 2017;7(1):17562.
- [8] Akinola I et al. Accelerated robot learning via human brain signals. *arXiv preprint*. 2019.
- [9] Kim SK, Kirchner EA. Detection of tactile-based error-related potentials (errps) in human-robot interaction. *Frontiers in Neurobotics*. 2023;17.
- [10] Bhattacharyya S, Konar A, Tibarewala DN. Motor imagery and error related potential induced position control of a robotic arm. *IEEE/CAA Journal of Automatica Sinica*. 2017;4(4):639–650.
- [11] Wang X, Chen HT, Wang YK, Lin CT. Implicit robot control using error-related potential-based brain-computer interface. *IEEE Transactions on Cognitive and Developmental Systems*. 2023;15(1):198–209.
- [12] Perrin X, Chavarriaga R, Colas F, Siegwart R, Millán J del. Brain-coupled interaction for semi-autonomous navigation of an assistive robot. *Robotics and Autonomous Systems*. 2010;58(12):1246–1255.
- [13] Ferracuti F, Freddi A, Iarlori S, Longhi S, Monteriù A, Porcaro C. Augmenting robot intelligence via EEG signals to avoid trajectory planning mistakes of a smart wheelchair. *Journal of Ambient Intelligence and Humanized Computing*. 2023;14(1):223–235.
- [14] Xu R. Path planning of mobile robot based on multi-sensor information fusion. *EURASIP Journal on Wireless Communications and Networking*. 2019;2019.
- [15] Tonin L, Beraldo G, Tortora S, Menegatti E. ROS-neuro: An open-source platform for neurobotics. *Frontiers in Neurobotics*. 2022;16:886050.
- [16] Omedes J, Schwarz A, Müller-Putz GR, Montesano L. Factors that affect error potentials during a grasping task: Toward a hybrid natural movement decoding BCI. *Journal of Neural Engineering*. 2018;15(4).
- [17] Hohaia W, Saurels BW, Johnston A, Yarrow K, Arnold DH. Occipital alpha-band brain waves when the eyes are closed are shaped by ongoing visual processes. *Scientific Reports*. 2022;12(1):1194.
- [18] Rabbitt RD, Damiano ER, Grant JW. Biomechanics of the semicircular canals and otolith organs. In: *The Vestibular System*. Springer New York, 2004, 153–201.

Expertise-Dependent Effects of Neurofeedback Design During Motor Imagery

M.Pachoud¹, M.Afankous², E.Segas², M.Izac^{1,3}, B.N'Kaoua^{1,3}, C.Jeunet^{1,2}

¹ Univ. Bordeaux, CNRS, EPHE, INCIA, UMR5287 F-33000 Bordeaux

² Univ. Bordeaux, SMART, F-33000 Bordeaux

³ Univ. Bordeaux, INSERM, BPH Research Center, UMR 1219, F33076 Bordeaux

E-mail: camille.jeunet@u-bordeaux.fr

ABSTRACT

Motor imagery (MI) combined with neurofeedback (NF) is a promising approach to enhance sport performance, yet its effectiveness remains heterogeneous. Athletic expertise emerges as a primary candidate for this heterogeneity, potentially determining how athletes engage with different feedback designs to modulate sensorimotor rhythms (SMRs). Whether explanatory feedback (neural processes) and evaluative feedback (performance magnitude) differentially benefit athletes as a function of expertise remains unknown. We examined these designs across novice, intermediate, and expert basketball athletes imagining free throws across four feedback conditions: no feedback (None), an evaluative target (Evaluative), an explanatory topographic map (Explanatory), or both (Both). While intermediates and experts improved when feedback was present relative to None, novices showed no benefit from feedback presence alone. Crucially, gaze-performance coupling revealed a double dissociation: novice success increased with fixation on explanatory information, whereas expert success increased with fixation on evaluative information. These findings support the implementation of adaptive feedback designs to optimize MI-NF training.

INTRODUCTION

MI, the mental simulation of movement without overt execution, is widely used to enhance sport performance [1]. By engaging neural networks that substantially overlap with those recruited during actual movement, MI can support skill refinement and motor learning [2,3]. However, MI outcomes are heterogeneous, and expertise is a major contributor to this variability, shaping both behavioural effects and underlying neural recruitment [4,5].

A central distinction concerns MI modality. Visual imagery emphasizes the visual and spatial characteristics of the action from an internal or external perspective [6], whereas kinaesthetic imagery focuses on somatic and proprioceptive sensations, the “feel” of movement [7]. These modalities recruit partially distinct neural substrates: visual MI preferentially engages occipito-parietal regions, whereas kinaesthetic MI more strongly recruits sensorimotor networks [1]. Expertise also affects MI quality: experts typically show greater

temporal congruence between imagined and executed movements, more accurate kinaesthetic representations, and more efficient neural activation patterns consistent with refined internal models [4,5,8–10].

Real-time NF may help address MI variability by providing immediate information about ongoing brain activity [11,12]. In MI settings, electroencephalography (EEG)-based NF can monitor imagery-related signatures such as event-related desynchronisation (ERD) in sensorimotor rhythms (SMR; 12–15 Hz) over motor cortices [13,14]. While NF has shown potential to improve sport performance, its effectiveness depends strongly on methodological rigor and protocol/interface design quality, including compliance with CRED-NF recommendations [15–17]. This suggests that outcomes reflect not only individual traits, but also how feedback information is communicated, interpreted, and used. In motor learning, a well-established distinction separates explanatory feedback (information about the processes driving that performance), from evaluative feedback (information about the magnitude of performance) [18,22]. These two feedback types differ not only in content but in the cognitive demands they place on the learner.

In the context of MI-NF, this distinction maps naturally onto explanatory topographic feedback, which communicates the spatial pattern of neural activity, and evaluative intensity-based feedback, which provides information on SMR modulation magnitude. While their relative utility as a function of expertise remains underexplored in MI-NF contexts, evidence from motor learning suggests that the relevance of each feedback type may shift with skill level [18,22]. Here, we hypothesised that novices would benefit primarily from explanatory feedback as a scaffold for building an internal representation of the target neural state, whereas experts, already possessing stable sensorimotor representations, would benefit most from evaluative feedback as a precise signal to monitor and calibrate their ongoing SMR modulation. To test these hypotheses, novice, intermediate, and expert basketball athletes performed MI of free throws under four feedback conditions: None (no feedback), Explanatory (a simplified topographic map visualising sensorimotor and occipital activity), Evaluative (a unidimensional dynamic target reflecting SMR modulation magnitude), and Both (combined). To dissociate feedback presence effects from usage effects, gaze behaviour was recorded

via eye tracking and integrated into a hybrid generalised linear mixed model (GLMM), enabling trial-by-trial success to be modelled as a function of both feedback condition and fixation duration.

MATERIALS AND METHODS

Participants: Twenty-four right-handed athletes were recruited, though one was excluded due to excessive noise in the recording. The final sample comprised 23 athletes (13 females, 10 males) naive to NF protocols. Participants were categorised into three competitive levels: Novices (n=8; 23.1 ± 2.0 years), Intermediates (n=8; 26.9 ± 7.4 years), and Experts (n=7; 24.1 ± 3.7 years). One participant was excluded due to movement artifacts. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and received a favorable opinion from a national ethics committee (Comité de Protection des Personnes, CPP 2022-A00626-37). All participants signed an informed consent form.

Procedure: Following a familiarisation phase involving the physical execution of basketball free throws and a transition to motor imagery (MI), a 2-minute eyes-open resting EEG baseline was recorded. Participants then completed 8 NF runs, counterbalanced across four feedback conditions (2 runs per condition): None (no feedback), Evaluative (target size proportional to SMR-ERD amplitude), Explanatory (cortical mapping of SMR and occipital activity), and Both. Each run consisted of 10 MI trials lasting 10 seconds each (20 trials in total per condition), separated by a 6–7 second inter-trial interval. The entire experimental protocol, encompassing both the baseline recording and the MI-NF runs, was administered using the Neurathletics software (Neurathletics, France).

Data Acquisition and Processing: EEG activity was recorded at 250 Hz using an 8-electrode Unicorn Hybrid Black system (mastoid reference)(gTec, Austria). Signals were processed in real-time via OpenViBE (1–100 Hz bandpass, 50 Hz notch, artifact rejection $> \pm 100 \mu V$). Region of Interest (RoI) power was extracted for the SMR rhythm (12–15 Hz; C3, Cz, C4) and occipital activity (8–30 Hz; PO7, Oz, PO8). Channel-wise power was estimated using $P(t) = \log(1 + x(t)^2)$. Data were smoothed (1s sliding windows, 0.25s step) and expressed as a temporal percentage variation relative to the baseline to drive the NF interface. Gaze behaviour was tracked using Tobii Pro Glasses 3 (Tobii AB, Sweden) extracting absolute fixation durations on Areas of Interest (AOIs).

Performance Metrics and Statistical Analyses: Trial success was defined as a mean decrease in SMR power relative to baseline ($ERD < 0\%$). Occipital activity served a purely informative function. The probability of success for each trial was modeled using a binomial Generalised Linear Mixed Model (GLMM) to assess the impact of visual configurations and their actual usage: $\text{logit}(P(\text{success})) = \beta_0 + (\text{Level} * \text{Evaluative} * \text{Explanatory}) + (\text{Level} * \text{FixEvaluative}) + (\text{Level} * \text{FixExplanatory}) + (1 | \text{Subject})$.

Explanatory) + (Level * FixEvaluative) + (Level * FixExplanatory) + (1 | Subject).

Fixation durations were nested-median-centered to isolate the usage effect exclusively when feedback was present. Post-hoc analyses (Wald z-tests, Holm correction) evaluated the impact of feedback presence, expertise level differences, and fixation trends (emtrends) on trial probability of success.

RESULTS

Descriptive Overview of Behavioural Performance and Individual Variability Fig. 1 and Tab. 1 summarise the participant score (number of successful trials out of 20 per condition). Collapsing across expertise levels, mean scores were 8.2 ± 5.1 in None, 10.3 ± 5.9 in Evaluative, 10.7 ± 6.3 in Explanatory, and 9.8 ± 5.5 in Both, with substantial between-participant variability across feedback conditions and within each expertise group.

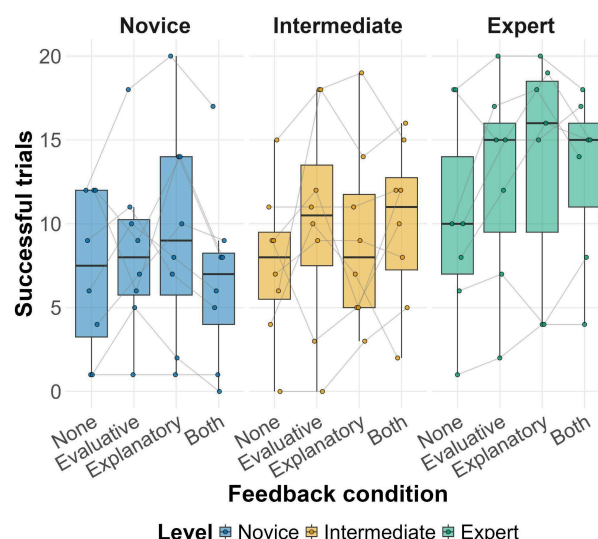


Figure 1: Participant scores across neurofeedback conditions and expertise level.

Table 1: Mean participant scores across neurofeedback conditions and expertise level (Mean $\pm$ SD).

Feedback	Novice	Intermediate	Expert	All levels
None	7.1 $\pm$ 4.8	7.6 $\pm$ 4.5	10.1 $\pm$ 6.2	8.2 $\pm$ 5.1
Evaluative	8.4 $\pm$ 5.0	10.1 $\pm$ 6.4	12.6 $\pm$ 6.2	10.3 $\pm$ 5.9
Explanatory	9.5 $\pm$ 6.4	9.1 $\pm$ 5.4	13.7 $\pm$ 6.8	10.7 $\pm$ 6.3
Both	6.8 $\pm$ 5.3	10.0 $\pm$ 4.8	13.0 $\pm$ 5.1	9.8 $\pm$ 5.5

When examined by expertise level, scores in None condition were 7.1 ± 4.8 for Novices, 7.6 ± 4.5 for Intermediates, and 10.1 ± 6.2 for Experts. When a feedback was present, experts tended to show higher average scores (Evaluative: 12.6 ± 6.2 ; Explanatory: 13.7 ± 6.8 ; Both: 13.0 ± 5.1) than novices (Evaluative:

8.4 ± 5.0; Explanatory: 9.5 ± 6.4; Both: 6.8 ± 5.3) and intermediates (Evaluative: 10.1 ± 6.4; Explanatory: 9.1 ± 5.4; Both: 10.0 ± 4.8). Notably, the combined feedback condition did not consistently yield the highest performance within groups (e.g., Novices: Both < Explanatory; Experts: Both slightly < Explanatory).

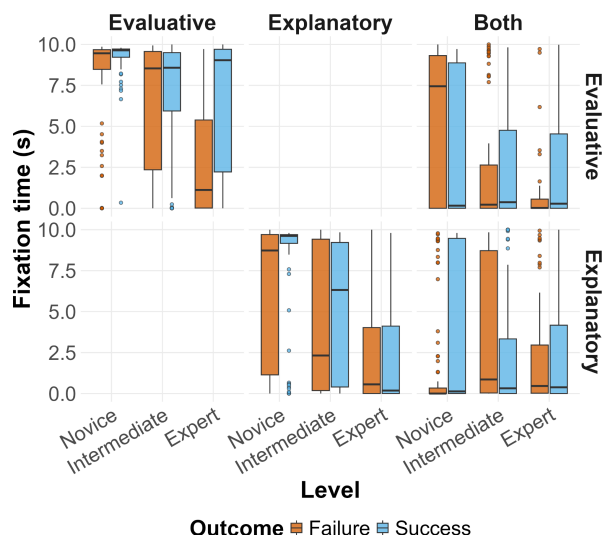


Figure 2: Distribution of fixation duration across feedback condition, trial outcome and expertise level.

Table 2: Mean fixation duration (s) across feedback condition, (Mean ± SD) trial outcome and level of expertise.

Condition	Novice	Intermediate	Expert
Eval - Solo - Fail	8.25 ± 2.70	6.41 ± 3.87	2.88 ± 3.54
Eval - Solo - Succ	9.18 ± 1.30	7.02 ± 3.50	6.61 ± 3.92
Eval - Both - Fail	5.27 ± 4.34	2.25 ± 3.60	0.90 ± 2.13
Eval - Both - Succ	3.84 ± 4.36	2.60 ± 3.61	2.62 ± 3.75
Exp - Solo - Fail	6.31 ± 4.08	4.32 ± 4.28	2.48 ± 3.49
Exp - Solo - Succ	8.14 ± 3.25	5.04 ± 4.02	2.19 ± 3.30
Exp - Both - Fail	1.37 ± 3.04	3.56 ± 4.10	2.25 ± 3.39
Exp - Both - Succ	3.24 ± 4.41	2.26 ± 3.36	2.37 ± 3.49

Fig. 2 and Tab. 2 report fixation durations toward each feedback element during the 10-s trials, within each expertise level and sorted by trial outcome. Two complementary tendencies were apparent. Novices showed longer fixation on the explanatory feedback during successful versus failed trials, both when presented alone (Success: 8.14 ± 3.25 s; Failure: 6.31 ± 4.08 s) and when combined with the evaluative feedback (Success: 3.24 ± 4.41 s; Failure: 1.37 ± 3.04 s). Conversely, experts showed longer fixation on the evaluative feedback during successful versus failed trials, both in the Evaluative-only condition (Success:

6.61 ± 3.92 s; Failure: 2.88 ± 3.54 s) and in Both (Success: 2.62 ± 3.75 s; Failure: 0.90 ± 2.13 s). Experts' fixation on the explanatory feedback remained comparatively low and similar across outcomes (Explanatory-only Success: 2.19 ± 3.30 s; Failure: 2.48 ± 3.49 s). Overall, these descriptive patterns suggest that the mere presence of the different feedback designs, as well as the consequences of actively fixating on them, impact performance differently across expertise levels.

Statistical Modeling of Success Probability: To evaluate the impact of feedback presence versus usage on SMR modulation performance during motor imagery across expertise levels, trial-level success probability was modeled using a hybrid binomial GLMM (Tab. 3).

Table 3: Fixed-effect estimates from the hybrid binomial GLMM predicting trial success.

Effect	β	p
(Intercept)	-0.805	0.090
Intermediate	0.185	0.783
Expert	0.790	0.256
Evaluative	0.304	0.278
Explanatory	-0.043	0.910
Evaluative × Explanatory	-0.470	0.347
Fixation Evaluative	0.012	0.784
Fixation Explanatory	0.106	0.011*
Intermediate × Evaluative	0.319	0.398
Expert × Evaluative	0.584	0.159
Intermediate × Explanatory	0.283	0.545
Expert × Explanatory	1.040	0.031*
Intermediate × Evaluative × Explanatory	0.432	0.498
Expert × Evaluative × Explanatory	-0.005	0.994
Intermediate × Fixation Evaluative	0.063	0.287
Expert × Fixation Evaluative	0.114	0.084
Intermediate × Fixation Explanatory	-0.063	0.254
Expert × Fixation Explanatory	-0.077	0.215

Note: * $p < .05$; . $p < .10$.

This model included the following factors: expertise level (Novice, Intermediate, Expert), feedback presence (Explanatory [yes/no] × Evaluative [yes/no]), fixation time on each feedback type (centered around the median usage). Analysis of the intercept ($\beta = -0.805$, $p = 0.090$) revealed that in the reference condition (Novices, None), the estimated success probability (approximately 0.31) marginally trended below chance levels. Besides,

the mere presence of the feedback interfaces did not significantly improve the performance of novices (Evaluative presence: $\beta = 0.304$, $p = 0.278$; Explanatory presence: $\beta = -0.043$, $p = 0.910$), nor did their combination yield a significant interaction effect ($\beta = -0.470$, $p = 0.347$). Novices' performance rather depended on how NF was used: longer fixation durations on the explanatory feedback significantly improved performance ($\beta = 0.106$, $p = 0.0115$), showing that the behavioural utility of explanatory information in novices emerged with active engagement rather than mere presence.

Impact of Expertise Level on SMR Modulation Across Feedback Conditions: As the GLMM was parameterized with Novices in the None condition as the reference level, the reported p-values primarily capture expertise-related differences in SMR modulation performance strictly under the no-feedback condition. To fully map the differential impact of expertise across NF designs, we therefore computed post-hoc estimated marginal means (EMMs) contrasts within each feedback condition at median usage (Tab. 4).

Table 4: Simple effect of expertise level within each feedback condition (Holm-adjusted).

Feedback	Level	Δ log-odds	p
None	Novice - Intermediate	-0.185	0.783
None	Novice - Expert	-0.790	0.767
None	Intermediate - Expert	-0.605	0.767
Evaluative	Novice - Intermediate	-0.504	0.456
Evaluative	Novice - Expert	-1.374	0.163
Evaluative	Intermediate - Expert	-0.870	0.432
Explanatory	Novice - Intermediate	-0.468	0.523
Explanatory	Novice - Expert	-1.830	0.046*
Explanatory	Intermediate - Expert	-1.363	0.112
Both	Novice - Intermediate	-1.219	0.156
Both	Novice - Expert	-2.409	0.004*
Both	Intermediate - Expert	-1.190	0.156

Note: * $p < .05$; . $p < .10$.

Experts and novices did not differ significantly in None ($\beta = -0.790$, $p = 0.767$) or Evaluative ($\beta = -1.374$, $p = 0.163$). A significant advantage for experts emerged only when explanatory feedback was present, alone ($\beta = -1.830$, $p = 0.046$) and combined with the evaluative feedback ($\beta = -2.409$, $p = 0.004$).

Impact of Feedback Conditions on SMR Modulation Across Expertise Levels: We next tested feedback

effects on performance by comparing each feedback condition directly against None within each expertise group, (Tab. 5). This analysis revealed a stratified profile. For Novices, none of the feedback configuration differed from the absence of feedback (Evaluative: $\beta = 0.304$, $p = 0.833$; Explanatory: $\beta = -0.043$, $p = 0.910$; Both: $\beta = -0.208$, $p = 0.900$), reinforcing the conclusion that novice benefit is not driven by the mere presence of a feedback. For Intermediates, performance significantly improved in Evaluative ($\beta = 0.623$, $p = 0.027$) and Both ($\beta = 0.826$, $p = 0.016$), whereas Explanatory alone was not significant ($\beta = 0.240$, $p = 0.387$). For Experts, all three feedback conditions significantly improved performance relative to None (Evaluative: $\beta = 0.888$, $p = 0.004$; Explanatory: $\beta = 0.998$, $p = 0.002$; Both: $\beta = 1.411$, $p = 0.001$). Together, these results show that feedback presence is sufficient to yield robust gains in experts, partially sufficient in intermediates (primarily when evaluative information is present), and insufficient in novices at median usage.

Table 5: Simple effect of feedback condition within each expertise level (Holm-adjusted).

Level	Feedback	Δ log-odds	p
Novice	Evaluative	0.304	0.833
Novice	Explanatory	-0.043	0.910
Novice	Both	-0.208	0.900
Intermediate	Evaluative	0.623	0.027*
Intermediate	Explanatory	0.240	0.387
Intermediate	Both	0.826	0.016*
Expert	Evaluative	0.888	0.004*
Expert	Explanatory	0.998	0.002*
Expert	Both	1.411	<.001*

Note: * $p < .05$; . $p < .10$.

Gaze Allocation and SMR-regulation performance: Beyond the impact of the mere presence of the feedback conditions on performance, which varies according to expertise level, the impact of active gaze allocation, reflecting the distribution of attentional resources, was also investigated. Statistical modeling confirmed a clear double dissociation in gaze-performance coupling across expertise levels. For Novices, fixation duration on the explanatory feedback significantly predicted success ($\beta = 0.106$, OR = 1.11, $p = 0.0115$), corresponding to an 11% increase in success odds per additional second, whereas fixation on the evaluative feedback did not ($\beta = 0.012$, $p = 0.7839$). Conversely, for Experts, the pattern was reversed: fixation on the explanatory feedback was not predictive ($\beta = 0.029$, $p = 0.5321$), while fixation on the evaluative feedback

robustly predicted success ($\beta = 0.126$, $OR = 1.13$, $p = 0.0099$), yielding a 13% increase in odds per additional second. Intermediates displayed a transitional profile, with a marginal positive coupling for evaluative feedback fixation ($\beta = 0.075$, $p = 0.0524$) and no significant relationship for explanatory feedback ($\beta = 0.043$, $p = 0.2398$).

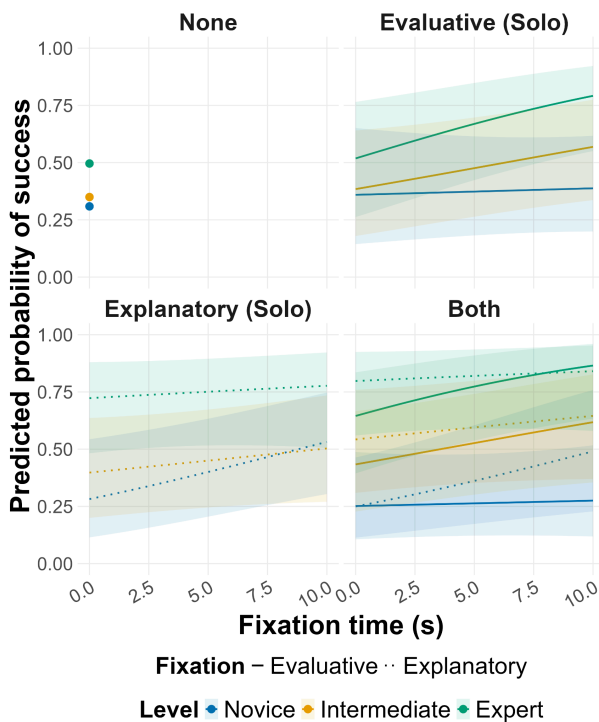


Figure 3: Estimated Probability of Success Based on Evaluative and Explanatory Fixation Times Across Expertise Levels.

Table 6: Fixation slopes by expertise level and feedback

Level	Feedback	Slope ($\Delta \log\text{-odds} / +1s$)	p
Novice	Evaluative	0.012	0.784
Novice	Explanatory	0.106	0.011*
Intermediate	Evaluative	0.075	0.052
Intermediate	Explanatory	0.043	0.240
Expert	Evaluative	0.126	0.010*
Expert	Explanatory	0.029	0.532

Note: * $p < .05$; . $p < .10$. No multiple-comparison correction applied.

Overall, the specific visual information stream driving trial success differed fundamentally with expertise: Novices performed better when actively monitoring explanatory topographic information, whereas Experts and Intermediates demonstrated greater success with sustained fixation on evaluative feedback.

DISCUSSION

The stratified profile described above is consistent with the broader variability of NF responsiveness reported in MI-NF studies [17]. Although all participants were instructed to perform kinaesthetic MI, the present results suggest that novices and experts rely on different informational routes to achieve SMR modulation. Novice athletes, still developing kinaesthetic representations [9], may benefit from explanatory information as a scaffold. This visual support enables them to actively adapt their mental strategies through a trial-and-error process, monitoring the relative involvement of sensorimotor and occipital activity to limit drift toward predominantly visual imagery [1,7,20]. Importantly, while novices spent the most time actively fixating NF overall, they tended to neglect the explanatory feedback when the evaluative feedback was simultaneously present. Because prolonged fixation on the explanatory map is the primary driver of their success, this attentional shift toward the evaluative feedback proved directly detrimental to their ability to self-regulate SMR.

Experts showed the opposite pattern: trial-by-trial success was specifically coupled with monitoring evaluative feedback, while monitoring explanatory feedback did not predict outcome. This is consistent with experts' more stable internal motor representations and the utility of a unidimensional signal for the efficient fine-tuning of an already well-defined sensorimotor strategy during MI [4,5,10,21]. Rather than sustained inspection, the optimal approach for experts seems to involve brief visual sampling of the explanatory feedback for contextual validation, while allocating the majority of their attentional resources to monitoring the evaluative signal.

Intermediates exhibited a transitional profile, consistent with a progressive maturation of kinaesthetic representations. This group represents a critical pivot for adaptive NF as it marks the stage where the feedback design must shift from explanatory to evaluative feedback.

CONCLUSION

The present study demonstrates that NF design is not a static requirement but a dynamic one that must evolve in alignment with the user's expertise. While novices require isolated explanatory feedback to guide the initial adaptation of their mental strategy, experts optimize SMR modulation through evaluative feedback that allows for the precise fine-tuning of a well-established internal motor model.

This shift in informational needs, from the descriptive "how" for novices to the quantitative "how much" for experts, highlights that the benefit of NF depends on how well the information matches the athlete's current stage of skill acquisition. When NF design is mismatched to the user's level, the information fails to be actionable and may instead act as an attentional

distraction [16]. Future developments should therefore prioritize adaptive interfaces capable of identifying objective markers of user proficiency (such as neural signature or gaze behaviour) to adjust feedback design accordingly. Such dynamic systems would ensure that NF consistently provides optimal support for users learning how to self-regulate specific brain activity.

ACKNOWLEDGEMENTS

This work received financial support from the French government in the framework of the University of Bordeaux's IdEx "Investments for the Future" program / GPR BRAIN2030 as well as the programme ANR-21-EXES-0004 - France 2030.

CONFLICT OF INTERESTS

Certain members of the research team are involved in the development of the Neurathletics application utilized in the present study. The authors declare no further commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

- [1] Guillot A, Collet C, Nguyen VA, Malouin F, Richards C, Doyon J. Brain activity during visual versus kinesthetic imagery: an fMRI study. *Human Brain Mapping*. 2009;30(7):2157-2172.
- [2] Jeannerod M. Neural simulation of action: a unifying mechanism for motor cognition. *NeuroImage*. 2001;14(1 Suppl):S103-S109.
- [3] Munzert J, Lorey B, Zentgraf K. Cognitive motor processes: the role of motor imagery in the study of motor representations. *Brain Research Reviews*. 2009;60(2):306-326.
- [4] Milton J, Solodkin A, Hluštík P, Small SL. The mind of expert motor performance is cool and focused. *NeuroImage*. 2007;35(2):804-813.
- [5] Olsson CJ, Jonsson B, Larsson A, Nyberg L. Motor representations and practice affect brain systems underlying imagery: an fMRI study of internal imagery in novices and active high jumpers. *Open Neuroimaging Journal*. 2008;2:5-13.
- [6] Hardy L, Callow N. On the efficacy of external and internal imagery: a review, theoretical critique, and recommendation. *Journal of Sport & Exercise Psychology*. 1999;21(2):71-90.
- [7] Callow N, Hardy L. The relationship between the use of kinaesthetic imagery and different visual imagery perspectives. *Journal of Sports Sciences*. 2004;22(2):167-177.
- [8] Guillot A, Collet C. Duration of mentally simulated movement: a review. *Journal of Motor Behavior*. 2005;37(1):10-19.
- [9] Lebon F, Byblow WD, Collet C, Guillot A, Stinear CM. The modulation of motor cortex excitability during motor imagery depends on imagery quality. *European Journal of Neuroscience*. 2012;35(2):323-331.
- [10] Orlandi A, Proverbio AM. Bilateral engagement of the occipito-temporal cortex in response to dance kinematics in experts. *Scientific Reports*. 2019;9(1):1000.
- [11] Sitaram R, Ros T, Stoeckel L, Haller S, Scharnowski F, Lewis-Peacock J, et al. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*. 2017;18(2):86-100.
- [12] Jeunet C, Glize B, McGonigal A, Batail JM, Micoulaud-Franchi JA. Using EEG-based brain computer interface and neurofeedback targeting sensorimotor rhythms to improve motor skills: theoretical background, applications and prospects. *Neurophysiologie Clinique*. 2019;49(2):125-136.
- [13] Babiloni C, Del Percio C, Iacoboni M, Infarinato F, Lizio R, Marzano N, et al. Golf putt outcomes are predicted by sensorimotor cerebral EEG rhythms. *Journal of Physiology*. 2008;586(1):131-139.
- [14] Gruzelier JH. EEG-neurofeedback for optimising performance. I: A review of cognitive and affective outcome in healthy participants. *Neuroscience & Biobehavioral Reviews*. 2014;44:124-141.
- [15] Jeunet C, N'Kaoua B, Subramanian S, Hachet M, Lotte F. Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns. *PLoS ONE*. 2015;10(12):e0143962.
- [16] Alkoby O, Abu-Rmieleh A, Shriki O, Todder D. Can we predict who will respond to neurofeedback? A review of the inefficiency problem and existing predictors for successful EEG neurofeedback learning. *Neuroscience*. 2018;378:155-164.
- [17] Yu X, Zhang Q, Yan J, Xiang MQ, Fu W. The effect of EEG neurofeedback training on sport performance: a systematic review and meta-analysis. *Scandinavian Journal of Medicine & Science in Sports*. 2025;35(1):e70055.
- [18] Schmidt, R.A., & Lee, T.D. (2011). *Motor Control and Learning: A Behavioral Emphasis* (5th ed.). Human Kinetics.
- [19] Renard Y, Lotte F, Gibert G, Congedo M, Maby E, Delannoy V, et al. OpenViBE: an open-source software platform to design, test, and use brain-computer interfaces. *Frontiers in Neuroscience*. 2010;4:22.
- [20] White A, Hardy L. An in-depth analysis of the uses of imagery by high-level slalom canoeists and artistic gymnasts. *The Sport Psychologist*. 1998;12(4):387-403.
- [21] Wolpert DM, Ghahramani Z, Jordan MI. An internal model for sensorimotor integration. *Science*. 1995;269(5232):1880-1882.
- [22] Sigrist, R., Rauter, G., Riener, R., & Wolf, P. (2013). Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review*, 20(1), 21–53.

Towards an Integrated Model of Onset Detection and Multiclass Recognition

D.M. Carmona-Peña¹, C.A. Reyes-García¹, A.A Torres-García¹

¹Biosignals Processing and Medical Computing Laboratory, Instituto Nacional de Astrofísica,
Óptica y Electrónica, Puebla, Mexico

E-mail: david.carmona@inaoe.mx

ABSTRACT: Brain-Computer Interfaces (BCIs) based on imagined speech offer a promising alternative for intuitive communication; however, the transition to real-time asynchronous systems is hindered by the "onset detection" problem. This study evaluates the efficacy of Instant Wavelet Energies (IWE), Empirical Mode Decomposition (EMD), and Generalized Hurst Exponent (GHE) for identifying the transition between Idle State Segments (ISS) and Imagined Word Segments (IWS). We analyzed three datasets (DS1, DS2, and DS3) using Support Vector Machines (SVM), Random Forest (RF), k-NN, and Logistic Regression (LogReg). Our results demonstrate that binary intent detection consistently outperforms multiclass word recognition, with the latter remaining a significant challenge. The MCC remains < 0.35 for most configurations, indicating that while these features are effective for intention "gating"; however, they exhibit a lack of the requisite specificity to achieve decisive discrimination.

INTRODUCTION

A Brain-Computer Interface (BCI) is a system that establishes a direct communication channel between bioelectrical brain activity and external devices, enabling the control of assistive devices or prostheses without the intervention of conventional neuromuscular pathways [1]. These interfaces are fundamentally based on the acquisition and processing of biological signals, predominantly electroencephalograms (EEG) to decode cognitive processes. Their application is crucial in the development of assistive technologies aimed at restoring autonomy in individuals with severe motor disabilities.

In the state of the art, BCI systems are primarily categorized into synchronous and asynchronous modes. Most current developments operate under a synchronous paradigm, where the timing and speed of communication are predetermined by the experimental protocol [2, 3]. In these systems, users generate commands exclusively within temporal windows defined by external stimuli, such as P300 evoked potentials or Steady-State Visual Evoked Potentials (SSVEP) [2].

However, this dependence on external stimuli entails significant limitations. It not only requires precise visual stimulation hardware (e.g., LED matrices) that compli-

cates the technical setup [4], but also induces considerable visual fatigue after prolonged periods of use [5].

In contrast, asynchronous (or self-paced) systems offer a more natural interaction, allowing the user to issue commands voluntarily at any time. Since they do not rely on external stimuli, the system must continuously analyze the EEG stream to detect cognitive intent. The critical challenge in these systems is the "onset detection" problem, which consists of accurately discriminating between an active cognitive task and an inactive or resting state [6]. In the literature, this problem has been defined as intentional versus non-intentional control, primarily within the context of motor imagery [7, 8]. To address this asynchrony, various taxonomies of approaches have been developed, ranging from power threshold analysis and state-based models to the use of gating classifiers or evidence accumulation detectors [9–13].

As an alternative, the use of imagined speech has been proposed. While we acknowledge the pioneering advances of Chang et al. [14–16] in using electrocorticography (ECoG) to achieve superior Information Transfer Rates (ITR), our work focuses on accessibility and non-invasiveness. While Chang's methods require high-density surgical implants in patients with specific pathologies, our proposal seeks to optimize the detection of inactivity states through EEG. This enables a low-cost BCI technology with lower clinical risk, which remains an open challenge given the low signal-to-noise ratio inherent to non-invasive methods.

This neuroparadigm in EEG signals ranges from the identification of basic phonetic units such as vowels [17] and phonemes [18] to the decoding of syllables [19] and full words with semantic meaning [20, 21], as well as the identification of emotions transmitted through speech in EEG signals [22]. Despite its potential, current literature lacks a robust approach for detecting imagined speech segments independently of the specific words of interest (e.g., "up", "down").

The contribution of this study lies in the proposal of a two-phase BCI structure. In this scheme, the first stage acts as a crucial stochastic filter that conditions the system's flow: the transition to secondary classification only occurs after effective isolation of the intention segment. This hierarchical model aims to optimize the allocation of computational resources and highly specialized algo-

rhythms, prioritizing the analysis of time windows with substantial informational weight, such as imagined word segments (IWS), and separating them from idle state segments (ISS).

This work builds upon the feature extraction framework proposed by Hernández-del-Toro et al. [20], applying a 1–50 Hz bandpass filter and evaluating its effectiveness on three different datasets across three distinct tasks: (1) discrimination between imagined word segments (IWS) and inactive state segments (ISS), (2) multiclass classification of imagined words in Spanish, and (3) a combined classification of words versus the inactive state, which more closely resembles a real-world application. Model performance is assessed using the Matthews Correlation Coefficient (MCC). Unlike the F1-score, whose values are not directly comparable across problems with different numbers of classes, MCC maintains a fixed scale of $[-1, +1]$ where zero corresponds to random prediction regardless of the class distribution. This property makes it particularly suitable for jointly evaluating the binary task [23] and the multiclass tasks [24], both of which exhibit class imbalance due to the inclusion of the inactive state.

MATERIALS

Datasets:

The imagined speech datasets employed in this study correspond to those previously used by Hernández-del-Toro et al. [20], with the exception of their third database, which has been replaced by the dataset from Nieto et al. [25], totaling three databases (DS1, DS2, and DS3). Datasets DS1 and DS2 comprise the imagined words "up," "down," "left," "right," and "select" ("arriba", "abajo", "izquierda", "derecha", "seleccionar"). Meanwhile, DS3 includes the commands "up," "down," "left," and "right" ("arriba", "abajo", "izquierda", "derecha").

The third database from [20], discarded due to technical inconsistencies identified during the data adaptation phase, specifically regarding the synchronization and alignment of event markers with the internal targets of the recorded signals.

DS1: Data were recorded using the EMOTIV EPOC+ device (14 channels, 128 Hz sampling rate) with the participation of 27 subjects. For each participant, 20 repetitions per word were obtained, resulting in a total of 100 samples per subject [26].

DS2: Recorded using the EMOTIV EPOC+ device (14 channels, 128 Hz sampling rate) with 27 subjects distinct from those in DS1. For each subject, 32 repetitions per word were obtained, totaling 160 samples per subject. In this data acquisition protocol, words were presented in a randomized order [26].

DS3: This dataset comprises multimodal recordings of electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG) signals acquired with the BioSemi ActiveTwo high-resolution

biopotential measurement system. The configuration consisted of 128 active EEG channels and 8 external active EOG/EMG channels, with 24-bit resolution and an initial sampling rate of 1024 Hz. Subsequently, the data were decimated by a factor of four, resulting in a final sampling rate of 256 Hz. This set integrates data from 10 subjects across three different sessions. For each session, 50 repetitions per word were obtained, recording signals for imagined speech, spoken speech, and visualization, for a total of 200 samples per session per subject [25].

METHOD

Data Cleaning and Epoching:

Before the preprocessing stage, a data cleaning and segmentation phase was performed to isolate the epochs of interest from the continuous recordings. This process was implemented using MATLAB for all three datasets, ensuring a standardized approach to identifying and extracting trials based on event markers. Each essay was structured to cover the period leading up to the imagined speech, the imagined speech period, and the subsequent rest phase:

DS1: A total of 100 trials were extracted per subject with a mean duration of $4.3 \text{ s} \pm 1 \text{ s}$.

DS2: A total of 160 trials were extracted per subject with a mean duration of $5.6 \text{ s} \pm 1 \text{ s}$.

DS3: From the original 5,640 trials and over 9 hours of multimodal recordings, a specific subset was isolated. For this study, only the imagined speech condition was selected, excluding spoken speech and visualization trials. This resulted in 200 trials per subject across three sessions (50 repetitions per word), each with a standardized duration of 4.5 s.

The choice of preprocessing techniques was dictated by the electrode density of each system. For DS1 and DS2 (14 channels), CAR was applied as the primary spatial filtering strategy due to its robustness and suitability for low-density montages, where the application of ICA is limited by insufficient dimensionality for reliable source separation.

In contrast, for DS3 (128 EEG channels plus auxiliary channels capturing EOG/EMG), ICA was employed as a preprocessing step for artifact removal, leveraging the higher dimensionality and the availability of reference signals to achieve a more accurate decomposition of neural and non-neural sources.

Following the segmentation stage, each trial was delimited by specific start and end markers identifying the Imagined Words Segment (IWS) and its corresponding word label. To prepare these segments for feature extraction, the processing and validation pipeline (Figure 1) was implemented. This workflow encompasses signal filtering, experimental scenario definition, sliding window segmentation, and the final machine learning evaluation.

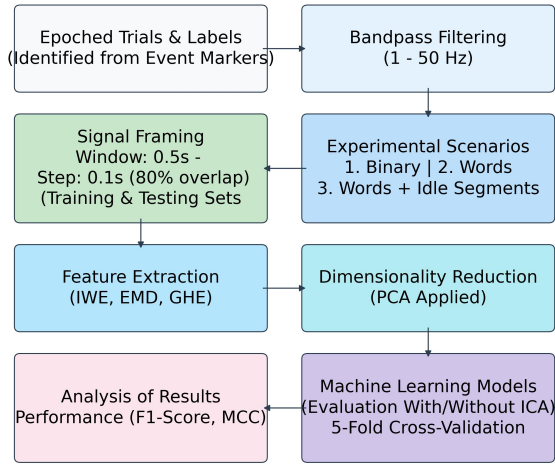


Figure 1: Block diagram of the methodology for imagined speech detection and classification. The process spans from initial epoching to the final machine learning evaluation, comparing performance with and without PCA across different experimental tasks.

This bandpass filter was applied because previous frequency analysis revealed significant saturation below 1 Hz. Additionally, this filter effectively eliminates power line interference (50/60 Hz) and high-gamma frequencies, which are outside the scope of this study.

Unlike the original work, we evaluate the performance of the three scenarios under the extracted feature sets. The sliding window segmentation (Signal Framing) and model training were conducted independently for each scenario using a hybrid workflow: feature extraction was executed in MATLAB (exporting data in .h5 format), while model training and validation were implemented in Python.

Experimental Scenarios Detail:

IWS vs. ISS (Binary): For the training phase, a specific division is made prior to Signal Framing: Class 0 (ISS) aggregates segments recorded before and after word imagination, while Class 1 (IWS) includes all imagined speech segments. For the testing phase, signal framing is applied to the entire trial; the label for each window is assigned based on a majority voting criterion, determined by the predominant label within the 0.5s window duration.

Words (Multi-class): This scenario exclusively isolates the imagined speech segments. Labels are assigned according to the specific word (e.g., "up", "down"), resulting in a number of classes that depends on the database (DS1/DS2/DS3). Signal Framing is applied only to these epochs for both training and testing sets.

Words + idle (Words + Inactive): This scenario follows a logic similar to the binary one but with higher granularity. Imagined speech segments retain their individual words labels (word classes), while all rest segments are consolidated into a final category designated as Idle State.

Feature extraction:

1.-Wavelet Transform (DWT):

A four-level DWT was implemented using the bior2.2 mother wavelet. This technique decomposes the signal into 1 set of approximation coefficients (low frequencies) and 4 set of detail coefficients (high frequencies) 1 per level. To quantify the power distribution in the time-frequency domain, the Instant Wavelet Energie (IWE) was calculated for each of the obtained coefficients [20]. This procedure generates a vector of 5 features per channel.

2. Empirical Mode Decomposition (EMD):

EMD is an adaptive technique that decomposes complex signals into basis functions called Intrinsic Mode Functions (IMF) and a final residue. Unlike other methods, EMD does not rely on a predefined basis function, making it ideal for non-linear and non-stationary EEG signals. Following the decomposition, the two IMFs exhibiting the greatest Euclidean distance relative to the original signal were selected, ensuring the capture of the most physically relevant components[20]. For each of these two IMFs, six non-linear metrics were calculated: Teager Energy (TE) and Instantaneous Energy (IE). Higuchi and Katz fractal dimensions. Generalized Hurst Exponent (GHE) for orders $q = 1$ and $q = 2$. This stage contributes a total of 12 features per channel (6 metrics $\times$ 2 IMFs).

3. Fractal Analysis and Generalized Hurst Exponent (GHE):

Given the non-stationary nature and dynamic complexity of EEG signals, the GHE was employed to quantify self-similarity and temporal scaling properties. The GHE extends the classic Hurst concept by evaluating different statistical moments of signal fluctuations, allowing for the identification of persistent, random, or anti-persistent behaviors.

Exponents were calculated for two specific orders:

$H(q = 1)$: Characterizes the dynamics associated with absolute fluctuations of the signal[20]. $H(q = 2)$: Linked to second-order dynamics and the variance of fluctuations[20].

The analysis of these two orders captures complementary information regarding the multifractal structure of the EEG signal, generating 2 additional features per channel.

Feature Vector Summary:

By integrating the aforementioned techniques (5 DWT + 12 EMD + 2 GHE), a total of 19 features per channel is obtained. Based on the electrode configurations of each database, the resulting feature vector per trial is:

DS1 and DS2: 266 features (14 channels $\times$ 19).

DS3: 2432 features (128 channels $\times$ 19).

Classification:

The classification phase was conducted using four distinct machine learning architectures. RF was implemented as an ensemble of 100 decision trees. The kNN

algorithm was configured with $k = 100$ neighbors. For the SVM, a non-linear decision boundary was established using the Radial Basis Function (RBF) kernel with a default regularization parameter ($C = 1.0$). Finally, LogReg was employed as a probabilistic classifier with an L2 penalty for weight regularization; notably, the maximum number of iterations was increased to 10,000 to guarantee model convergence across all experimental scenarios.

Data Leakage Prevention:

Because the original trials underwent differentiated segmentation (as described in Experimental Scenarios Detail), a rigorous data integrity control was implemented. Strict trial-level separation was guaranteed: any trial selected for the test set was entirely excluded from the training database. This procedure prevents the model from being evaluated on overlapping segments or signals derived from data already seen during the learning phase, thereby ensuring the generalization capability of the results when presented with unseen data.

Table 1: Experimental structure and training configuration per dataset and scenario

Scenario / Task	Feature Set	PCA	ML Models
Binary	All	Yes / No	4 / 4
	IWE, EMD, GHE	No	4 × 3
Words*	All	Yes / No	4 / 4
	IWE, EMD, GHE	No	4 × 3
Words-all**	All	Yes / No	4 / 4
	IWE, EMD, GHE	No	4 × 3

Note: A total of 60 models were trained per subject across the three datasets (DS1, DS2, and DS3) * 5 classes for DS1/DS2 and 4 for DS3 ** 6 classes for DS1/DS2 and 5 for DS3 .

To evaluate the performance of the classifiers across the different imagined speech scenarios, two metrics were used: the F1-score and the Matthews Correlation Coefficient (MCC).

RESULTS AND DISCUSSION

The application of Principal Component Analysis (PCA) in this study was intended to manage the high dimensionality of the features extracted from the EEG signals, aiming to mitigate noise and reduce spatial redundancy. However, a comparative analysis between the results obtained using the original feature set (Table 2) and those following the application of PCA (Table 3) reveals a distinct performance pattern. This behavior appears to be strictly dependent on the channel density and the specific nature of the experimental task.

Statistical validity was assessed using the Wilcoxon signed-rank test with Bonferroni correction. For DS1 and DS2, no significant differences were found ($p \geq 0.05$), justifying the use of the original feature set to preserve data integrity. In DS3, despite a marginal improvement in the binary scenario with PCA, the change was not statistically significant. This suggests that inter-subject and inter-session variability remain the dominant factors in performance, outweighing the benefits of linear dimen-

sionality reduction even in high-density (128-channel) configurations.

Table 2: Best performing classifier for each scenario using the combined feature set (All) without PCA.

Dataset / Scenario	Best Classifier	MCC $\pm$ SD
DS1 (Binary)	LogReg	0.250 $\pm$ 0.136
DS1 (Words)	LogReg	0.423 $\pm$ 0.169
DS1 (Words + Inactive)	LogReg	0.049 $\pm$ 0.072
DS2 (Binary)	LogReg	0.153 $\pm$ 0.118
DS2 (Words)	SVM	0.032 $\pm$ 0.059
DS2 (Words + Inactive)	SVM	0.071 $\pm$ 0.053
DS3 (Binary)	SVM	0.299 $\pm$ 0.116
DS3 (Words)	LogReg	-0.009 $\pm$ 0.048
DS3 (Words + Inactive)	SVM	0.131 $\pm$ 0.050

Table 3: Best performing classifier for each scenario using the combined feature set (All) with PCA.

Dataset / Scenario	Best Classifier	MCC $\pm$ SD
DS1 (Binary)	LogReg	0.253 $\pm$ 0.141
DS1 (Words)	LogReg	0.402 $\pm$ 0.168
DS1 (Words + Inactive)	LogReg	0.050 $\pm$ 0.075
DS2 (Binary)	LogReg	0.143 $\pm$ 0.117
DS2 (Words)	SVM	0.030 $\pm$ 0.057
DS2 (Words + Inactive)	SVM	0.069 $\pm$ 0.053
DS3 (Binary)	LogReg	0.317 $\pm$ 0.114
DS3 (Words)	RF	-0.006 $\pm$ 0.040
DS3 (Words + Inactive)	SVM	0.128 $\pm$ 0.050

The superior performance achieved in DS1 (MCC = 0.423) should be interpreted considering that the data were acquired through a continuous stimulation protocol without randomization of stimuli. This sequential structure likely enhanced the stability of neural patterns, enabling the classifier to optimally leverage the integrity and relevance of the original extracted features.

For DS2, the LogReg model in the Binary task without PCA achieved an average F1-score of 0.704 ± 0.052 across 27 subjects, with a corresponding MCC of 0.153 ± 0.118 . This significant discrepancy between metrics suggests that the model is biased toward the majority class, which in this case represents the inactive states. This phenomenon is attributed to the acquisition methodology, where the duration of the imagined speech trials was self-regulated by the subjects through a keypress at the beginning and end of the mental task, resulting in a temporal class imbalance. Similarly, for the Words + Inactive scenario using SVM, the observed F1-score of 0.256 ± 0.055 and MCC of 0.071 ± 0.053 further indicate that the classifier predominantly identifies the majority class.

In contrast, DS3 featured stricter control over the timing of the imagined speech task; however, a class imbalance persists. In the Binary task with PCA, the LogReg model with PCA yielded the highest performance with an MCC of 0.317 ± 0.114 and an associated F1-score of 0.746 ± 0.043 . For the Words + Inactive dataset, the SVM model with PCA achieved an MCC of 0.128 ± 0.050 and an F1-score of 0.339 ± 0.036 . These results demonstrate that while improved temporal control enhances classification

quality, the underlying class imbalance continues to influence the relationship between these performance metrics.

Evaluation by Feature Type:

To identify the individual contribution of each extraction method (IWE, EMD, GHE), independent tests were conducted. The comparative performance across different datasets and tasks is summarized in Table 4.

Table 4: Top performing configuration (Classifier and Feature Extraction method) for each dataset and experimental scenario (without PCA).

Dataset / Scenario	Rank 1: Model + Feature (MCC $\pm$ SD)
DS1 (Binary)	LogReg + IWE (0.242 $\pm$ 0.157)
DS1 (Words)	LogReg + IWE (0.422 $\pm$ 0.156)
DS1 (Words + Inactive)	LogReg + EMD (0.053 $\pm$ 0.075)
DS2 (Binary)	LogReg + IWE (0.157 $\pm$ 0.114)
DS2 (Words)	SVM + GHE (0.027 $\pm$ 0.037)
DS2 (Words + Inactive)	RF + GHE (0.054 $\pm$ 0.042)
DS3 (Binary)	LogReg + IWE (0.325 $\pm$ 0.112)
DS3 (Words)	LogReg + IWE (0.003 $\pm$ 0.046)
DS3 (Words + Inactive)	SVM + EMD (0.133 $\pm$ 0.039)

Table 4 reports the best-performing configuration (classifier and feature extraction method) for each dataset and classification scenario in the absence of PCA. Logistic Regression combined with IWE features consistently ranks first in five of the nine evaluated combinations, suggesting that this pair generalizes well across binary tasks. The lowest non-trivial performance (0.003 $\pm$ 0.046) corresponds to DS3 in the Words scenario, highlighting the substantial inter-dataset variability that characterizes imagined-speech decoding. For the most ecologically realistic task (Words + Inactive), the best configurations achieve modest MCC values (0.053 for DS1, 0.054 for DS2, and 0.133 for DS3), with SVM combined with EMD features producing the highest result. These figures confirm that the inclusion of the inactive class introduces substantial difficulty due to class imbalance, and that no single classifier–feature combination dominates across all datasets and scenarios.

A noteworthy observation is the relatively high MCC obtained on DS1 in the Words scenario (0.422 $\pm$ 0.156 using LogReg + IWE), which exceeds the values reported for the other two datasets in the same scenario by a wide margin. As previously described, DS1 was acquired using a non-randomized (sequential) stimulation protocol, in contrast to the randomized protocols employed in DS2 and DS3. This sequential structure likely provided a more stable cognitive framework and reduced cognitive load for the participants, favoring the consistency of EEG patterns and allowing linear models such as Logistic Regression to identify discriminative features more effectively. This result should therefore be interpreted with caution: the higher MCC reflects the specific acquisition protocol rather than an intrinsic advantage of the classification pipeline, and it is not directly comparable to the values obtained on DS2 and DS3.

Across all datasets, Logistic Regression and SVM emerged as the most robust classifiers. This consistency,

combined with the strong performance of IWE and EMD features, suggests that discriminative information related to imagined speech can be effectively captured through linear relationships in the time–frequency domain. The competitive performance of EMD features in the Words + Inactive scenario further indicates their capacity to extract non-stationary components of the EEG signal that retain class-discriminative information even under the more challenging conditions imposed by class imbalance.

CONCLUSION

Results demonstrate that, regardless of whether the complete feature space or isolated subsets are employed, superior performance—quantified by the Matthews Correlation Coefficient (MCC)—is consistently achieved for the intention identification task (IWS vs. ISS) relative to multiclass scenarios. This trend is particularly salient when contrasting the binary onset task against the continuous classification paradigm (Words + Inactive), which serves as the most proximal model for real-time BCI applications.

The evaluation in the multiclass word scenario confirmed that the proposed feature sets are primarily optimized for detecting the cognitive intent of speech (IWS vs. ISS). This experiment corroborated that these descriptors are not designed to efficiently discriminate between individual words, proving insufficient for an adequate classification of imagined speech words. While the observed MCC values consistently remain below 0.35—which exceeds the chance threshold but recognizes non-optimal performance—these findings demonstrate that the features implemented from the baseline work of [20] provide a viable operational basis for the first gating stage. Consequently, this scenario underscores that the path to enabling a successful transition toward word classification does not lie in the simple tuning of classifiers, but in the strategic evolution of this first phase of intent detection.

To perfect this isolation of IWS from ISS and break current limits, future research must maintain the gating architecture but expand the analytical framework beyond isolated-channel paradigms. While evidence demonstrates that the current feature sets, specifically IWE and EMD, are effective as initial filters, it is essential to leverage the intrinsic high-dimensionality and spatiality of the EEG manifold without imposing restrictive linearizations. Directly integrating spatial covariance through adaptive methods and multiresolution analysis thereby allows transforming the BCI from a conventional pattern classifier into a dynamic system, capable of faithfully decoding the complex cortical activity of speech within non-invasive neural interfaces.

ACKNOWLEDGMENTS

The authors thankfully acknowledge the computer resources, technical expertise and support provided by the Laboratorio Nacional de Supercómputo del Sureste de

México, SECIHTI member of the network of national laboratories. I wish to express my gratitude for the postgraduate scholarships awarded by SECIHTI, Mexico. These scholarships have enabled me to conduct the research presented in this work.

REFERENCES

- [1] Moreno I et al. Los sistemas de interfaz cerebro-computadora basado en EEG: Características y aplicaciones. *I+D Tecnológico*. 2019;15(2):13–26.
- [2] Chailloux Peguero JD, Mendoza-Montoya O, Antelis JM. Single-option P300-BCI performance is affected by visual stimulation conditions. *Sensors*. 2020;20(24):7198.
- [3] Waytowich N et al. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of Neural Engineering*. 2018;15(6):066031.
- [4] Fazel-Rezai R et al. P300 brain computer interface: Current challenges and emerging trends. *Frontiers in neuroengineering*. 2012;5:14.
- [5] Saha S et al. Progress in brain computer interface: Challenges and opportunities. *Frontiers in Systems Neuroscience*. 2021;15:578875.
- [6] Mitra K et al. Characterizing the onset and offset of motor imagery during passive arm movements induced by an upper-body exoskeleton. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2023, 3789–3794.
- [7] Birch GE, Bozorgzadeh Z, Mason SG. Initial online evaluations of the LF-ASD brain-computer interface with able-bodied and spinal-cord subjects using imagined voluntary motor potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2002;10(4):219–224.
- [8] Shahini N et al. Automatically identified EEG signals of movement intention based on CNN network (End-To-End). *Electronics*. 2022;11(20):3297.
- [9] Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [10] Blankertz B et al. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2003;11(2):127–131.
- [11] Obermaier B, Müller GR, Pfurtscheller G. "virtual keyboard" controlled by spontaneous EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2003;11(4):422–426.
- [12] Perdakis S, Leeb R, Millán JdR. Context-aware adaptive spelling in motor imagery BCI. *Journal of Neural Engineering*. 2016;13(3):036018.
- [13] Faulkner HG et al. Sequential hypothesis testing for automatic detection of task-related changes in cerebral perfusion in a brain–computer interface. *Neuroscience Research*. 2015;100:29–38.
- [14] Silva AB et al. The speech neuroprosthesis. *Nature Reviews Neuroscience*. 2024;25(7):473–492.
- [15] Metzger SL et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*. 2023;620(7976):1037–1046.
- [16] Moses DA et al. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature Communications*. 2019;10(1):3096.
- [17] DaSalla CS et al. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Networks*. 2009;22(9):1334–1339.
- [18] Shevchenko AO, Vartanov AV. Comparison of the mechanisms of phonemic awareness and internal pronunciation of phonemes and syllables: EEG and fMRI study. *Russian Psychological Journal*. 2022;19(4):186–210.
- [19] Suyuncheva A et al. Reconstruction of words, syllables, and phonemes of internal speech by EEG activity. In: *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics*. Springer International Publishing: Cham, 2021, 319–328.
- [20] Hernández-Del-Toro T, Reyes-García CA, Villaseñor-Pineda L. Toward asynchronous EEG-based BCI: Detecting imagined words segments in continuous EEG signals. *Biomedical Signal Processing and Control*. 2021;65:102351.
- [21] Cooney C, Folli R, Coyle D. A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech. *IEEE Transactions on Biomedical Engineering*. 2022;69(6):1983–1994.
- [22] Chang J et al. Generative listener EEG for speech emotion recognition using generative adversarial networks with compressed sensing. *IEEE Journal of Biomedical and Health Informatics*. 2024;28(4):2025–2036.
- [23] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6.
- [24] Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*. 2004;28(5–6):367–374.
- [25] Nieto N et al. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Scientific Data*. 2022;9(1):52.
- [26] Torres-García AA. Análisis y clasificación de electroencefalogramas (EEG) registrados durante el habla imaginada. Español. Tesis de doctorado. Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Santa María Tonantzintla, Puebla, México. Apr. 2016. [Online]. Available: <https://repositorio.inaoe.mx/handle/123456789/100>.

Using brain network features in motor imagery-based BCI for stroke rehabilitation

Bousfiha C¹, Gabot C¹, Baili D¹, Couton M¹, Adwane G³, Obadia M³, Zavanone C⁴, Dupont S⁴, Marcilliere Q⁵, Bayen E⁵, Bougrain L^{1,2}, Bartolomeo P¹, de Vico Fallani F¹

¹Paris Brain Institute - ICM, Sorbonne Universite, CNRS, Inria, Inserm, AP-HP, Pitie Salpetriere Hospital, Paris, France

²Universite de Lorraine, CNRS, LORIA, F-54000, Nancy, France

³Neurology Department, Hôpital Fondation A. de Rothschild, Paris, France

⁴AP-HP-Service de Soins de Suite et Réadaptation, Hôpital Pitié-Salpêtrière, Paris, France

⁵AP-HP-Service de Médecine Physique et Réadaptation, Hôpital Pitié-Salpêtrière, Paris, France
E-mail: camile.bousfiha@icm-institute.org

ABSTRACT: Motor imagery brain-computer interfaces are promising tools for post-stroke rehabilitation, but most systems rely on spectral power features and are often tested in highly selected patient populations. Because stroke affects large-scale brain networks, connectivity-based markers may provide complementary information for decoding and monitoring rehabilitation. This pilot study evaluated the feasibility of an intensive 6-week MI-BCI protocol combined with functional electrical stimulation in five chronic stroke patients with heterogeneous and often severe motor impairment. Functional connectivity analyses revealed the emergence of motor imagery-related modulations over centro-frontal and centro-parietal regions. These findings support the feasibility of intensive MI-BCI rehabilitation in chronic stroke and suggest that functional connectivity may be a relevant biomarker for adaptive BCI strategies.

INTRODUCTION

Brain-computer interfaces (BCIs) have emerged as a promising approach to promote motor recovery after stroke by coupling motor imagery with contingent feedback, such as functional electrical stimulation, robotic assistance, or sensory feedback [1]. By directly linking brain activity to external effectors, BCIs aim to reinforce use-dependent plasticity and restore motor function through repeated activation of sensorimotor networks. This approach is particularly relevant in stroke rehabilitation, where impaired voluntary movement limits the delivery of conventional task-oriented therapies, especially in patients with severe motor deficits.

Despite encouraging results [2, 3], the transfer of BCI-based rehabilitation from controlled research settings to routine clinical practice remains limited. An important step is to evaluate BCI feasibility in patients reflecting the clinical complexity of stroke, including associated cognitive or language deficits, attentional fluctuations, fatigue, and comorbidities. These factors may contribute to al-

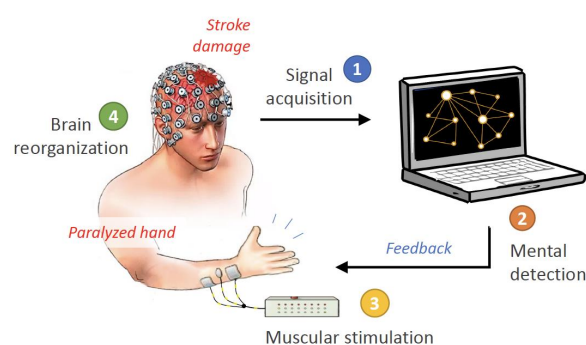


Figure 1: Schematic representation of the BCINET protocol. 1) EEG signals are acquired and transformed in functional brain networks via imaginary coherence. 2) Node-strengths are extracted from the resulting time-varying brain networks and considered as classification features. 3) A functional electrical stimulator (FES) commands the hand opening in case of correct detection. 4) This feedback is hypothesized to influence the brain network reorganization across sessions. More details ClinicalTrials.gov (NCT06951035)

tered sensorimotor rhythm modulation and increased BCI inefficiency [4]. They also highlight the need for decoding approaches that are robust to altered oscillatory activity and better adapted to heterogeneous stroke population.

In parallel, stroke is increasingly conceptualized as a network-level disorder rather than a purely focal lesion [5] Functional recovery depends on the reorganization of large-scale brain networks, including intra- and inter-hemispheric sensorimotor interactions [6]. In this framework, the activity of isolated cortical regions may provide only a partial description of the underlying neuroplastic processes. Connectivity-based decoding strategies may therefore be used in BCI protocols to integrate information across multiple cortical regions, thus reducing sensitivity to focal signal alterations [7]. Such network-level biomarkers could also facilitate adaptive and patient-specific training, which is essential for translating BCIs

Table 1: Patient characteristics

Patient	Age (years)	Stroke side	Time since stroke (months)	UE-FMA
P01	47	L	12	13
P02	62	R	63	49
P03	78	L	45	18
P04	51	R	17	19
P05	55	R	16	18

into clinical practice (Fig. 1). Among connectivity estimators, measures based on the imaginary part of coherency (ImCoh) have been recommended to reduce spurious zero-lag coupling and improve the physiological interpretability of EEG networks [8]. This choice is supported by previous network-based work showing that motor imagery after stroke involves altered functional brain network organization [9]. In MI-BCI settings, phase-synchronization measures, including ImCoh, have been shown to capture task-related sensorimotor modulations [10]. More recently, the FUCONE framework showed that functional connectivity estimators can provide complementary information for MI-BCI classification [11]. In this context, we designed an intensive and individualized rehabilitation protocol combining motor imagery-based BCI training with neuromodulation. The present pilot study aimed to evaluate the feasibility of this approach and to explore the neurophysiological changes associated with intensive, connectivity-based BCI rehabilitation.

MATERIALS AND METHODS

Study design and population: Five patients with chronic stroke were included in this pilot phase aimed at assessing the feasibility and tolerability of an intensive MI-BCI rehabilitation protocol (6 weeks, 3 sessions/week), prior to a subsequent phase investigating the additional effect of non-invasive brain stimulation. The study was registered on ClinicalTrials.gov (NCT06951035). Thus, consecutive chronic stroke patients with upper limb deficit were recruited (Table.1). The mean age was 58.6 years (range 47–78) and time since stroke ranged from 12 to 63 months. Three patients had right hemisphere lesions and two left hemisphere lesions. Four patients presented severe distal motor impairment (UE-FMA < 32), three of whom had no voluntary hand opening. Two patients presented aphasia but were able to understand the instructions and complete the protocol.

BCI protocol: Sessions were conducted in a quiet, shielded environment with patients seated in front of a screen table and the paretic arm supported on an arm-rest. EEG was recorded using a Brain products amplifier with active electrodes (impedance < 10 kΩ). FES was delivered using the Fesia Grasp system [12] placed on the paretic forearm.

The BCI paradigm consisted of a MI-baseline design [13] in which only the MI condition was presented and repeated 18 times per run. Each trial began with a 5-second

baseline resting period (hollowed circle). This was followed by a 4-second cue period during which a red circle appeared on the screen, instructing the patient to perform motor imagery. The cue period was followed by a 3-second feedback phase, during which FES was delivered when MI was correctly detected. Finally, an intertrial interval of 2 to 3 seconds (random jitter) was introduced to reduce temporal predictability and anticipation effects. Closed-loop BCI experiment was performed with Openvibe [14]. Each session consisted of one calibration run (only positive feedback) following 8 training run, corresponding to 144 MI trials in total. This number was selected to ensure a rehabilitation-oriented dosage exceeding 100 task repetitions per session [15].

EEG signal acquisition and processing: EEG signals were recorded with a 64-channel BrainAmp system (referenced to mastoids signals) with a 500 Hz frequency sampling. Left or right electromyogram signals were recorded on the affected arm to ensure that subjects were not moving their forearm while performing the tasks. Signal processing was computed with MNE python library. Online feature extraction and decoder training was performed with Happyfeat software [16].

Feature extraction: For the MI condition, features were computed from the last 3 seconds of the cue period to capture stable MI-related activity. For the Rest condition, the first 3 seconds of the baseline were used, ensuring sufficient distance from both the upcoming cue and the preceding MI and feedback periods. Each epoch was then divided into shorter time segments to enable reliable estimation of power spectral density and functional connectivity. A sliding-window approach was used, with windows of 0.5 seconds (250 samples) and a 36% overlap, corresponding to 90 shared samples between consecutive segments. As a result, each trial was decomposed into eight overlapping windows. PSD was estimated using an autoregressive Burg method. The model order was set to 25, and the spectral estimation was computed using 250 frequency points. A common average reference was applied. The imaginary coherence (ImCoh) was used to quantify phase coupling while removing zero-lag interactions:

$$\text{ImCoh}_{ij}(f) = \left| \frac{\text{Im}(S_{ij}(f))}{\sqrt{S_{ii}(f)S_{jj}(f)}} \right| \quad (1)$$

where $S_{ij}(f)$ is the cross-spectrum between signals i and j at frequency f , and $S_{ii}(f)$ and $S_{jj}(f)$ are the corresponding power spectral densities.

Node strength (NS) was computed for each electrode as the sum of its connectivity values with all other electrodes:

$$NS_i(f) = \sum_{j=1}^N \text{ImCoh}_{ij}(f) \quad (2)$$

where N is the total number of electrodes, and $\text{ImCoh}_{ij}(f)$ denotes the imaginary coherence between electrodes i and j at frequency f . This allows to reduce dimensionality and improving robustness for classification [7].

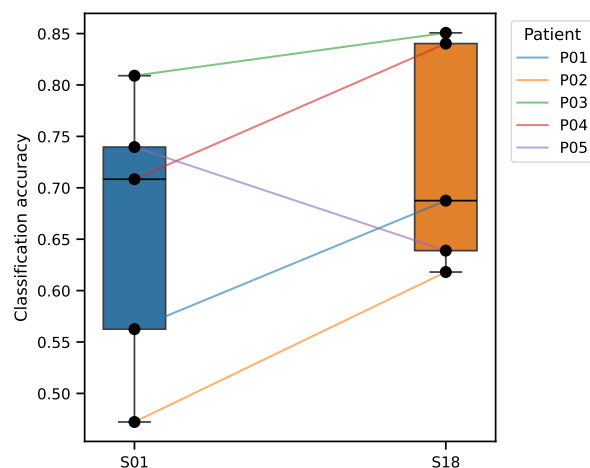


Figure 2: Evolution of classification accuracy between session 1 (S01) and session 18 (S18). Each line represents an individual patient.

Feature selection and decoder training: For each session, the linear discriminant analysis (LDA) decoder was updated at three predefined moments during the session in order to account for non-stationarities in EEG signals, including mental fatigue, strategy changes, and signal drifts. A first decoder was trained after an initial calibration run with a FES at each trial. The model was then retrained after two additional control runs, and a final update was performed after three further runs.

At each training step, candidate features were ranked using the r^2 metric computed between MI and Rest conditions [17]. The most discriminative electrode–frequency features within sensorimotor cortex (FC, C, and CP regions, including channels 1–6 and midline sites) and alpha/beta frequency band (8 to 30Hz) were then selected independently within each feature domain. Separate LDA classifiers were trained using different balanced combinations of PSD and functional connectivity features derived from ImCoh node strength. Specifically, models were computed using equal numbers of features across domains, including configurations with one to three electrode–frequency pairs per modality (i.e., 1 PSD and 1 FC, 2 PSD and 2 FC, and 3 PSD and 3 FC features). The best-performing classifier, based on classification accuracy within the current batch, was selected to control the subsequent runs.

Statistical analysis: To assess the evolution of electrophysiological features across rehabilitation, statistical comparisons between MI and baseline conditions were performed using a non-parametric cluster-based permutation test [18]. Analyses were conducted on calibration runs recorded at the beginning and at the end of the intervention. These runs used standardized MI-FES sequences with systematic feedback, but without online decoder-dependent control, allowing the assessment of MI-related electrophysiological activity under comparable calibration conditions. Test statistics were computed for each electrode–frequency pair and clustered across neighboring electrodes and adjacent frequency bins based

on spatial and spectral adjacency. Statistical significance was estimated by randomly permuting trial labels between conditions and comparing the observed cluster statistics to the null distribution of the maximum cluster statistic.

RESULTS

BCI performance: All patients successfully completed rehabilitation, with no dropouts. As shown in figure 2, mean classification accuracy over subjects increased from 0.66 at Session 1 (range: 0.47–0.81) to 0.73 at Session 18 (range: 0.62–0.85). Four out of five patients showed an improvement in BCI performance over the intervention period, with the largest gains observed in patients with initially lower accuracy. A wilcoxon signed-rank test did not confirm a significant difference due to a small number of patients.

Evolution of functional connectivity features: Figure 3 illustrates representative frequency profiles of PSD and NS-ImCoh during rest and motor imagery, highlighting that connectivity-derived features may provide information partially distinct from power-based markers. Figure 4 shows the evolution of ImCoh node strength modulations between Session 1 and Session 18, with topographic maps displaying r^2 values from calibration runs restricted to statistically significant clusters. At Session 1, no significant clusters were observed. At Session 18, clusters with higher r^2 values were identified over centro-frontal and centro-parietal sensorimotor regions within the alpha–beta range in four of the five patients.

DISCUSSION

This pilot study evaluated the feasibility of an intensive motor imagery BCI protocol integrating functional connectivity features in chronic stroke patients. Participants were recruited consecutively without selection based on BCI aptitude and included individuals with severe distal motor impairment and spasticity, profiles often excluded from BCI studies. Two patients initially failed to reach the conventional threshold for effective BCI control, but by the final sessions all participants achieved classification accuracies above 60% [19]. This progression may reflect preserved neural representations of motor imagery despite the absence of voluntary movement [20, 21].

All participants completed the intervention without dropout. Although the sample size was limited, these observations suggest that intensive BCI-based rehabilitation protocols may be feasible in a broader stroke population than typically included in experimental studies, an important issue for the clinical translation of BCI neurorehabilitation.

Electrophysiological analyses revealed the emergence of functional connectivity modulations after training. No significant clusters were detected during the initial calibration session, whereas significant NS ImCoh changes

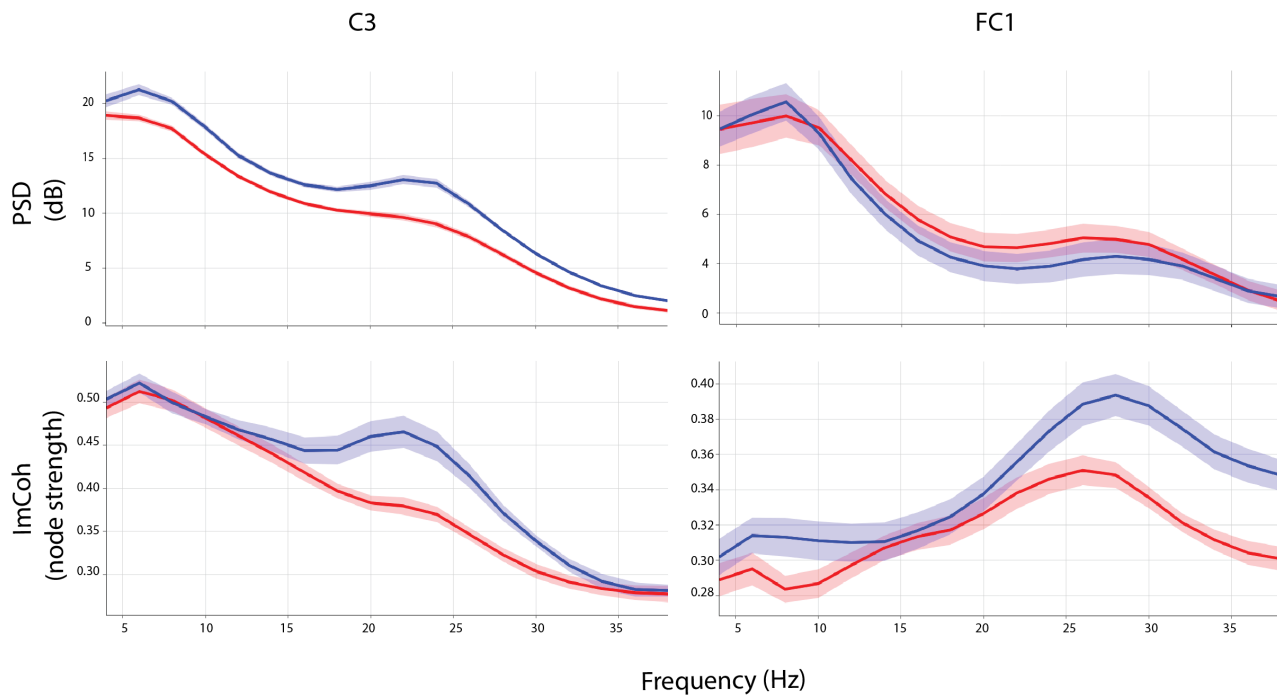


Figure 3: Two representative cases illustrate that PSD and NS-ImCoh can capture both redundant and complementary information. At C3, motor imagery (red) induced alpha/beta ERD in PSD (top row) together with beta-range decreases in NS-ImCoh (bottom row). At FC1, PSD showed mainly beta ERD, whereas NS-ImCoh revealed beta/gamma-range connectivity decreases. Shaded areas indicate standard error of the mean across trials.

were observed at the final session in four out of five patients. Although exploratory, these findings are consistent with previous studies reporting network-level plasticity following BCI-based rehabilitation [3, 22]. Stroke recovery is increasingly understood as a network reorganization process [5, 6], and the progressive emergence of connectivity-based discriminative features may reflect adaptations within distributed sensorimotor networks. These results also support the potential complementarity of connectivity metrics with traditional power-based features commonly used in motor imagery BCIs.

While the FUCONE framework [11] supports the offline relevance of functional connectivity estimators for MI-BCI classification, NS-ImCoh should be considered one possible network-based representation rather than a definitive choice. Other connectivity metrics, such as wPLI, may also reduce sensitivity to volume conduction, whereas source-space reconstruction could improve the anatomical interpretability of the inferred networks. Effective connectivity approaches could further address information flow within post-stroke sensorimotor networks [5]. Beyond pairwise connectivity, recent work on neuronal avalanches [23] also suggests that large-scale spatiotemporal brain dynamics may provide informative biomarkers for BCI training. These complementary approaches will be relevant for future analyses, while NS-ImCoh was selected here as a conservative, low-dimensional, and online-compatible first step.

Next, a second phase of the study will investigate whether combining BCI training with non-invasive brain stimula-

tion can produce synergistic effects on functional connectivity within sensorimotor networks.

CONCLUSION

These preliminary findings support the feasibility of intensive longitudinal network-based MI-BCI rehabilitation in chronic stroke patients with severe motor impairments. Connectivity features derived from imaginary coherence captured task-related sensorimotor network modulations and may represent promising biomarkers for improving BCI decoding and supporting adaptive rehabilitation strategies.

ACKNOWLEDGEMENTS

For this first preliminary work we would like to acknowledge the help of 1) Tristan Venot, Arthur Desbois, and Laurent Hugueville for the engineering support and 2) Marie Villain, Erin Soller, Sergio Talegon-de la Fuente, Celestine Allombert for the experiment assistance. This study was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement N° 864729) and by the Agence Nationale de la Recherche under the France 2030 program (Grant Agreement ANR-23-IACL-0008). The authors declare that they have no competing interests.

REFERENCES

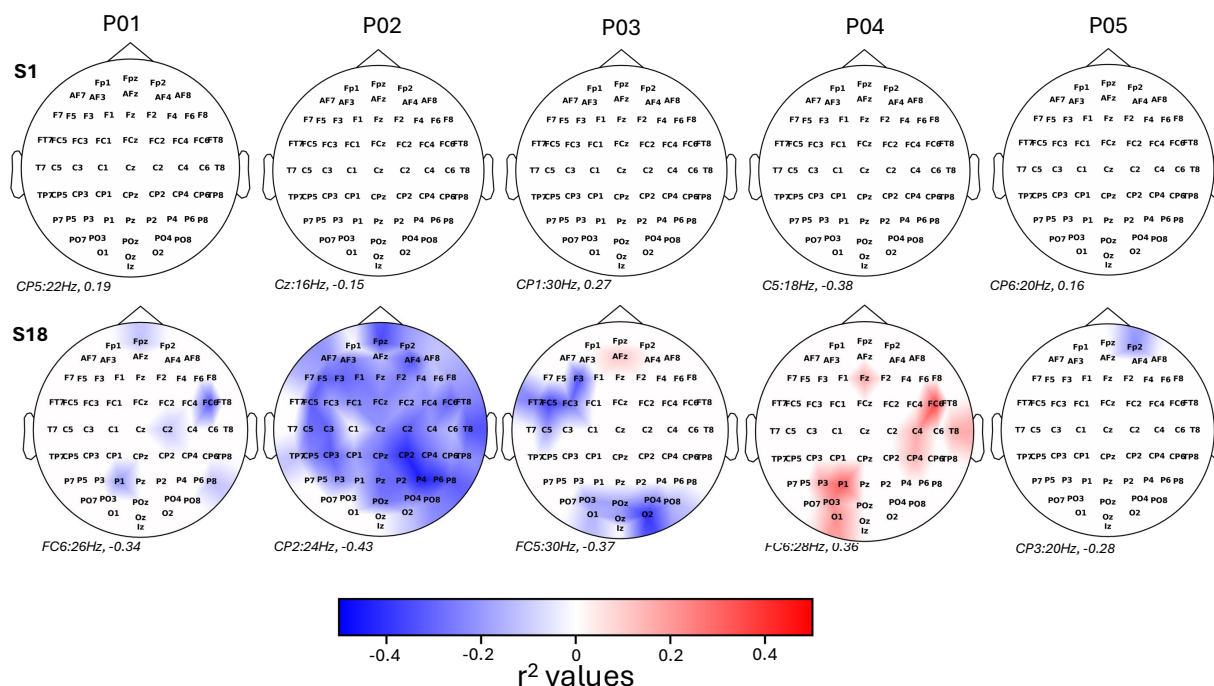


Figure 4: Evolution of the most discriminative brain network features (NS-ImCoh) as revealed by a cluster-based statistical analysis. Only significant clusters are displayed in the topographic maps. For Session 1 and Session 18, the electrode showing the strongest significant modulation was identified, and the corresponding frequency bin was selected. Topographic maps illustrate the spatial distribution of the signed r^2 values at the selected frequency. Warm colors indicate higher NS-ImCoh values during motor imagery relative to rest, whereas cool colors indicate lower NS-ImCoh values during motor imagery.

[1] Mane R, Chouhan T, Guan C. BCI for stroke rehabilitation: Motor and beyond. *Journal of Neural Engineering*. 2020;17(4):041001.
[2] Coscia M et al. Neurotechnology-aided interventions for upper limb motor rehabilitation in severe chronic stroke. *Brain*. 2019;142(8):2182–2197.
[3] Pichiorri F et al. Brain-computer interface boosts motor imagery practice during stroke recovery. *Annals of Neurology*. 2015;77(5):851–865.
[4] Kancheva I, Van Der Salm SMA, Ramsey NF, Vansteensel MJ. Association between lesion location and sensorimotor rhythms in stroke – a systematic review with narrative synthesis. *Neurological Sciences*. 2023;44(12):4263–4289.
[5] De Vico Fallani F, Richiardi J, Chavez M, Achard S. Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2014;369(1653):20130521.
[6] Grefkes C, Fink GR. Reorganization of cerebral networks after stroke: New insights from neuroimaging with connectivity approaches. *Brain*. 2011;134(5):1264–1276.
[7] Gonzalez-Astudillo J, Cattai T, Bassignana G, Corsi MC, De Vico Fallani F. Network-based brain-computer interfaces: Principles and applications. *Journal of Neural Engineering*. 2021;18(1):011001.
[8] Nolte G, Bai O, Wheaton L, Mari Z, Vorbach S, Hallett M. Identifying true brain interaction from EEG data

using the imaginary part of coherency. *Clinical Neurophysiology*. 2004;115(10):2292–2307.
[9] De Vico Fallani F et al. Multiscale topological properties of functional brain networks during motor imagery after stroke. *NeuroImage*. 2013;83:438–449.
[10] Cattai T, Colonnese S, Corsi MC, Bassett DS, Scarano G, De Vico Fallani F. Phase/Amplitude Synchronization of Brain Signals During Motor Imagery BCI Tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2021;29:1168–1177.
[11] Corsi MC, Chevallier S, Fallani FDV, Yger F. Functional Connectivity Ensemble Method to Enhance BCI Performance (FUCONE). *IEEE Transactions on Biomedical Engineering*. 2022;69(9):2826–2838.
[12] Santamaría-Vázquez M, Ortiz-Huerta JH, Martín-Odriozola A, Saiz-Vazquez O. Improvement of Motor Imagination and Manual Ability Through Virtual Reality and Selective and Nonselective Functional Electrical Stimulation: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*. 2024;13:e63329.
[13] Rimbart S, Trocellier D, Lotte F. Impact of the baseline temporal selection on the ERD/ERS analysis for Motor Imagery-based BCI. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE: Sydney, Australia, Jul. 2023, 1–4.
[14] Renard Y et al. OpenViBE: An Open-Source Software Platform to Design, Test, and Use Brain-Computer

Interfaces in Real and Virtual Environments. *Presence: Teleoperators and Virtual Environments*. 2010;19(1):35–53.

[15] Lang CE et al. Dose response of task-specific upper limb training in people at least 6 months poststroke: A phase II, single-blind, randomized, controlled trial. *Annals of Neurology*. 2016;80(3):342–354.

[16] Desbois A, Venot T, De Vico Fallani F, Corsi MC. HappyFeat—An interactive and efficient BCI framework for clinical applications. *Software Impacts*. 2024;19:100610.

[17] Blankertz B, Dornhege G, Krauledat M, Müller KR, Curio G. The non-invasive Berlin Brain–Computer Interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*. 2007;37(2):539–550.

[18] Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*. 2007;164(1):177–190.

[19] Müller-Putz GR, Scherer R, Brunner C, Leeb R, Pfurtscheller G. Better than random? A closer look on BCI results. *International Journal of Bioelectromagnetism*. 2008.

[20] Di Rienzo F, Collet C, Hoyek N, Guillot A. Impact of Neurologic Deficits on Motor Imagery: A Systematic Review of Clinical Evaluations. *Neuropsychology Review*. 2014;24(2):116–147.

[21] Guillot A, Di Rienzo F, Collet C. The Neurofunctional Architecture of Motor Imagery. In: *Advanced Brain Neuroimaging Topics in Health and Disease - Methods and Applications*. InTech, May 2014.

[22] Biasucci A et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nature Communications*. 2018;9(1):2421.

[23] Mannino C, Sorrentino P, Chavez M, Corsi MC. Neuronal avalanches as a predictive biomarker for guiding tailored BCI training programs. *Imaging Neuroscience*. 2026.

MindRun: an Accessible Multi-Player BCI Game for Collaborative Motor Imagery Training

Alessio Palatella¹, Stefano Tortora^{1,2}, Emanuele Menegatti^{1,2}, Luca Tonin^{1,2}, Maryam Alimardani³

¹Department of Information Engineering, University of Padova, Padova, Italy

²Padova Neuroscience Center, University of Padova, Padova, Italy

³Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

E-mail: alessio.palatella@phd.unipd.it

ABSTRACT: Motor imagery (MI) brain-computer interfaces (BCIs) often suffer from limited engagement and reliance on high-cost electroencephalography (EEG) systems. We present MindRun, a continuous MI-BCI game supporting both single and multi-player interactions using a low-cost eight-channel EEG headset. The system was evaluated in a feasibility study with 10 healthy participants who engaged in both collaborative and solo game modes. EEG signals were decoded using a Riemannian geometry-based framework, yielding probabilistic outputs for continuous avatar control. In collaborative mode, the classifier combined individual outputs through a weighted probabilistic fusion to achieve stable joint control. Classification accuracy was significantly above chance level, and participants were able to achieve goal-directed steering in the game. These results demonstrate that continuous single and collaborative MI-BCI gaming is feasible with low-cost EEG hardware, highlighting the importance of interaction design and probabilistic control strategies for accessible real-time MI-BCI systems.

INTRODUCTION

Brain-computer interfaces (BCIs) have advanced considerably over the past decades, enabling applications in communication, control, and neurorehabilitation for both clinical and non-clinical populations [1]. Among the various paradigms, systems based on motor imagery (MI) are the most widely adopted because they allow users to voluntarily modulate their brain activity without external stimulation, which is relevant for applications such as prosthetic control and post-stroke neuro-rehabilitation. However, despite technological advances, user training remains one of the most critical bottlenecks limiting the practical deployment of motor imagery BCI systems [2]. The ability to reliably modulate neural activity is not innate, but requires sustained practice and cognitive effort [3]. Training is often mentally demanding, fatiguing, and frustrating, which can reduce motivation and increase dropout rates [4]. Enhancing user engagement during BCI training is therefore not only advantageous, but essential for long-term usability and widespread adoption. Gamification has emerged as a promising strategy to enhance motivation and improve user experience within

BCI training protocols [5]. Video games inherently incorporate structured goals, adaptive feedback, reward mechanisms, and immersive environments, transforming repetitive training tasks into meaningful interactive experiences [6]. Several studies have demonstrated that game-based BCI training can increase engagement and, in some cases, improve performance outcomes. For instance, Wang et al. [7] developed a virtual reality (VR) game that simulates real-life motor tasks to deliver adaptive MI-BCI neurofeedback, reporting higher event-related desynchronization (ERD) activation and improved classification accuracy. Similarly, Bermúdez i Badia et al. [8] introduced a VR-based MI-BCI system with personalized motor tasks and adaptive training, reporting higher levels of user enjoyment and acceptance. Mercado et al. [9] developed *FarmerKeeper*, a runner-style BCI video game designed for neurofeedback training in children with autism, demonstrating improved usability and user experience compared to a commercial BCI game while supporting sustained attention during sessions.

Despite these promising developments, most approaches focus primarily on single-user scenarios and individual training tasks, with limited consideration of social interaction during training. Social interaction is a powerful driver of motivation in learning and skill acquisition contexts, fostering emotional engagement, sustained attention, and intrinsic motivation [10]. Collaborative training introduces shared objectives, coordination, and interdependence, which can enhance both enjoyment and persistence. Therefore, integrating social collaboration into the BCI training protocols may represent an important step toward more sustainable and user-centered training approaches. However, empirical investigations and technological implementations in this area remain scarce. Bonnet et al. [11] were among the first to explore collaborative BCI paradigms by designing *BrainArena*, a two-player football-style video game fully controlled by MI-BCI. In this game, each participant moved a ball left or right by imagining the corresponding hand movements. The classifier outputs of both players were then combined to determine the ball's movement, supporting both collaborative and competitive modes. Their results indicated that multi-user BCI setups can be more engaging than tra-

ditional single-user protocols, with comparable levels of performance and motivation across interaction modes. Building on these early findings, Palatella et al. [12] investigated a collaborative training strategy based on a collaborative brain-machine interface (cBMI), in which two participants learned to operate a real robotic arm through cooperative control. Experiments involving three pairs of subjects over a five-day training period provided preliminary evidence that a collaborative framework may accelerate skill acquisition and help reducing frustration levels typically experienced by naive users.

Despite these explorations, accessibility remains a major barrier in socially engaging BCI research. The adoption of multi-user BCIs is particularly limited, in part because their implementation requires hyperscanning setups, often involving two research-grade electroencephalography (EEG) devices [13], which are typically costly and time-consuming to configure. To enable a larger adoption of collaborative BCI devices, reliable consumer-grade EEG systems and classifiers that can function effectively with a reduced number of electrodes are needed. In this work, we introduce *MindRun*¹, a video game that can be controlled using single or dual low-cost EEG headsets equipped with only eight channels. Through optimized signal processing and interaction design, we demonstrate that both single-user and collaborative BCI gameplay through MI can be achieved effectively despite the reduced channel density. The plug-and-play nature of the commercial headset, combined with our classifier architecture, substantially reduces setup time and technical complexity, facilitating faster deployment and improving accessibility beyond specialized research facilities.

The primary contribution of this study is a proof-of-concept continuous-control MI-BCI gaming architecture explicitly designed to support both individual and cooperative interaction using low-density EEG hardware. To achieve stable shared control in collaborative scenarios, we implemented a probabilistic multi-user fusion strategy based on softmax-weighted Riemannian decoding outputs, ensuring robust performance without degradation. By combining accessible hardware, cooperative gameplay, and task-driven assessment, this work advances BCI research toward scalable, ecologically valid, and user-centered real-world applications.

SYSTEM DEVELOPMENT

BCI Game Description: *MindRun* simulates a snowball rolling down a mountain valley. Along the path, rows of trees appear as obstacles, leaving a gap on either the right or left side of the screen. Participants must guide the snowball through the gap by performing the MI task. The game is organized into trials, each including one obstacle segment spanning a floor width of 60 *Uu* (Unity units). The snowball's maximum velocity is capped to ensure a minimum inter-trial interval of 10 s. The open-

ing between the trees corresponds to 20% of the floor width. Its outer boundaries never exceed the floor limits, while its inner boundaries never cross the center of the screen, ensuring that the opening remains entirely on either the left or the right side while maintaining a constant width. Within each trial, obstacle rows are presented in randomized order, with minor variations in scale, rotation, number of trees and position to enhance realism.

To prepare participants for each trial, three thin colored lines (red, green, and blue) are displayed on the ground. A fog wall positioned at 200 *Uu* restricts the visibility of subsequent segments. The red line, located 30 *Uu* after each obstacle row, signals the beginning of a new trial. After an additional 30 *Uu*, a green line appears and serves as a fixation cue to prepare participants for the MI task [14]. Following a further 60 *Uu*, the blue line marks the onset of the task (Figure 1). At this point, participants must steer the snowball toward the opening located 180 *Uu* ahead. The fog wall is positioned beyond the obstacle distance so that limited visibility becomes possible only a little before crossing the blue line (Figure 1). This configuration prevents anticipatory artifacts during the MI phase while still providing a preparatory cue before task initiation.

After each trial, if the *Continuous Feedback (CF) Reset* option is enabled in the main menu (Figure 1), the snowball automatically returns to the center of the screen. Under this configuration, lateral movement is disabled until the blue line has been crossed, otherwise the subjects can steer the snowball for all the run duration. All data acquisitions were performed with this option enabled.

Successful passage of the snowball through the opening triggers positive visual feedback (i.e., an animation of a purple explosion). Conversely, collisions with the trees slightly push the snowball backward, reducing its velocity and initiating a blinking animation that persists until the red line is crossed. During this phase, the snowball can pass through the trees without further interaction. This mechanism prevents the snowball from becoming stuck, however, if contact occurs while correctly passing through the opening, no backward displacement is applied and the snowball maintains its velocity.

During calibration runs, the snowball autonomously moves toward the opening, whereas during evaluation runs participants actively control its movement. In the evaluation mode, the snowball increases in size by 5% after each successful passage, up to a maximum of 140% of its original size, thereby progressively increasing task difficulty as performance improves. Conversely, following a collision, the snowball decreases in size by 5%, but never below the default diameter of 1.2 *Uu*.

EEG Acquisition: Brain signals were acquired using the 8-channel Unicorn Hybrid Black EEG headset (g.tec Medical Engineering, Schiedlberg, Austria) at a sampling rate of 250 Hz. Electrodes were positioned over the sensorimotor cortex according to the international 10–20 EEG system at locations *Fz*, *C3*, *Cz*, *C4*, *Pz*, *PO7*, *Oz*, and *PO8*. For the collaborative gameplay, two identical EEG

¹A playable version of *MindRun* is available at <https://github.com/Brain-Machine-Interface-VU/MindRun>.

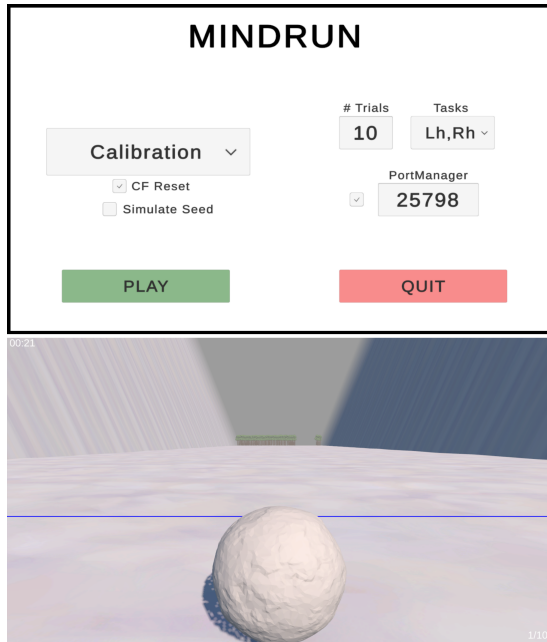


Figure 1: (Top) MindRun main menu. The interface allows users to select the run type, enable continuous feedback reset, recreate a game by specifying a seed in the PortManager, set the number of trials, choose the MI task, and configure the connection port for external scripts. If the connection manager is disabled, the game can be played using the arrow keys. (Bottom) Screenshot of a continuous feedback session. The top-left corner displays the elapsed time, while the bottom-right corner shows the number of completed trials out of the total.

headsets with the same configuration were used.

Riemannian BCI Classifier: A two-class MI paradigm was used, in which the participants performed kinesthetic imagination of the movements of the right and left hand. EEG signals were processed and classified using a Riemannian geometry-based pipeline designed for continuous control. To enhance local sensorimotor activity, EEG signals were first spatial filtered using a Laplacian filter. For each electrode, the spatial neighborhood was defined by the immediate adjacent electrodes along the horizontal and vertical axes of the grid. Because of the electrode layout, connections were restricted to directly adjacent electrodes and were not allowed to cross intermediate rows. Subsequently, the signals were bandpass filtered using a second-order Butterworth filter between 6 Hz and 24 Hz, with a stopband between 14 Hz and 18 Hz, to isolate the α and β frequency bands relevant for MI. The filtered signals were segmented using a sliding window of 1 s with a shift of 40 ms (update rate of 25 Hz).

For each window, spatial covariance matrices (SCM) were computed and regularized using Ledoit–Wolf normalization [15] to ensure that the resulting matrices were well-conditioned and remained symmetric positive definite (SPD). These SCMs were then used to train a Riemannian classifier based on the Fisher geodesic minimum distance to the mean (FgMDM) [16].

To enable continuous horizontal control of the snowball in the game, the classifier outputs were smoothed using

an exponential integration function [17], with smoothing parameter α ranging between 0.93 and 0.99. Due to the intrinsic nature of the classifier, which converts distances to class centroids into posterior probabilities, and the limitation of the exponential smoothing function, the output values rarely reach extreme probabilities close to 0 or 1. Consequently, to allow the snowball to reach the extreme horizontal positions, a hard decision rule was applied to the classifiers output, in which the class with the highest posterior probability was assigned probability 1 and the other class probability 0.

To prevent small fluctuations around the decision boundary from producing abrupt directional changes and jittery movements of the snowball, an additional thresholding rule was introduced. Specifically, probabilities below 0.51 were considered uncertain and reassigned to 0.50, thus directing the snowball toward the center of the screen and stabilizing its motion. With this approach, the value of $(1 - \alpha)$ in the integration function represents the contribution of each window to the movement of the snowball toward the right or left side.

In the collaborative condition, each subject retained an individual classifier. Their outputs were combined using a weighted mean before applying the same hard-decision remapping and integration procedures. The weighting strategy was designed to promote robustness and decoding stability rather than maximize ensemble accuracy. Subject-specific weights were derived from a composite performance score S_i computed on calibration data using probabilistic performance metrics including negative log-likelihood, Brier score, AUC, and expected calibration error, which are widely used measures for evaluating probabilistic classifiers [18, 19]. The collaborative weights were then obtained using a softmax normalization:

$$w_i = \frac{\exp(\beta S_i)}{\sum_{j=1}^M \exp(\beta S_j)},$$

where M is the number of users, S_i is the subject score and β controls the sharpness of the weighting. This strategy ensures stable and normalized weighting while preventing hard selection of a single participant.

VALIDATION STUDY

Participants: A total of 10 healthy participants (mean age: 23 ± 3 years, 4 males and 6 females) were recruited and randomly assigned to two groups. The first group included six players, randomly paired into three dyads for collaborative play. The second group consisted of four single players for solo play. None of the participants had previous experience with a BCI system. All participants completed two training sessions conducted on separate days with a 3-day interval between sessions. Written informed consent was obtained from participants after providing detailed information about the study goals, data management policy, and privacy protection measures. The study was approved by the Research Ethics Review Committee of the Faculty of Science, Vrije Uni-

versiteit Amsterdam.

Experimental Procedure: The experimental protocol comprised two sessions, separated by three days. Each session consisted of 10 runs of the proposed game 2, with each run including 10 trials, equally and randomly distributed between right- and left-hand MI. Two types of runs were defined: calibration runs, during which data were collected to train the decoder and feedback was systematically positive, and evaluation runs, in which the trained decoder was integrated into the proposed BCI framework, allowing participants to control the in-game avatar (i.e., snowball).

Session 1 was identical for all 10 participants who all played the game individually. Six calibration runs were first conducted to train the Riemannian BCI decoder, followed by four evaluation runs to assess its performance. In Session 2, participants were divided into two groups. Depending on the assigned group, they performed both the calibration and evaluation phases either individually or collaboratively. The session began with four additional calibration runs. The decoder was then retrained using the six calibration runs from Session 1 together with the four newly recorded runs to account for potential inter-session feature shifts. Participants subsequently completed six evaluation runs.

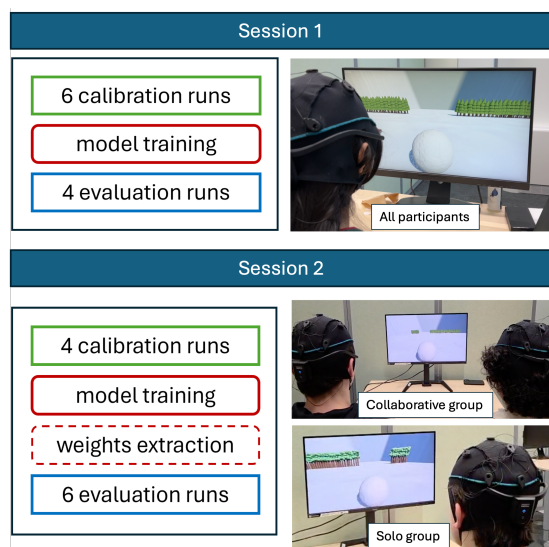


Figure 2: Experimental procedure illustrating distribution of calibration and evaluation runs across the two sessions. Weights extraction for BCI classification was applied only to the collaborative group. Videos of calibration and collaborative evaluation runs are available at <https://cloud.dei.unipd.it/index.php/s/gHiaG3JJjSaGkag>

Evaluation metrics: To evaluate the system and gameplay, two complementary sets of metrics were defined: in-game performance and BCI decoding accuracy. In-game performance was quantified as the lateral distance between the snowball and the center of the target gap at the moment of crossing, normalized by the total floor width. Since the gap corresponds to 20% of the floor width and is fully located on either the left or right side,

successful passage through the gap corresponds to normalized distances below 10%. The performance of the BCI decoder was assessed using three metrics. *Trial success rate* was defined as the percentage of trials in which the snowball successfully passed through the opening. This metric captures the effective control achieved by the participant within the game environment. *Sample accuracy* was defined as the proportion of covariance matrices correctly assigned to the intended MI class during the continuous feedback period over the threshold rule. Finally, *rejection rate* was calculated as the proportion of classifier outputs whose posterior probability fell below the uncertainty threshold of 0.51. Lower rejection rates, therefore, indicate more confident classifier outputs and stable control signals.

RESULTS

Figure 3 summarizes in-game performance and BCI decoding accuracies for collaborative and single players across the two sessions. No significant differences were observed between Session 1 and Session 2 in window accuracy (Wilcoxon signed-rank test: $p = 0.85$), trial success rate ($p = 0.32$), or distance-to-gap ($p = 0.70$), indicating stable performance for all participants.

In-Game Performances: Across both sessions, median distances in both solo- and collaborative-play conditions were generally close to the 10% gap threshold, and no statistically significant difference was observed between the two conditions (Mann–Whitney U tests: $p = 0.76$). This indicates that users were able to intentionally control the snowball’s trajectory and achieve goal-directed steering in both modes of the game.

BCI Decoding Accuracy: BCI classification accuracy remained consistently above the chance level (50%) across participants and sessions, with a mean of 0.61 ± 0.07 (Wilcoxon signed-rank test comparing accuracies vs. chance: $p = 0.002$, $r = 0.89$). All participants performed above chance level, indicating that the Riemannian geometry-based decoder reliably discriminated short-time MI states. In contrast, the overall trial success rate remained moderate, consistent with the gap distance results, particularly during Session 1, reflecting the stricter nature of trial metrics. Brief classification errors or transient uncertainty periods can invalidate an entire trial, making trial success rate a conservative measure of control performance. Rejection rates were moderate to high, effectively reducing rapid oscillations near the decision boundary. Crucially, elevated rejection rates did not impair task performance, as indicated by the distance-to-gap metric.

DISCUSSION

The purpose of this study was not to optimize MI classification accuracy, but to introduce and validate an integrated continuous MI-BCIs game framework supporting both single and collaborative interaction using a com-

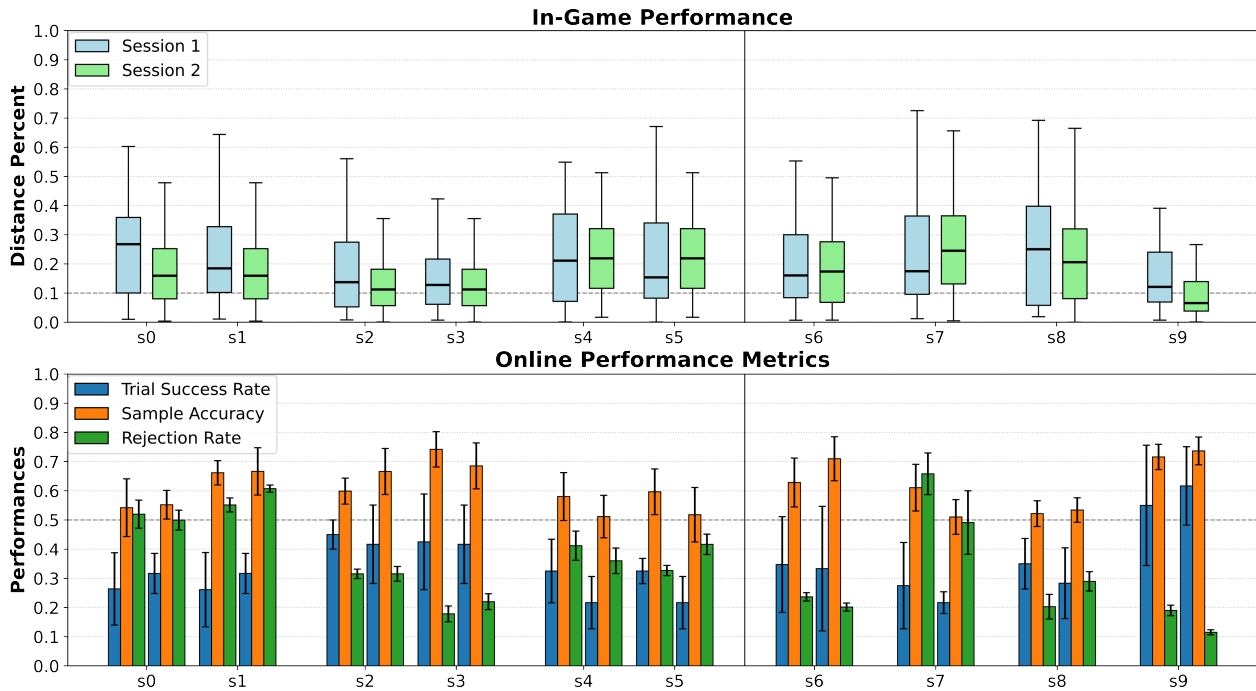


Figure 3: In-game performances (top) and BCI decoding accuracies (bottom) for collaborative dyads (s0–s1, s2–s3, s4–s5) and single players (s6, s7, s8, s9) across the two experimental sessions. In-game performance was quantified by the lateral distance between the snowball and the center of the target gap, expressed as a percentage of the total floor width (successful passage at 10%) (Bottom) BCI decoding performance was measured using three metrics: trial success rate, sample accuracy (chance level at 50%), and rejection rate.

mercially available EEG system. The observed above-chance sample accuracy demonstrates that a low-density consumer-grade EEG headset provides sufficient signal fidelity for Riemannian MI decoding. Game performance and BCI decoding results collectively show that MI control can be reliably translated into stable, continuous avatar steering in an interactive gaming environment. Furthermore, the collaborative fusion strategy enabled shared control without compromising system stability, confirming the feasibility of multiplayer MI-BCI interaction within a lightweight configuration. These findings confirm that functional continuous MI control is achievable with reduced channel density; however, direct comparisons with high-density systems are needed to objectively assess efficiency and potential performance trade-offs. While limited due to the different application scenarios and control paradigms, the results of current study can be compared to those of Palatella et al. [12] and Bonnet et al. [11], both introducing a collaborative MI-BCI using research fidelity EEG. When considering sample accuracy, the results obtained in this study are comparable to those reported by Bonnet et al. [11]. Specifically, they reported that overall sample accuracies across subjects ranged from 67% to 81%, with considerable variability between subjects. In contrast, Palatella et al. [12] reported trial accuracies that were highly dependent on the session, with a median value of approximately 45% and a range from 17% to 73%, which are also comparable to the results of the present study. However, it should be noted that our framework relies

on a more lightweight configuration, using a low-density EEG headset with fewer electrodes and a lower sampling rate. An additional analysis was performed by recomputing offline classification after removing the Fz channel, which is more susceptible to ocular contamination. As accuracies and rejection rates remained unchanged across participants (61.7% vs. 61.6% paired t-test: $t = 0.21$, $p = 0.84$ and 35.5% vs. 35.1%; paired t-test: $t = 0.35$, $p = 0.73$), the results suggest that decoding performance was not driven by EOG-related activity.

Importantly, decoding performance translated into measurable goal-directed in-game behavior. The results of the distance-to-gap metric confirm that participants were able to generate sustained and goal-directed steering rather than isolated classification events. This distinction is critical because successful BCI gaming requires temporal stability and controllability [5], not merely statistical separability of classes. Particularly in the collaborative gameplay, the probabilistic fusion of individual BCI outputs proved technically viable. The weighted softmax strategy maintained stable shared control without collapsing into dominance by a single user or introducing instability. This result supports the feasibility of multiplayer MI-BCI paradigms beyond competitive designs, extending toward cooperative interaction models.

The lower trial success rate compared to the sample accuracy should be interpreted within the task design context. The 20% gap width imposed strict spatial precision, making trial success an intentionally demanding outcome measure. Consequently, trial success rate functions as

a conservative measure of sustained control rather than instantaneous decoding capability. Similarly, higher rejection rates result from explicit design choices that prioritize smooth interaction over aggressive classification. In continuous control environments, stability and predictability are often more important than maximizing raw classifier output. The integration and thresholding mechanisms were therefore implemented as interaction design decisions aimed at improving control stability and usability. Nevertheless, the rejection mechanism does not resolve the intermediate control state. The rejection interval only filters uncertain outputs and is not a true “rest” class. Users must still maintain outputs near the decision boundary, representing a framework limitation.

By integrating social gameplay elements with an accessible hardware platform, this work contributes toward more scalable and ecologically valid BCI systems. The findings challenge the assumption that continuous BCI control necessarily requires high-density research-grade EEG setups and instead highlight the importance of interaction, classification robustness, and usability in real-world applications. Future research should extend this framework to larger cohorts and longitudinal protocols to evaluate learning effects, motivational impact, and the long-term benefits of collaborative training paradigms. To this end, we are currently collecting data from a larger sample size across 12 sessions, enabling the analysis of longitudinal learning dynamics between single-player and collaborative MI-BCI training conditions.

ACKNOWLEDGMENT

The work is partially funded by the Piano Nazionale di Ripresa e Resilienza (PNRR), Project PE8 “AGE-IT” Spoke 9 and Progetti di Rilevante Interesse Nazionale (PRIN 2022, 2022BCZ52A)-Ministero dell’Università e della Ricerca. This work reflects only the authors’ views and opinions, neither the European Union and European Commission can be considered responsible for them.

REFERENCES

- [1] Hramov A, Maksimenko V, Pisarchik A. Physical principles of brain–computer interfaces and their applications for rehabilitation, robotics and control of human brain states. *Physics Reports*. 2021.
- [2] Mladenović J. Standardization of protocol design for user training in eeg-based brain–computer interface. *Journal of Neural Engineering*. 2020;18.
- [3] Tortora S et al. Neural correlates of user learning during long-term bci training for the cybathlon competition. *Journal of NeuroEngineering and Rehabilitation*. 2022;19.
- [4] Roc A et al. A review of user training methods in brain computer interfaces based on mental tasks. *Journal of Neural Engineering*. 2020;18.
- [5] Atilla F, Postma M, Alimardani M. Gamification of motor imagery brain-computer interface training protocols: A systematic review. *Computers in Human Behavior Reports*. 2024;16:100508.
- [6] Vermeir JF, White M, Johnson D, Crombez G, Ryckeghem DVV. The effects of gamification on computerized cognitive training: Systematic review and meta-analysis. *JMIR Serious Games*. 2020;8.
- [7] Wang K et al. Adaptive neurofeedback training using a virtual reality game enhances motor imagery performance in brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2025;33:2956–2966.
- [8] Badia SBI, Morgade AG, Samaha H, Verschure P. Using a hybrid brain computer interface and virtual reality system to monitor and promote cortical reorganization through motor activity and motor imagery training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2013;21:174–181.
- [9] Mercado J, Espinosa-Curiel I, Escobedo L, Tentori M. Developing and evaluating a bci video game for neurofeedback training: The case of autism. *Multimedia Tools and Applications*. 2018;78:13675–13712.
- [10] Wang C, Huang L. A systematic review of serious games for collaborative learning: Theoretical framework, game mechanic and efficiency assessment. *Int. J. Emerg. Technol. Learn.*. 2021;16.
- [11] Bonnet L, Lotte F, Lécuyer A. Two brains, one game: Design and evaluation of a multiuser bci video game based on motor imagery. *IEEE Transactions on Computational Intelligence and AI in Games*. 2013;5(2):185–198.
- [12] Palatella A, Forin P, Tortora S, Menegatti E, Tonin L, et al. An alternative training protocol for a motor imagery bmi based on a collaborative approach. In: -. 2024.
- [13] Xue Z, Zhang Y, Li H, Chen H, Shen S, Du H. Instrumentation, measurement, and signal processing in electroencephalography-based brain-computer interfaces: Situations and prospects. *IEEE Transactions on Instrumentation and Measurement*. 2024;73:1–28.
- [14] Tonin L, Leeb R, Sobolewski A, Millán JdR. An online eeg bci based on covert visuospatial attention in absence of exogenous stimulation. *Journal of neural engineering*. 2013;10(5):056007.
- [15] Chen Y, Wiesel A, Eldar YC, Hero AO. Shrinkage algorithms for mmse covariance estimation. *IEEE transactions on signal processing*. 2010;58(10):5016–5029.
- [16] Barachant A, Bonnet S, Congedo M, Jutten C. Riemannian geometry applied to bci classification. In: *International conference on latent variable analysis and signal separation*. 2010, 629–636.
- [17] Tonin L, Bauer FC, Millán JdR. The role of the control framework for continuous teleoperation of a brain–machine interface-driven mobile robot. *IEEE Transactions on Robotics*. 2019;36(1):78–91.
- [18] Filho TS, Song H, Perello-Nieto M, Santos-Rodríguez R, Kull M, Flach PA. Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*. 2021;112:3211–3260.
- [19] Waghmare KG, Ziegel J. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and Its Application*. 2025.

Benchmarking Feature-Level and Decision-Level Fusion Strategies for Hybrid EEG-fNIRS Cognitive State Detection

Ainara Carpio-Chicote^{1,2,3}, Tudor-Andrei Deaconu⁵, Enrique J. Gómez^{1,2,3,4}, George Kontaxakis⁵

¹Biomedical Engineering and Telemedicine Centre, ETSI Telecomunicación

²Center for Biomedical Technology, Universidad Politécnica de Madrid

³imas12 Hospital 12 de Octubre Madrid, Spain

⁴CIBER-BBN, Madrid, Spain

⁵Information Processing and Telecommunications Center, ETSI Telecomunicación, Universidad Politécnica de Madrid

Email: ainara.carpio.chicote@fgupm.upm.es

ABSTRACT: The integration of Electroencephalography (EEG) and functional Near-Infrared Spectroscopy (fNIRS) has emerged as a robust framework for passive Brain-Computer Interface (pBCI) applications, particularly in the field of cognitive workload monitoring. This study benchmarks feature-level (early) and decision-level (late) fusion strategies using a subject-independent Leave-One-Subject-Out (LOSO) validation framework. Leveraging an open-access dataset of 26 participants performing an n-back task (0-back vs. 2/3-back), we developed a MATLAB-Python pipeline to contrast unimodal and multimodal performance. Our results demonstrate that while EEG provides high temporal responsiveness, its inter-subject variability limits generalization (65.9% LOSO accuracy). In contrast, fNIRS offers stable hemodynamic markers (76.9% accuracy). Multimodal integration significantly outperformed unimodal baselines, with decision-level fusion achieving the highest subject-independent performance (85.3% accuracy, 0.977 AUC) using an SVM-Poly meta-classifier. These findings demonstrate that hybrid EEG-fNIRS integration provides a more reliable framework than unimodal approaches for detecting distinct cognitive states. While further validation is needed for granular workload levels, this study establishes a robust baseline for the development of future pBCI applications in operational environments. **KEYWORDS:** Hybrid EEG-fNIRS, Multimodal Fusion, Cognitive Workload.

INTRODUCTION

Monitoring cognitive states in high-risk operational scenarios is a fundamental challenge for the development of reliable Brain-Computer Interfaces (BCIs) [1] [2]. In these environments, the saturation of working memory (WM) resources can lead to critical performance degradation [1]. While traditional BCIs focus on active control, passive BCIs (pBCIs) aim to monitor mental states—such as cognitive workload—without requiring conscious effort from the user [3]. However, the deployment of pBCIs

in real-world settings is often hindered by the limitations of unimodal recordings. Electroencephalography (EEG), while providing excellent temporal resolution, is highly susceptible to artifacts, whereas functional near-infrared spectroscopy (fNIRS) offers hemodynamic stability but suffers from an inherent metabolic delay [4], [5], [6].

Hybrid EEG-fNIRS systems have been proposed to overcome these modality-specific weaknesses by leveraging neurovascular coupling (NVC) [7]. Although the potential of hybrid systems is well-documented, the optimal architectural strategy for integrating these different data streams, specifically regarding feature-level versus decision-level fusion, remains a key area of investigation in neural engineering [8]. This paper establishes a performance benchmark for multimodal fusion strategies by evaluating their ability to detect working memory engagement using the n-back test [9]. By contrasting minimal (0-back) versus high (2/3-back) cognitive demand [10], the aim is to identify the most robust integration framework, providing a baseline for the transition of pBCI technologies from controlled laboratory settings to demanding operational applications.

MATERIALS AND METHODS

Dataset and Experimental Paradigm:

We utilized the open-access hybrid EEG-fNIRS dataset provided by Shin et al. (2018) [11]. The data consists of simultaneous recordings from 26 healthy participants performing a mental workload task based on the n-back paradigm.

EEG activity was recorded at 1,000 Hz using 30 active electrodes (BrainAmp) following the 10-5 system. Simultaneously, fNIRS data were acquired at 10.4 Hz (NIRScout) through 36 channels formed by 16 sources and 16 detectors. Sensors were strategically distributed across frontal, motor, parietal, and occipital regions, with a high-density focus on the prefrontal cortex (16 fNIRS channels) to monitor executive demand (see Fig. 1).

The n-back paradigm is widely used to assess working

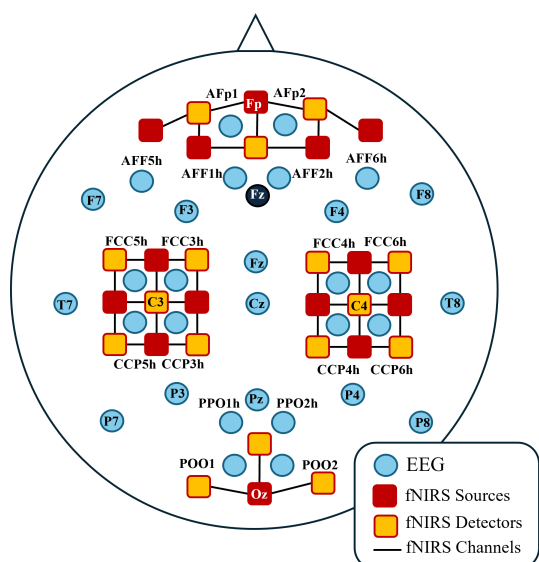


Figure 1: Hybrid EEG-fNIRS montage

memory, requiring participants to determine whether the current stimulus matches one presented n steps earlier in the sequence (see Fig. 2). Cognitive load increases with n : the 0-back condition serves as a low-load baseline, while 2-back and 3-back impose higher demands through continuous maintenance and updating of information. The experimental design included three levels of difficulty:

- 0-back (Control): Minimal WM demand, used as the baseline.
- 2-back and 3-back (High Load): Increased demand for maintenance, updating, and inhibition of information.

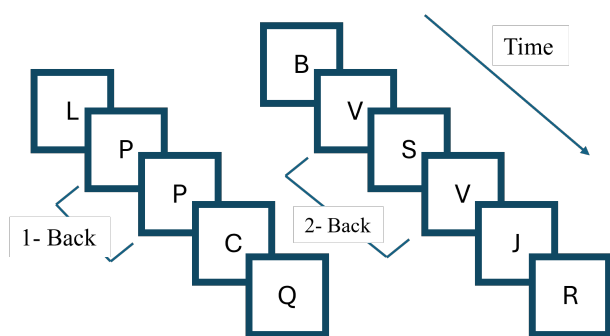


Figure 2: N-back paradigm

For the purpose of this study, we defined a binary classification task comparing the 0-back condition (low load) against the combined 2-back and 3-back conditions (high load). Due to the differences in intensities of the difficulty of the levels, this approach maximizes the physiological contrast, allowing for a clear evaluation of multimodal integration strategies.

Preprocessing Pipeline:

The preprocessing pipelines for both EEG and fNIRS were conducted using the BBCI Toolbox in MATLAB, and are described below.

EEG Preprocessing:

The raw EEG signal (1,000 Hz) was band-pass filtered between 0.5 and 40 Hz using a zero-phase Butterworth filter. Then, artifact rejection was performed via Independent Component Analysis (ICA) using the FastICA algorithm [12]. To ensure objectivity, artifactual components were automatically identified and removed using the Multiple Artifact Rejection Algorithm (MARA) [13], supplemented by statistical thresholds (variance/kurtosis > 3 SD). Finally, the cleaned signal was re-referenced to a Common Average Reference (CAR) [14]. The summarized pipeline can be seen in Fig. 3.

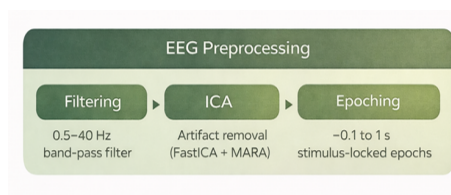


Figure 3: EEG Processing Pipeline

fNIRS Preprocessing:

The fNIRS raw light intensities had already been transformed into relative concentration changes of oxygenated (ΔHbO) and deoxygenated (ΔHbR) hemoglobin via the Modified Beer-Lambert Law (MBLL) and downsampled to 10 Hz.

To eliminate cardiac and respiratory artifacts, a 4th-order zero-phase IIR low-pass filter with a 0.2 Hz cutoff was applied. Each block-level epoch was baseline-corrected using a 3-second window (from -5 s to -2 s prior to block onset) to remove slow physiological drifts and isolate task-related hemodynamic responses. The summarized fNIRS preprocessing pipeline can be seen in Fig. 4.

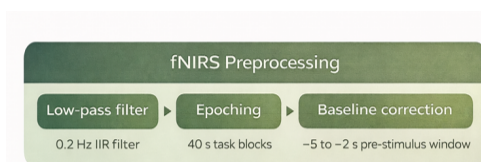


Figure 4: fNIRS Processing Pipeline

Segmentation and Multi-level Alignment:

Due to the differences in temporal resolutions, a block-aggregation strategy was adopted. For EEG, spectral features (PSD) and statistical metrics were first computed for each trial-locked epoch (-0.1 to 1.0 s). These features were then averaged across all trials within a 40 s block. This approach ensures that high-frequency spectral information (e.g., beta and gamma bands) is preserved before aggregation, providing a synchronized feature set that matches the 40 s fNIRS block duration.

Feature Extraction: Following block-level aggregation, a multi-domain feature set was extracted to capture the complex dynamics of cognitive load. For each task block, a unified feature vector was created.

- EEG Features: Three domains were calculated for each channel:
 - Spectral: Power Spectral Density (PSD) in Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz) and Gamma (30–40 Hz) bands using Welch’s method.
 - Temporal: P300 component amplitude and latency within a 300–600 ms post-stimulus window.
 - Statistical: Mean, variance, skewness, and kurtosis.
- fNIRS Features: For each channel and chromophore (ΔHbO and ΔHbR), four metrics were extracted: mean concentration, standard deviation, Area Under the Curve (AUC), and the hemodynamic slope estimated via first-order polynomial fitting.

Neurovascular Coupling (NVC) Analysis:

To ensure that the classification accuracy comes from legitimate physiological changes rather than statistical artifacts, a Neurovascular Coupling (NVC) analysis was performed as a biological validation step [15].

- Correlation Analysis: Spearman’s correlation (r_s) was calculated between EEG Alpha power (a marker of cortical inhibition) and HbO concentration in the prefrontal cortex. The aim was to identify the classic inverse relationship: a decrease in Alpha power (event-related desynchronization) correlated with an increase in the hemodynamic response during high-load (2/3-back) conditions.
- Granger Causality: A Granger Causality test was applied to verify that changes in neuronal electrical activity temporally precede fluctuations in hemoglobin metabolism with a physiological delay of 2–5 seconds, validating the neurovascular activation chain.

To account for the metabolic delay of the fNIRS signal, the correlation analyses considered a physiological lag of 2–5 seconds. The use of 40 s blocks further stabilizes the comparison, focusing on sustained hemodynamic trends rather than instantaneous changes.

Classification Framework and Fusion Strategies:

The classification pipeline was implemented in Python using scikit-learn and XGBoost. To ensure a rigorous comparison between unimodal and multimodal approaches, a consistent evaluation protocol was performed in all experiments.

Unimodal Baselines:

Two independent pipelines were developed for EEG and fNIRS [16].

To avoid overfitting due to high-dimensional data, a supervised feature selection using SelectKBest with an ANOVA F-test was implemented. This allowed identifying the most discriminative features (those with the highest variance between cognitive load levels). To reduce dimensionality and mitigate overfitting, EEG features were

selected using an ANOVA F-test, retaining features with a significance threshold of $p < 0.05$.

To ensure robustness, a grid search for hyperparameter tuning was performed, evaluating a diverse set of classifiers. The specific configurations for the best performing models are summarized in Table 1.

Table 1: Classification Models and Hyperparameter Configurations

Model	Core Parameters	Details/Kernel
LDA	Solver: SVD	No shrinkage.
RF	Estimators: 100	Gini; Max depth: None.
XGBoost	LR: 0.1	Max depth: 6; Sub. 0.8.
SVM-RBF	$C \in [0.1, 10]$, γ	RBF kernel.
SVM-Poly	Deg: 3; $C = 1.0$	Poly (Meta-classifier).

Feature-Level Fusion (Early Fusion):

EEG and fNIRS feature vectors were concatenated into a high-dimensional joint representation [17]. To prevent overfitting and exploit cross-modal interactions, we applied dimensionality reduction and hyperparameter tuning within the training folds.

Decision-Level Fusion (Late Fusion):

This modular approach combined the probability outputs of the best-performing unimodal classifiers. These probabilities were used as inputs for a meta-classifier (e.g., SVM-Poly), which generated the final prediction [18]. This strategy preserves the unique characteristics of each signal while leveraging their complementary predictive power.

Validation Protocol:

To ensure a comprehensive evaluation, we employed a dual-validation scheme.

First, a stratified block-level hold-out split (80/20) was used to establish a performance baseline for the hybrid features. To stabilize the high-dimensional feature space, a global scaling transformation (Z-score) and supervised feature selection (SelectKBest, $p < 0.05$) were applied based on the training pool distribution.

Second, to assess subject-independent generalization, a LOSO cross-validation was performed. In this scheme, the model was iteratively trained on $N - 1$ subjects and tested on the remaining participant. This protocol evaluates the BCI’s ability to handle inter-subject variability, which is the primary challenge in real-world deployment. The overall classification and fusion pipeline is illustrated in Fig. 5.

RESULTS

Unimodal Performance and Cross-Subject Robustness:

The first stage of our evaluation established the baseline performance of EEG and fNIRS in a subject-independent scenario. To ensure the results represent the system’s ability to generalize to new users, we prioritized Leave-One-Subject-Out (LOSO) cross-validation over simple hold-out splits.

EEG-based Classification:

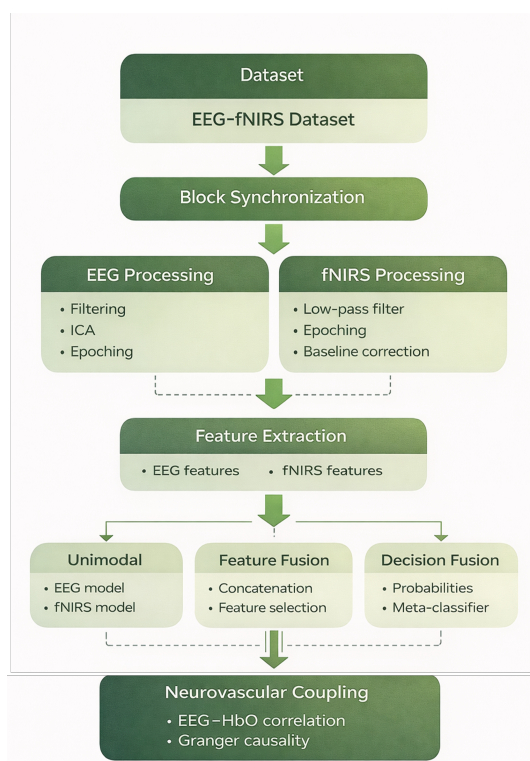


Figure 5: Classification framework

The RF classifier achieved a peak accuracy of 87.2% on the hold-out set. However, performance significantly decreased under LOSO validation to a mean accuracy of 65.95% ($\pm 20.6\%$) and an AUC of 0.709. This drop, combined with the high standard deviation, reflects the high inter-subject variability of electrophysiological signals. While EEG captures fast cognitive changes, its patterns vary significantly between individuals.

fNIRS-based Classification:

Hemodynamic features, evaluated via SVM-RBF, showed a lower peak accuracy on the hold-out set (82.9%) but demonstrated better stability across subjects. The LOSO mean accuracy remained at 76.92% ($\pm 12.8\%$) with an AUC of 0.805. These results suggest that prefrontal hemodynamic responses—specifically oxygenation trends—are more consistent markers of cognitive workload across different participants than EEG spectral densities.

Table 2: Performance Benchmark: Unimodal vs. Multimodal Fusion Strategies. Results reported as Mean (%) $\pm$ SD for Acc. and Mean $\pm$ SD for AUC.

Strategy	Hold-out Acc (%) $\pm$ SD	LOSO Acc (%) $\pm$ SD	LOSO AUC $\pm$ SD
<i>Unimodal</i>			
EEG	87.2 $\pm$ 5.4	66.0 $\pm$ 20.6	0.71 $\pm$ 0.25
fNIRS	82.9 $\pm$ 4.2	76.9 $\pm$ 12.8	0.81 $\pm$ 0.18
<i>Multimodal</i>			
Feature Fusion	90.1 $\pm$ 3.1	80.1 $\pm$ 11.7	0.86 $\pm$ 0.16
Decision Fusion	89.7 $\pm$ 2.8	85.3 $\pm$ 19.1	0.98 $\pm$ 0.05

Multimodal Feature-Level Fusion (Early Fusion):

This strategy integrates EEG and fNIRS information at

Unimodal LOSO Performance: Accuracy Distribution (N=26)

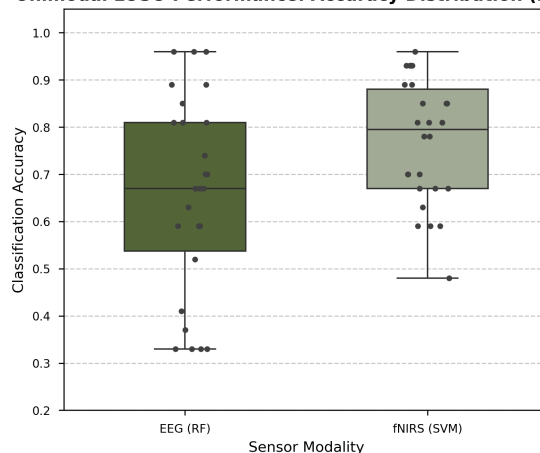


Figure 6: Comparison of unimodal classification accuracy and variability under LOSO cross-validation.

the input stage by concatenating their feature vectors into a single, high-dimensional joint representation. This approach allows the models to learn cross-modal dependencies directly from the synchronized neurophysiological data.

As shown in Table 3, the XGBoost classifier achieved a peak accuracy of 90.1% on the hold-out set. However, the subject-independent LOSO validation, which represents a more realistic scenario for new users, obtained an accuracy of 80.1%.

Table 3: Feature-Level Fusion Performance (Early Fusion). Metrics reported as Mean (%) $\pm$ SD for Acc. and Mean $\pm$ SD for AUC.

Model	Hold-out Acc (%) $\pm$ SD	LOSO Acc (%) $\pm$ SD	LOSO AUC $\pm$ SD
XGBoost	90.1 $\pm$ 3.1	80.1 $\pm$ 11.7	0.86 $\pm$ 0.16
Random Forest	84.4 $\pm$ 4.5	75.3 $\pm$ 14.1	0.81 $\pm$ 0.19
SVM-RBF	85.8 $\pm$ 3.8	74.2 $\pm$ 15.5	0.80 $\pm$ 0.21

The feature importance analysis (Fig. 7) confirms that both EEG and fNIRS contribute significantly to the classification. Frontal Alpha power and prefrontal HbO emerged as the most relevant features. This is consistent with the role of the prefrontal cortex in working memory tasks.

Fig. 7 illustrates the relative contribution (Gain) of the top features in the hybrid model. The decision-making process is clearly driven by a balanced integration of both modalities: while Frontal Alpha power (EEG) serves as a primary neural trigger, prefrontal HbO metrics (fNIRS) provide metabolic stability to the classification. This confirms the biological relevance of our features, as the most discriminative variables are located over the prefrontal cortex, the main hub for working memory processing.

Decision-Level Fusion (Late Fusion):

In this strategy, we combined the posterior probabilities from the best unimodal classifiers (EEG-RF and fNIRS-SVM). As shown in Table 4, the SVM with a Polynomial kernel (degree 3) emerged as the most robust meta-

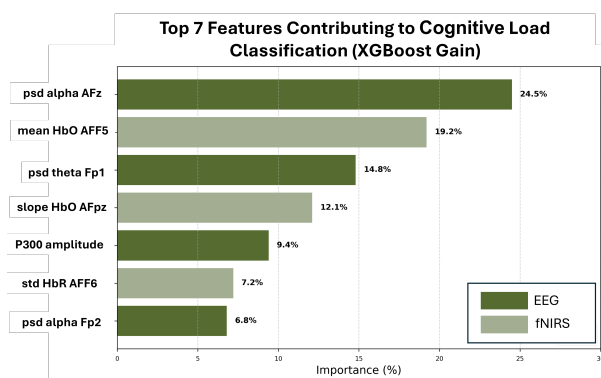


Figure 7: Feature importance ranking (XGBoost Gain) for the early fusion model.

classifier. Under LOSO validation, it achieved a mean accuracy of 85.3% ($\pm 19.1\%$) and an AUC of 0.977 (± 0.05). This high AUC value demonstrates a strong ability to separate workload states across the entire cohort as illustrated by the ROC curve in Fig. 8.

Table 4: Decision-Level Fusion Performance. Comparison of different meta-classifiers. Metrics reported as Mean (%) $\pm$ SD for Acc. and Mean $\pm$ SD for AUC.

Meta-Classifier	Hold-out Acc (%) $\pm$ SD	LOSO Acc (%) $\pm$ SD	LOSO AUC $\pm$ SD
SVM-Poly (d=3)	89.7 $\pm$ 2.8	85.3 $\pm$ 19.1	0.98 $\pm$ 0.05
XGBoost	88.5 $\pm$ 3.5	81.2 $\pm$ 14.5	0.91 $\pm$ 0.12
Random Forest	86.4 $\pm$ 4.1	79.8 $\pm$ 16.2	0.88 $\pm$ 0.15

A key finding is the balanced contribution of both modalities. As detailed in Table 5 and Fig. 9, the SVM-Poly meta-classifier assigns nearly equal weights to each signal (48.1% for EEG and 51.9% for fNIRS). This improvement happens because the model combines the strengths of both signals: the fast response of the EEG and the stability of the fNIRS. By integrating both electrical and hemodynamic data, the model can compensate for the weaknesses of each modality, leading to a more reliable prediction than unimodal methods.

Table 5: Relative contribution of EEG and fNIRS probabilities to the final decision-level fusion across different meta-classifiers.

Meta-Classifier	EEG Influence (%)	fNIRS Influence (%)
SVM-Poly (d=3)	48.1	51.9
XGBoost	51.8	48.2
Random Forest	55.2	44.8
LDA	45.5	54.5

DISCUSSION

The unimodal results highlight a clear difference between the two modalities. EEG shows a high potential to detect changes but fails to generalize well to new users (approximately 66% accuracy). This is likely due to the natural differences in brain shape and electrical patterns between individuals. In contrast, fNIRS acts as a more sta-

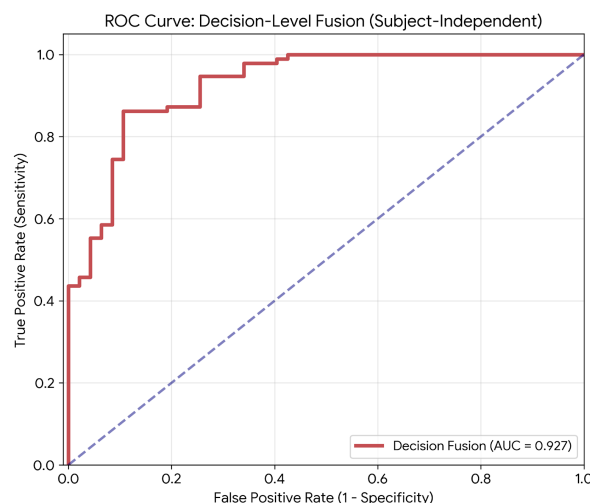


Figure 8: ROC curve analysis for the top-performing Decision-Level Fusion model.

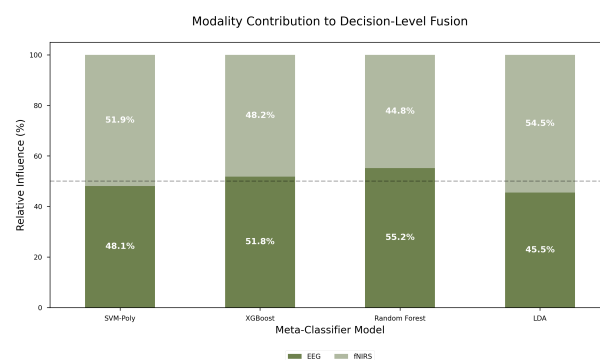


Figure 9: Relative influence of EEG and fNIRS probabilities across different meta-classifiers.

ble marker of workload across the group (77% accuracy), even though it is a slower signal.

The results of the Early Fusion strategy (90.1% peak accuracy) confirm that EEG and fNIRS work better together by capturing complementary aspects of working memory. The XGBoost model effectively identifies moments where both signals align, such as when a drop in Alpha waves (EEG) occurs at the same time as an increase in oxygenation (fNIRS). However, the 10% drop in LOSO accuracy shows that combining too many variables can make the model too specific to the training group. Regarding Late Fusion, the high AUC (0.977) should be viewed carefully. This result is partly due to our test comparing two very different states (very low effort vs. high effort). While this shows the system is very effective at detecting clear workload changes, its main strength is how the stability of fNIRS helps correct the variability of the EEG. This makes the hybrid approach more reliable for potential pBCI applications. Although the use of ICA and MARA is mainly an offline approach, it allows us to establish a clear performance benchmark for fusion. For future real-time applications, these steps could be replaced by faster online artifact rejection methods without changing the core fusion logic.

This dual-source confirmation is essential for transitioning passive BCI (pBCI) systems from controlled laboratory settings to real-world, safety-critical applications.

CONCLUSIONS

In conclusion, our results show that hybrid EEG-fNIRS fusion helps to mitigate the high inter-subject variability of EEG by using fNIRS as a stable reference. Our findings establish that:

1. **Multimodal Reliability:** Hybrid architectures consistently outperform unimodal baselines by using fNIRS as a stabilizing anchor for variable EEG signals.
2. **Generalizability:** Achieving 85% accuracy in subject-independent (LOSO) scenarios confirms that the system captures reliable biomarkers of cognitive load that work across different users.
3. **Future Work:** While the system is highly effective at detecting clear workload transitions, future research should evaluate its performance in more granular states to further define its operational limits.

BIBLIOGRAPHY

- [1] Aricò P, Borghini G, Di Flumeri G, Sciaraffa N, Colosimo A, Babiloni F. Passive BCI in operational environments: Insights, recent advances, and future trends. *IEEE Transactions on Biomedical Engineering*. 2017;64(7):1431–1436.
- [2] Fazli S et al. Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage*. 2012;59(1):519–529.
- [3] Aricò P, Borghini G, Di Flumeri G, Sciaraffa N, Babiloni F. Passive BCI beyond the lab: Current trends and future directions. *Physiological Measurement*. 2018;39(8):08TR02.
- [4] Chiarelli AM, Zappasodi F, Di Pompeo F, Merla A. Simultaneous functional near-infrared spectroscopy and electroencephalography for monitoring of human brain activity and oxygenation: A review. *Neurophotonics*. 2017;4(4):041411.
- [5] Shibasaki H. Human brain mapping: Hemodynamic response and electrophysiology. *Clinical Neurophysiology*. 2008;119(4):731–743.
- [6] Pinto-Orellana MA, Khan H, Ombao H, Mirtaheri P. Emerging neuroimaging approach of hybrid EEG-fNIRS recordings: Data collection and analysis challenges. *Data Science in Science*. 2024;3(1):2426785.
- [7] Yeung MK, Chu VW. Viewing neurovascular coupling through the lens of combined EEG–fNIRS: A systematic review of current methods. *Psychophysiology*. 2022;59(6):e14054.
- [8] Lee HT, Shim M, Liu X, et al. A review of hybrid EEG-based multimodal human–computer interfaces using deep learning: Applications, advances, and challenges. *Biomedical Engineering Letters*. 2025;15:587–618.
- [9] Kane MJ, Conway ARA, Miura TK, Colflesh GJH. Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2007;33(3):615–622.
- [10] Aghajani H, Garbey M, Omurtag A. Measuring mental workload with EEG+fNIRS. *Frontiers in Human Neuroscience*. 2017;11:359.
- [11] Shin J et al. Open access dataset for EEG and fNIRS single-trial classification. *Scientific Data*. 2018;5(1):1–16.
- [12] Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. *Neural Networks*. 2000;13(4-5):411–430.
- [13] Marriott Haresign I et al. Automatic classification of ICA components from infant EEG using MARA. *Developmental Cognitive Neuroscience*. 2021;52:101024.
- [14] Ludwig KA, Miriani RM, Langhals NB, Joseph MD, Anderson DJ, Kipke DR. Using a common average reference to improve cortical neuron recordings from microelectrode arrays. *Journal of Neurophysiology*. 2009;101(3):1679–1689.
- [15] Croce P, Zappasodi F, Merla A, Chiarelli AM. Exploiting neurovascular coupling: A Bayesian sequential Monte Carlo approach applied to simulated EEG–fNIRS data. *Journal of Neural Engineering*. 2017;14(4):046029.
- [16] Qiu L et al. Multi-modal integration of EEG-fNIRS for characterization of brain activity evoked by preferred music. *Frontiers in Neurorobotics*. 2022;16:823435.
- [17] Khan MU, Hasan MAH. Hybrid EEG-fNIRS BCI fusion using multi-resolution singular value decomposition (MSVD). *Frontiers in Human Neuroscience*. 2020;14:599802.
- [18] Rabbani MHR, Islam SMR. Deep learning networks based decision fusion model of EEG and fNIRS for classification of cognitive tasks. *Cognitive Neurodynamics*. 2024;18:1489–1506.

Scene-based Stimulation in the Context of a Visual ERP BCI

Amar Enkhbat¹, Michael Tangermann¹, Jordy Thielen¹,

¹Donders Centre for Cognition, Radboud University Nijmegen, Netherlands

E-mail: amar.enkhbat@donders.ru.nl

ABSTRACT: Visual brain-computer interfaces (BCIs) provide a promising communication channel for individuals with severe motor impairments. One well-known example is the event-related potential (ERP) matrix speller, which enables users to select characters on a computer screen. However, such systems typically require users to repeatedly shift attention between a screen-based interface and the physical environment. In this work, we propose an in-the-scene BCI paradigm, in which users interact directly with real-world objects rather than symbolic representations on a screen. We evaluated the proposed paradigm in offline experiments in which healthy participants ($N = 22$) selected target objects under two conditions: a screen-based BCI and an in-the-scene laser-based BCI. Comparisons of the elicited ERPs showed differences between the two stimulation conditions, including distinct scalp activation patterns and significantly larger N1 amplitudes in the in-the-scene condition. These findings indicate that neural responses elicited by real-world object stimulation differ from those obtained with conventional screen-based BCIs. This suggests that decoding methods may require adaptation for in-the-scene BCI systems.

INTRODUCTION

A brain-computer interface (BCI) is a type of assistive technology that decodes brain activity directly, bypassing the brain's neural pathways that control other forms of muscular activity. For people with severe speech or muscular disability such as paralysis or locked in syndrome, BCI is a promising technology that could allow them to communicate with the external world, improve independence and quality of life.

One such assistive device is the classic event-related potential (ERP) matrix speller originally described by Farwell and Donchin [1]. The 26 letters of the alphabet, together with several other symbols and commands are displayed on a computer screen that serves as a keyboard. The user focuses attention to a symbol they wish to select while rows and columns of the keyboard are highlighted. Upon highlighting letters, ERPs are elicited and captured by the electroencephalogram (EEG). By analyzing these responses across multiple stimulus presentations, the BCI determines the character attended by the user.

Although the ERP speller enables users to communicate their intentions through typing, its long-term adoption has been limited [2]. Our hypothesis is that adoption could

be limited by the sometimes cumbersome screen-based selection procedure. For example, when interacting with devices or objects in a physical environment, users must repeatedly shift attention between the screen interface and the physical environment. Additionally, during communication, users are restricted to characters or options that are pre-specified on the screen interface. The speed of interacting with a real-world object (e.g., a glass of water) often requires the user to spell out the name of the object and the type of interaction desired. Because selecting a single character typically takes several seconds, describing objects such as a glass of water or actions such as drinking from it, may take over a minute. This tedious process could lead to frustration, mental fatigue and reduced long-term usability [3].

To address these limitations, we explore the concept of an *in-the-scene BCI* originally proposed by Kolkhorst et al. [4, 5]. In their study, objects in the real-world were selected directly, circumventing the need of a screen. This was achieved by placing objects around the participant and highlighted them using a single laser pointer mounted on a moving robotic arm. The authors hypothesized, that such a system might reduce cognitive workload because users do not need to repeatedly switch attention between a screen interface and the surrounding environment. Instead, object selection could occur directly by attending to the desired item. For simplicity, we refer to the in-the-scene condition as the *laser* condition for the remainder of this paper.

However, it remains unclear whether highlighting real-world objects with varying visual and reflective properties results in different ERP responses compared to conventional screen-based stimuli that are typically presented on a homogeneous background. In this study, we use data from an experiment where participants selected objects using both, a screen condition and a laser condition. Within-subjects, we compare respective evoked ERP components regarding amplitude, latency and scalp activation characteristics.

MATERIALS AND METHODS

Participants: Twenty-five healthy participants aged between 17 and 71 years (9 male, 16 female; mean age: 38.75, SD: 16.3) participated for a single session after pre-screening and providing written informed consent. Exclusion criteria included inability to understand instructions in English, a history of color blindness, neu-

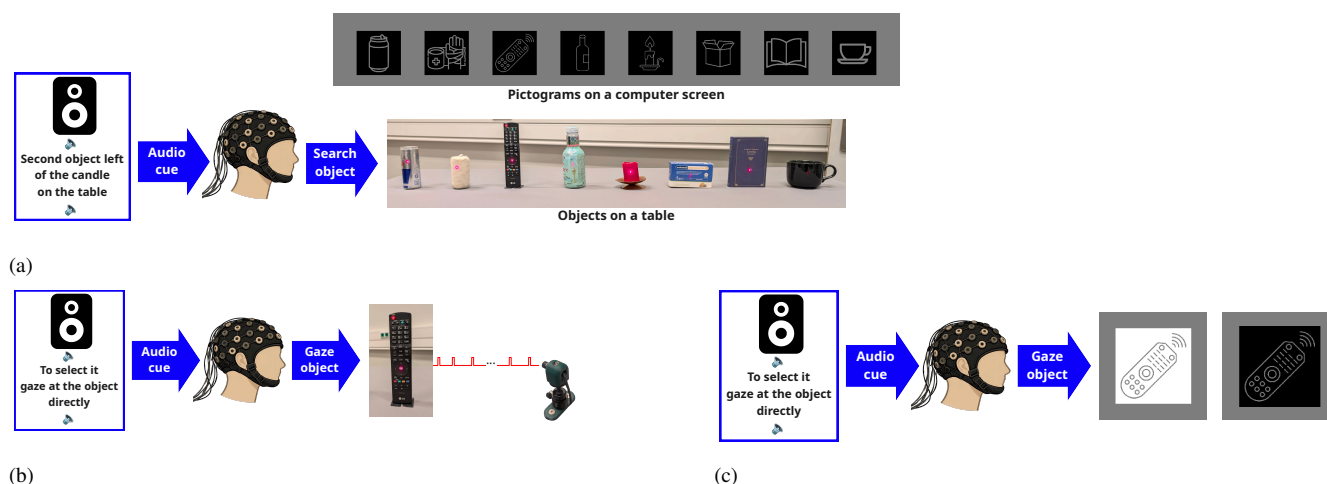


Figure 1: **Experiment overview.** (a) An audio cue instructs participants to find the target object on the table. (b) In the laser condition, participants gaze at the red laser-dot projected on the target. Highlighting is done by switching lasers on and off. (c) In the screen condition, participants gaze at the target object's pictogram on a screen. Pictograms are highlighted by alternating between white and black squares.

rological treatment, brain surgery, or epilepsy. All participants had normal or corrected-to-normal vision. Participants were compensated for their contribution to the study. The experiment was approved by the Ethical Committee of the Faculty of Social Sciences at Radboud University Nijmegen.

Materials: EEG data was recorded with 64 channel Ag/AgCl active electrodes positioned according to the 10-10 system, and amplified with a BrainProducts BrainAmp amplifier at a sampling rate of 1000 Hz. The reference electrode was placed on the nose and the ground electrode was positioned at AFz. The impedance of all electrodes was kept under 10 k Ω . Event markers were recorded and synchronized using Lab Streaming Layer (LSL). EEG acquisition and marker recording were performed using the LSL Connector for the BrainAmp amplifier.

We designed an experiment that includes the two conditions: a standard screen-based BCI and a laser-based BCI. Eight everyday objects were placed in front of the participant (can, bandage, remote, bottle, candle, box, book, cup; see Figure 1a). The objects were positioned in a line with 18 cm gap between their respective highlighting locations. Participants were seated approximately 1.5 m from the objects.

Eight laser pointers were placed to the left of the participant. Each laser was manually aligned to point at the center of mass of each object. The lasers were transistor-to-transistor logic (TTL) laser diodes (Class IIIa, 5 mW, 650 nm wavelength) with a maximum modulation frequency of 50 kHz. The lasers were controlled using a Teensy 4.0 development board capable of sub-microsecond digital output.

The measured rise and fall times of the laser diodes were $5.62 \pm 5.26 \mu\text{s}$ and $410.1 \pm 9.6 \mu\text{s}$, respectively. Overall, the measured latency between the computer-issued command and laser diode activation was sub-millisecond.

The viewing angles of the laser reflections on the objects were $0.16 \pm 0.08^\circ$ horizontally and $0.33 \pm 0.06^\circ$ vertically. The size of the reflections varied slightly depending on the optical properties of the objects. The viewing angle between the leftmost and right-most objects was 45.6° .

A computer screen was positioned 10 cm behind and above the objects, displaying eight pictograms corresponding to the objects. The pictograms were displayed on a BenQ XL2420Z monitor (24-inch, 16:9 aspect ratio, 1920 $\times$ 1080 resolution, 60 Hz refresh rate) using PsychoPy [6]. They were arranged in a 1 $\times$ 8 matrix. The viewing angle of a pictograms was 1.48° in both horizontal and vertical directions. The viewing angle from the leftmost to the right-most pictogram was 9.5° .

An optical sensor was attached to the top-left corner of the screen to capture precise stimulus onset timings for marker synchronization. Display presentation latency tests revealed an average delay of 3.1 ± 0.62 ms.

Stimuli: In this study, we used the same stimulus onset asynchrony (SOA) of 250 ms and inter-stimulus interval (ISI) of 150 ms as Kolkhorst et al. (2020) [5]. To reduce the complexity of the setup and the trial duration, we replaced the single sensor robotic arm with a parallel highlighting by multiple lasers.

As highlighting neighboring stimuli can lead to co-amplification of responses [7], two *non-adjacent* objects (laser condition) or pictograms (screen condition) were highlighted simultaneously during each highlight event, yielding a target-to-non-target ratio of 1:3. Although this target probability is slightly higher than the recommended range of 0.1–0.2 for eliciting P300 [8], it allowed us to reduce Kolkhorst et al.'s trial duration by 50%. While the amplitude of the P300 component is known to increase with longer target-to-target intervals (TTI) [8, 9] and a low target probability [10–13], excessively long TTIs and very low target probabilities may

also lead to participant fatigue and a loss of attention. To balance these factors, we used a variable TTI with a minimum interval of 500 ms. An example stimulation sequence is shown in Figure 2. The highlighting sequences were pseudo-randomly generated and satisfied three constraints: (1) two highlights of the same object or pictogram were at least 500 ms apart, (2) two highlighted objects or pictograms were not spatially adjacent, and (3) each object/pictogram was highlighted exactly 12 times. This produced highlighting sequences of a fixed length ($T = 48$ highlights, 12 s) with controlled target probability, variable and unpredictable TTIs, and reduced the co-amplification by neighboring stimuli.

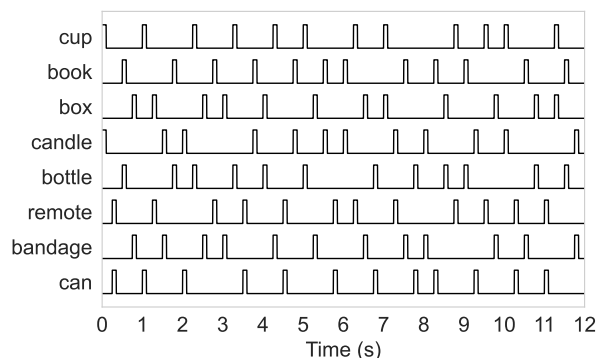


Figure 2: **Example stimulation sequence.** Each object and trial presented a unique pseudo-random sequence.

Experimental procedure: Before the experiment, the participants were familiarized with the task, object names, corresponding pictograms and the audio instructions which determined the target object or pictogram. Three practice trials were completed to make sure participants gazed at the correct object or pictograms.

Per trial, a participant's goal was to select an object placed on the table. There were 24 trials for each condition (6 for each of the objects) presented in a pseudo-random order. After 4 trials, the participants were allowed a self-paced break.

The target object's position was indicated through an audio cue (see Figure 1). It provided a logical explanation of the target object and emulated a real-world object search and selection. Additionally, the audio cue instructed to either gaze at the objects directly, or to the corresponding pictograms on the screen. The latter required participants to first find the object on the table, then focus their attention to the correct pictogram on the screen. The audio cue was repeated until participants pressed a button. It served to indicate that the target had been found, and initiated the stimulation after a random interval of 3–5 s. Participants were instructed to keep gazing at the target throughout the stimulation, which consisted of either a laser dot (see Figure 1b) or a highlighting of the pictogram (see Figure 1c).

Analysis: Data of three participants were excluded from data analysis. One participant withdrew from the experiment due to excessive eye discomfort. Data of two

participants were excluded because of technical problems related to the optical screen sensor in their session.

The remaining EEG data of 22 participants were band-pass filtered at 0.5–20 Hz using IIR forward-backward filtering. The data were then segmented to target and non-target epochs from 200 ms pre-stimulus to 600 ms post-stimulus. Epochs were baseline-corrected using the 250 ms pre-highlight interval. The epoched data were decimated to 250 Hz. Epochs with a peak-to-peak difference above $100 \mu\text{V}$ were rejected from analysis. In total, 15.9 % of epochs were excluded from further analysis.

We analyzed and compared the ERPs between the screen and laser conditions. Per participant, we estimated the amplitudes and latencies by region-of-interests (ROIs). The N1 ROI comprised of electrodes PO7, PO3, POz, PO4, PO8, PO9, O1, Oz, O2 and PO10, while the P2 ROI included FCz, Cz, CPz, Pz, POz and Oz. Amplitudes were defined as the minimum voltage within 0–250 ms for the N1 and the maximum voltage within 150–300 ms for the P2. Values were then averaged across electrodes within each ROI. To test for differences in ERP amplitudes and latencies, we used the two-sample pairwise t-test.

RESULTS

Target and non-target responses from both the screen and laser conditions were averaged across participants. The resulting event-related potentials (ERP) and corresponding scalp topographies are shown in Figure 3. Topographic maps were computed by averaging activity around the peak latency of each component. We observed clear differences in ERP waveforms and scalp activity patterns.

In the screen condition, the distinctive P1 component is observed between 112–132 ms with positive amplitudes over central electrodes and negative amplitudes over occipital electrodes. In the laser condition, a small P1 component is observed over central electrodes.

The N1 component is visible in both conditions but differences in the waveform morphology can be observed. In the screen condition, the N1 appears as a sharp negative peak, while in the laser condition it shows a broader negative deflection with a wider onset, which may suggest the presence of overlapping peaks. Additionally, the N1 amplitude appears larger in the laser condition compared to the screen condition.

The P2 component is visible in both conditions but its scalp activation patterns differ slightly. In the screen condition, P2 is primarily localized over parieto-occipital regions, with a slight extension toward central electrodes. In contrast, the laser condition shows P2 activity predominantly over central electrodes.

Temporal patterns also differ between the two conditions. In the screen condition, scalp activity initially appears in occipital regions, shifts toward central electrodes, and then returns toward occipital regions before gradually dissipating. In the laser condition, strong occipital activ-

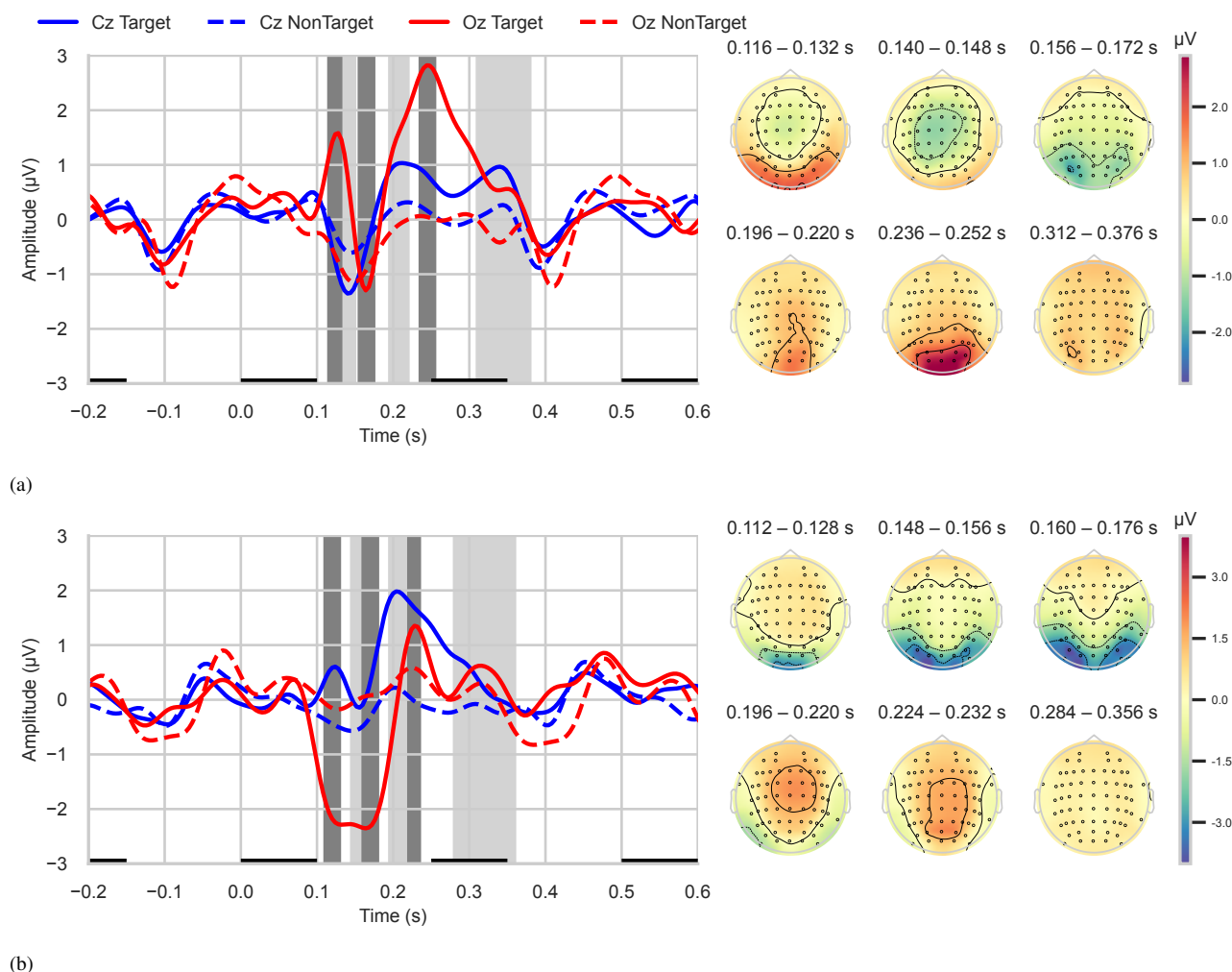


Figure 3: ERP timeseries and topoplots. Displayed are the grand average ERP timeseries (left) and topoplots (right) for the screen-based (a) and laser-based (b) conditions. Gray areas indicate the intervals across which the topoplots are shown. The black horizontal bars indicate highlighting intervals.

ity persists initially and subsequently shifts toward central electrodes before gradually dissipating.

For both conditions, no clearly distinguishable P3 component was observed despite the stimuli being designed to elicit a P300 component.

Component amplitude and latency analysis: Overall, ERP amplitudes observed in this study were lower than those typically reported for visual oddball paradigms. Figure 4 shows the N1 and P2 amplitudes and latencies across participants. Paired-samples t tests were conducted to compare ERP amplitudes and latencies between the laser and screen conditions. Effect sizes are reported as Cohen's d_z for paired designs. As shown in Table 1, the N1 amplitude was significantly more negative in the laser condition than in the screen condition ($M_{\text{diff}} = -1.47 \mu\text{V}$, 95% CI $[-2.35, -0.58]$, $t(21) = -3.44$, $p = .002$, $d_z = -0.73$). No significant condition differences were observed for N1 latency, P2 amplitude or P2 latency ($p > .05$). As mentioned in the methods section, we estimated the minimum and maximum values for ROIs instead of using only the Cz and Oz electrodes.

Our results suggest that the P2 responses are comparable between the screen and laser conditions despite differences in their scalp activation patterns.

DISCUSSION

This study compared ERP responses elicited by a conventional screen-based BCI and a scene-based BCI in which objects were highlighted using laser pointers.

The results indicate that the observed ERP waveforms differed between the two conditions. Although both conditions timing-wise used an identical stimulation protocol, the screen condition exhibited clearly distinguishable P1, N1, and P2 components, whereas the laser condition exhibited a broader N1 component followed by a reduced P2 component. One possible explanation for the missing P1 in the laser condition is that the broader N1 response may overlap with the P1 component, making it less distinguishable. In addition, differences were observed in the spatial distribution of the P2 component. In the screen condition, the P2 response was primarily localized over

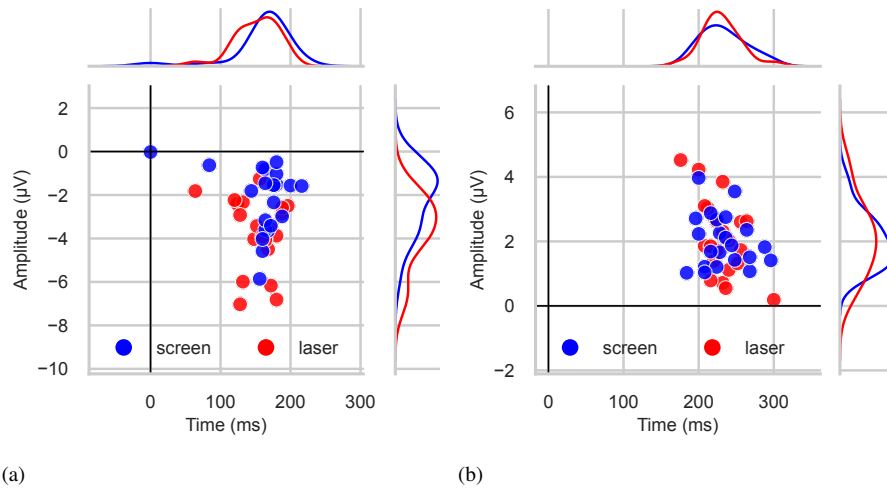


Figure 4: **Amplitude and latency distributions of ERP components.** Displayed are the ERP amplitude and latencies under laser and screen conditions. (a) N1 component amplitude and latency distribution. (b) P2 component amplitude and latency distribution.

Table 1: **Paired-samples t -tests comparing ERP measures between screen and laser conditions.** Note, M_{screen} and M_{laser} denote means for the screen and laser conditions, respectively. M_{diff} represents the mean paired difference (laser – screen). Confidence intervals correspond to the mean difference. Effect sizes are reported as Cohen’s d_z for paired designs.

Outcome	M_{screen} (SD)	M_{laser} (SD)	M_{diff}	95% CI	$t(21)$	p	d_z
N1 amplitude (μV)	-2.20 (1.51)	-3.66 (1.61)	-1.47	[-2.35, -0.58]	-3.44	.002	-0.73
N1 latency (ms)	160.55 (43.32)	151.82 (30.72)	-8.73	[-31.73, 14.28]	-0.79	.439	-0.17
P2 amplitude (μV)	2.02 (0.82)	2.13 (1.17)	0.11	[-0.33, 0.55]	0.53	.604	0.11
P2 latency (ms)	233.09 (30.06)	230.18 (25.74)	-2.91	[-19.38, 13.57]	-0.37	.717	-0.08

occipital regions, whereas in the laser condition it was most prominent over central electrodes. This shift in activation patterns may reflect differences in visual processing between controlled screen stimuli and stimuli presented in a more complex real-world environment.

We also observed a significant difference in N1 amplitude, which was larger in the laser condition than in the screen condition. This is consistent with findings reported by Kolkhorst et al. (2019) [4]. Assuming that the N1 component reflects early visual processing, this effect may be caused by differences in the visual characteristics of the stimuli. Despite being smaller, the laser dot may have appeared more intense or visually salient than the screen stimulus.

Although the stimulation protocol was based on an odd-ball paradigm, a clearly distinguishable P300 component was not consistently observed. One possible explanation is the relatively small visual angle of the stimuli, which may have reduced the signal-to-noise ratio of the P300 component. The waveforms provided in Figure 3 seem similar to those reported by the ISCEV standard for clinical visual evoked potentials [14, 15]. Specifically, at electrode Oz, the waveform of the laser condition resembles the pattern onset/offset VEP and the waveform of the screen condition resembles a flash VEP. However, as our protocols are substantially different from those used in the ISCEV standards, this similarity may not indicate a functional relationship and requires further investigation. Our data suggest that ERP responses in scene-based BCIs may differ in morphology and spatial distribution

from those observed in conventional screen-based systems. Consequently, decoding approaches for screen-based BCIs may need to account for these differences when applied to scene-based paradigms. This will be explored in future work.

Although the present study demonstrates that laser-based stimulation can elicit ERP responses with comparable intensities, it still needs to be tested, if in-the-scene BCIs can improve the usability or the workload in practical applications as this depends on factors beyond electrophysiological responses, including system complexity, setup time, the need for caregivers, user comfort, and the level of cognitive or physical workload (e.g., eye strain). Future work should investigate these aspects through dedicated usability studies.

In addition, the stimulation paradigm itself may influence usability and performance. The present study focuses on an ERP paradigm, but other BCI paradigms such as steady-state visually evoked potential (SSVEP) [16] or code-modulated visual evoked potential (c-VEP) [17] approaches may offer higher information transfer rates [18]. However, VEP-based paradigms are sensitive to stimulus properties such as size, color, luminance, and flicker frequency. In complex real-world environments, these properties may vary substantially, and their effects on VEP responses should be carefully investigated before adopting such paradigms in scene-based BCIs.

Finally, the hardware setup for scene-based BCIs may also play an important role. In the present work, objects were highlighted using laser pointers for its simplicity.

For practical applications, more advanced highlighting methods may be adopted.

CONCLUSION

This work presents an in-the-scene BCI using a laser-based implementation for highlighting real-world objects in parallel. EEG responses elicited by this approach were compared with those obtained using a conventional screen-based BCI. Statistical analysis of evoked ERP components revealed that the two conditions produced a comparable P2 amplitude and latency. While the N1 component did not show a significant difference in latency, a significantly larger N1 amplitude was found in the laser condition. These results suggest that laser-based highlighting elicits neural responses that are slightly different to those used in traditional ERP BCIs. Future work should investigate adapting decoding methods developed for screen-based paradigms for in-the-scene BCI systems.

ACKNOWLEDGEMENT

This project has received funding from the European Union's research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101118964.

REFERENCES

- [1] Farwell L, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*. 1988;70(6):510–523.
- [2] Wolpaw JR et al. Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. *Neurology*. 2018;91(3):e258–e267.
- [3] Pinegger A, Hiebel H, Wriessnegger SC, Müller-Putz GR. Composing only by thought: Novel application of the P300 brain-computer interface. *PLOS ONE*. 2017;12(9):e0181584.
- [4] Kolkhorst H, Veit J, Burgard W, Tangermann M. Heterogeneity of event-related potentials in a screen-free brain-computer interface. In: *Proceedings of the 8th Graz Brain-Computer Interface Conference 2019*. Graz, Austria, 2019, 228–233.
- [5] Kolkhorst H, Veit J, Burgard W, Tangermann M. A robust screen-free brain-computer interface for robotic object selection. *Frontiers in Robotics and AI*. 2020;7:38.
- [6] Peirce J et al. Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*. 2019;51(1):195–203.
- [7] Townsend G et al. A novel P300-based brain-computer interface stimulus presentation paradigm: Moving beyond rows and columns. *Clinical Neurophysiology*. 2010;121(7):1109–1120.
- [8] Duncan CC et al. Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*. 2009.
- [9] Gonsalvez, Polich. P300 amplitude is determined by target-to-target interval. *Psychophysiology*. 2002.
- [10] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*. 1988;70(6):510–523.
- [11] Donchin E, Coles MGH. Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*. 1988;11(3):357–374.
- [12] Picton TW. The P300 Wave of the Human Event-Related Potential. *Journal of Clinical Neurophysiology*. 1992;9(4):456.
- [13] Polich J. P300 Clinical Utility and Control of Variability. *Journal of Clinical Neurophysiology*. 1998;15(1):14.
- [14] Odom JV et al. ISCEV standard for clinical visual evoked potentials: (2016 update). *Documenta Ophthalmologica*. 2016;133(1):1–9.
- [15] Habjan M Šuštar et al. ISCEV standard for clinical visual evoked potentials (2025 update). *Documenta Ophthalmologica*. 2025;151(2):97–112.
- [16] Vialatte FB, Maurice M, Dauwels J, Cichocki A. Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. *Progress in Neurobiology*. 2010;90(4):418–438.
- [17] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): A literature review. *Journal of Neural Engineering*. 2021.
- [18] Neeling M de, Van Hulle MM. Single-paradigm and hybrid brain computing interfaces and their use by disabled patients. *Journal of Neural Engineering*. 2019;16(6):061001.

Connectivity Between Scalp EEG and Intracranial Recordings During 40 Hz Visual Stimulation

Mani Mirsaedi¹, Tjaša Mlinarič², Evelien Carrette³, Dirk Van Roost⁴, Paul Boon³, Alfred Meurs³,
Marc M. Van Hulle¹

¹Laboratory for Neuro- and Psychophysiology, KU Leuven, Leuven, Belgium

²Laboratory for Cognitive Neurology, KU Leuven, Leuven, Belgium

³Department of Neurology, Ghent University Hospital, Ghent, Belgium

⁴Department of Neurosurgery, Ghent University Hospital, Ghent, Belgium

E-mail: Mani.Mirsaedi@KULeuven.be

ABSTRACT: We investigated whether rhythmic visual stimulation at 40 Hz reveals connectivity-based relationships between scalp EEG and intracranial recordings. Two epilepsy patients with simultaneous EEG, ECoG, and SEEG recordings participated in regular, oddball, and irregular stimulation conditions. Connectivity was quantified using coherence and weighted phase lag index (wPLI). Rhythmic stimulation produced robust coupling between scalp and invasive electrodes, whereas irregular stimulation did not. These findings suggest that scalp EEG can reflect neural activity recorded in cortical and deeper brain structures during rhythmic stimulation.

INTRODUCTION

Scalp electroencephalography (EEG) provides a non-invasive measure of large-scale brain dynamics and is widely used to study rhythmic neural activity in humans. However, the extent to which EEG reflects activity originating from deeper brain structures remains debated [1]. Neural electric fields propagate through conductive tissue before reaching the scalp and are attenuated by distance as well as by the spatial organization of the underlying neuronal sources. The greater physical distance between deep neural generators and scalp electrodes further attenuates their measurable contributions to scalp recordings [2] and in particular, subcortical structures often exhibit distinct neuronal architectures that can generate weaker extracranial signals compared to cortical sources [3–5]

Despite these limitations, identifying signatures of deep neural activity in scalp EEG is of considerable interest for both clinical and research applications. One important example arises in the development of brain-computer interfaces (BCIs) based on implanted electrocorticography (ECoG) electrodes. For such systems, it would be highly valuable to determine before surgery whether a given individual is likely to generate neural signals that can support reliable BCI control [6]. Ideally, such a pre-implantation assessment would rely on non-invasive

recordings. Measurements obtained from scalp EEG could provide a rapid and practical approach for evaluating neural signal quality prior to implantation [6].

The ability to detect activity originating from deeper brain structures is also relevant for neurological disorders in which subcortical regions play a central role. For example, structures within the medial temporal lobe, including the hippocampus and amygdala, are some of the first regions to be affected by AD [7]. Because early stages of AD represent a potential window for intervention, identifying non-invasive biomarkers reflecting activity in these regions is of particular interest [8]. However, whether scalp EEG can reliably predict the presence of informative neural activity recorded intracranially remains an open question [9].

One approach to addressing the detectability of deep activity in scalp EEG is to induce a strong and well-defined oscillatory response and evaluate whether this activity can be observed simultaneously across recording modalities. In particular, visual stimulation around 40 Hz is known to produce robust gamma-band steady-state responses in the human brain [10]. If cortical and subcortical structures exhibit reliable 40 Hz activity during such stimulation, it becomes possible to evaluate whether corresponding signatures can be detected in scalp EEG.

Interpreting coupling between recording modes requires careful consideration of the mechanisms underlying correlated signals. Similar activity observed across electrodes may arise from two fundamentally different processes. In one case, passive spread of electrical potentials through conductive tissue leads to the same source being detected at multiple locations [11]. Under the quasi-static approximation of Maxwell's equations, such field spread results in signals that appear nearly simultaneous across sensors, producing relationships centered around zero phase difference [12]. In the other case, neural transmission between distinct neuronal populations generates a temporally structured activation sequence characterized by a consistent phase offset between signals [13].

In the present study, we investigated whether rhythmic visual stimulation at 40 Hz can reveal relationships be-

tween scalp EEG and intracranial recordings obtained simultaneously from cortical surface electrodes and depth electrodes. By combining coherence and phase-based connectivity measures, we examined whether activity observed at the scalp reflects signals recorded in cortical and subcortical regions, including hippocampal and other deep structures. Through this approach, we aim to assess the feasibility of detecting intracranial rhythmic activity from non-invasive scalp recordings and to characterize the nature of the coupling observed across recording modalities.

MATERIALS AND METHODS

Participants: For the current study, two epilepsy patients were recruited at Ghent University Hospital (UZ Gent) following approval from the Commission Medical Ethics of UZ Gent. All procedures were conducted in accordance with applicable ethical regulations for research involving human participants. Both patients were acutely implanted with subdural grids (ECoG) and/or depth electrodes (SEEG) as part of their clinical evaluation, while simultaneous scalp EEG recordings were obtained using surface electrodes. Prior to participation, they were informed about the study procedures, including data processing and storage, and provided written informed consent.

Experimental procedure and data acquisition: The experimental protocol comprised three stimulation conditions [14, 15]: Regular (40 Hz), Oddball (40 Hz with task), and Irregular. Each comprised a 1-minute baseline and two 2.5-minute stimulation blocks. Regular and Task conditions utilized 40 Hz square-waves (50% duty cycle), whereas the Irregular control featured single periods from random frequencies between 30–50 Hz. Participants maintained central fixation 55 cm from the LED panel. During the Task, participants mentally counted random 250 ms red LED flashes; behavioral data were not analyzed. LED luminance was 225–250 lumen. Simultaneous EEG, ECoG, and SEEG were acquired via an SD LTM 64 Express system (Micromed, Italy) at 256 Hz, referenced to average mastoids.

Electrode localization: For each participant, cortical reconstruction and volumetric segmentation were performed on the preimplantation MRI using the FreeSurfer image analysis suite (version 6.0) [16]. The post implantation CT scan was subsequently co-registered to the FreeSurfer-derived images using the Brainstorm toolbox with the SPM12 extension [17] [18]. Electrode coordinates were manually identified from CT artifacts and mapped onto the cortical surface. Cortical visualizations were generated using the Brainstorm toolbox.

Preprocessing: The preprocessing pipeline was implemented using the FieldTrip toolbox for MATLAB [19] and included re-referencing to minimize reference-dependent bias in subsequent connectivity analyses, band-pass filtering, and attenuation of 50 Hz line noise and its harmonics in the EEG, ECoG, and SEEG recordings.

For each participant, 5-minute stimulation epochs and a 1-minute baseline were extracted and aligned to stimulation onset. These were segmented into non-overlapping 5-second trials to enable artifact rejection while preserving temporal structure. Trials with variance exceeding the 80th percentile were excluded to minimize the influence of movement artifacts and non-physiological transients. The remaining artifact-reduced trials were retained for subsequent spectral and connectivity analyses.

Signal to noise ratio analysis: Given the heterogeneous electrode placement across participants, 40 Hz activation was evaluated at the level of individual electrodes. Power spectral density (PSD) was estimated for each 5-second trial and each channel using Welch's method with 1-second Hamming windows and 50% overlap [20]. To quantify entrainment, 40 Hz signal-to-noise ratio (SNR) was computed per electrode as $10\log_{10}(P_{\text{signal}}/P_{\text{noise}})$, using 40 Hz for signal and the mean of 38–39 Hz and 41–42 Hz for noise. Stimulation SNR distributions were compared against rest using one-sided nonparametric rank-based tests. Multiple comparisons were controlled via the Benjamini–Hochberg procedure ($\alpha = 0.05$), with effect sizes quantified as the area under the receiver operating characteristic curve (AUC).

Coherence: Coherence is a frequency-domain measure of synchronization between two signals and quantifies the consistency of their relative amplitude and phase relationship at a given frequency [21, 22]. When the phase difference between two signals remains stable across time, coherence approaches 1; when the relationship varies from moment to moment, coherence approaches 0. Thus, coherence reflects the strength of the linear relationship between two signals at a specific frequency [23].

Consider two signals $x(t)$ and $y(t)$. The magnitude-squared coherence (MSC) is defined as

$$C_{xy}^2(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)}, \quad (1)$$

where $S_{xy}(f)$ denotes the cross-spectrum between the signals and $S_{xx}(f)$ and $S_{yy}(f)$ denote their respective autospectra.

Weighted phase lag index (wPLI): While coherence provides an indication of consistent phase relationships between two signals, it does not distinguish the underlying nature of this relationship. The critical difference between volume conduction and functional connectivity lies in the phase lag between signals. Volume conduction induces near-zero phase differences across sensors, whereas genuine functional interactions are characterized by consistent non-zero phase lags [12, 13].

Let the cross-spectrum between signals x and y at frequency f be defined as

$$S_{xy}(f) = E\{X(f)Y^*(f)\}, \quad (2)$$

where $X(f)$ and $Y(f)$ are the Fourier transforms of the signals and $*$ denotes complex conjugation. The complex coherence is then

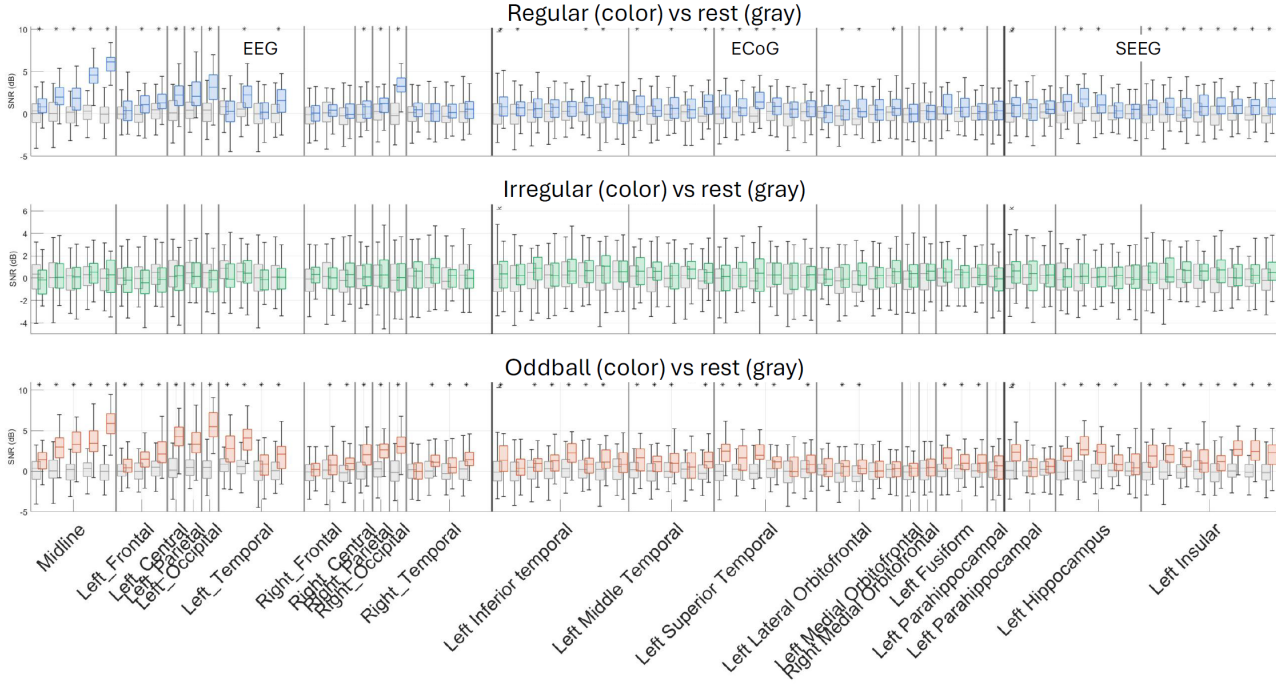


Figure 1: Electrode-wise 40 Hz signal-to-noise ratio (SNR) during regular stimulation relative to rest (gray) for Subject_01. Rhythmic stimulation produced strong electrophysiological responses detectable at the scalp level, with the strongest activity observed over central, parietal, and occipital electrodes. dots indicate significant differences between the stimulation and resting states.

$$\gamma_{xy}(f) = \frac{S_{xy}(f)}{\sqrt{S_{xx}(f)S_{yy}(f)}}. \quad (3)$$

Because the cross-spectrum is complex-valued, it can be written as

$$S_{xy}(f) = |S_{xy}(f)|e^{i\phi(f)}, \quad (4)$$

where $\phi(f)$ represents the phase difference between the signals. Volume conduction primarily contributes to the real part of $S_{xy}(f)$, corresponding to $\phi(f) \approx 0$.

To overcome this limitation, the weighted phase lag index (wPLI) relies on the imaginary component of the cross-spectrum. The wPLI is defined as [12]

$$\text{wPLI}(f) = \frac{E \{ |\text{Im}(S_{xy}(f))| \text{sign}(\text{Im}(S_{xy}(f))) \}}{E \{ |\text{Im}(S_{xy}(f))| \}}. \quad (5)$$

By weighting the imaginary part of the cross-spectrum, wPLI emphasizes consistent non-zero phase differences while down-weighting contributions close to zero phase lag [12]. Because purely zero-lag interactions yield $\text{Im}(S_{xy}(f)) = 0$, wPLI collapses toward zero in the presence of volume conduction. Consequently, wPLI provides a more conservative estimate of functional connectivity by selectively capturing phase-lagged synchronization.

In the present study, wPLI was used to further characterize the nature of the interactions identified by coherence analysis. After identifying electrode pairs exhibit-

ing significant coherence, the corresponding wPLI values were evaluated to determine whether the observed coupling reflected phase-lagged interactions or near-zero-lag synchronization.

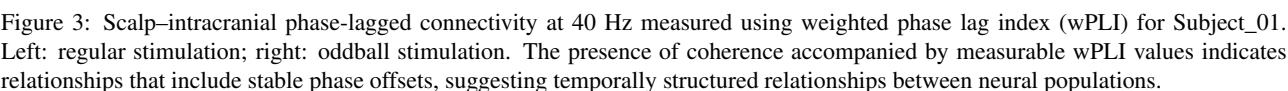
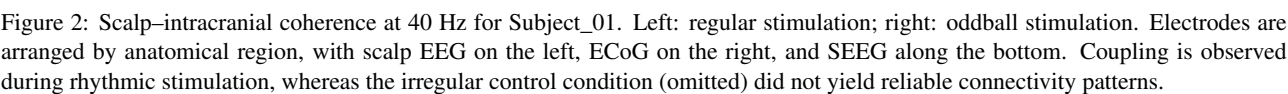
RESULTS

40 Hz signal-to-noise ratio: Regular 40 Hz stimulation produced clear increases in signal-to-noise ratio (SNR) relative to rest in both participants (Fig. 1). After FDR correction, 44 of 78 electrodes in Subject_01 and 43 of 58 electrodes in Subject_02 showed significantly higher SNR during the regular condition. Significant responses were present in both scalp and invasive recordings. Median SNR increased from 0.014 dB at rest to 0.753 dB in Subject_01 and from 0.190 dB to 1.762 dB in Subject_02. In contrast, irregular stimulation did not produce reliable entrainment. No electrodes survived FDR correction and median SNR values remained close to resting levels.

The oddball condition produced even stronger responses. After FDR correction, 59 of 78 electrodes in Subject_01 and 43 of 58 electrodes in Subject_02 showed significantly higher SNR than rest. In both participants, the median SNR increase was larger in scalp electrodes than in invasive recordings.

Overall, the results show a consistent condition-dependent pattern: regular and oddball stimulation produced robust 40 Hz entrainment across scalp and invasive electrodes, whereas irregular stimulation did not.

40 Hz coherence: Coherence analysis revealed syn-



Weighted phase lag index: Weighted phase lag index (wPLI) analysis revealed phase-lagged interactions dur-

The present analysis examined how rhythmic visual stimulation at 40 Hz is represented across scalp EEG and in-

tracranial recordings.

In both subjects, rhythmic stimulation produced strong electrophysiological responses detectable at the scalp level, with the strongest activity observed over posterior electrodes. This spatial distribution is consistent with the visual origin of the stimulation paradigm and aligns with previous work showing that visual flicker around 40 Hz produces strong gamma-band steady-state responses measurable in scalp EEG, particularly over occipital regions [10, 14]. These results indicate that rhythmic visual stimulation reliably induces large-scale oscillatory responses that can be detected using non-invasive recordings.

Invasive recordings confirmed that the 40 Hz response was not restricted to scalp measurements. Both subjects exhibited rhythmic activity across multiple intracranial electrodes. Although the precise anatomical distribution differed between subjects due to differences in electrode implantation, rhythmic responses were observed across several cortical and deeper brain regions sampled by the implanted electrodes. Previous intracranial studies have similarly demonstrated that rhythmic visual stimulation can entrain gamma-band activity across cortical regions and deeper structures including the hippocampus [15]. Together, these observations indicate that the scalp responses observed during rhythmic stimulation co-occur with activity recorded locally by intracranial electrodes.

Two factors motivated the focus on 40 Hz activity. First, the signal serves as a functional tracer, ensuring that the observed coupling is stimulus-driven rather than a random occurrence. Second, the continuous nature of this rhythmic activity provides the stationarity required for connectivity measures like coherence and wPLI to accurately estimate phase relationships—a task not possible with transient or random/resting state signals.

The irregular condition (30–50 Hz jitter) served as a stochastic control because its non-periodic nature prevents frequency-specific entrainment. It should be noted that irregular flickering differs from code-modulated (cVEP) paradigms, which use repeated m-sequences and would compromise their use as a stochastic control.

Connectivity analyses examined the relationship between scalp EEG and intracranial recordings during the stimulation conditions. Across both participants, coupling between scalp EEG and intracranial recordings was observed during the rhythmic stimulation conditions (Fig. 2), whereas the irregular control condition did not produce high-SNR scalp regions and therefore did not yield reliable connectivity patterns. This indicates that the interactions described below were specifically associated with the presence of rhythmic stimulation.

In Subject 01, interactions linked temporal scalp regions with temporal cortical recordings, including inferior temporal, middle temporal, and superior temporal cortex. Additional interactions connected frontal scalp regions with the lateral orbitofrontal cortex. In these region pairs, coherence occurred together with very low wPLI values, indicating relationships centered around zero phase dif-

ference between scalp and cortical signals.

Interactions between scalp EEG and SEEG contacts were also present in Subject 01, involving depth electrodes located within the insula, hippocampus, and parahippocampal region. For these region pairs, coherence was again accompanied by very low wPLI values, indicating relationships centered around zero phase difference between scalp activity and signals recorded in these deeper structures.

In Subject 02, interactions linked midline scalp regions with frontal cortical areas, including middle frontal, superior frontal, and precentral cortex. In these region pairs, coherence occurred together with very low wPLI values, indicating relationships centered around zero phase difference between scalp and cortical activity.

Interactions between scalp EEG and SEEG contacts were also observed in Subject 02, involving depth electrodes located within regions associated with focal cortical dysplasia (FCD). In these region pairs, coherence was accompanied by very low wPLI values, indicating relationships centered around zero phase difference between scalp signals and SEEG recordings.

Interactions characterized by coherence together with measurable wPLI values were also observed in the dataset, indicating relationships that include stable phase offsets between recording sites. Because wPLI emphasizes consistent non-zero phase differences, these interactions are less compatible with purely instantaneous field spread and instead suggest temporally structured relationships between neural populations.

Across both participants, the presence of significant coherence accompanied by very low wPLI values suggests a substantial contribution of volume conduction. Because volume conduction reflects the spread of electrical fields through brain tissue, skull, and scalp, this finding implies that scalp recordings may retain information about activity originating from deeper brain structures. Similar conclusions have been reported in previous studies using different methodological approaches [1, 9, 24]. Crucially, robust 40 Hz signals did not universally yield high connectivity. Multiple sites with high SNR displayed no significant coherence or wPLI with the scalp. This dissociation confirms that the observed coupling is not a trivial consequence of signal strength or volume conduction, but reflects specific functional interactions within an entrained neural circuit.

ACKNOWLEDGMENTS

This project has received funding from the European Union's research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101118964.

The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government.

Additional funding was provided by research grants re-

ceived from the Horizon 2020 research and innovation programme under grant agreement No. 857375, the special research fund of the KU Leuven (C24/18/098), the Belgian Fund for Scientific Research – Flanders (G0A4321N, G0C1522N, G031426N), and the Hercules Foundation (AKUL 043).

REFERENCES

- [1] M. Fahimi Hnazaee et al., “Localization of deep brain activity with scalp and subdural EEG,” *NeuroImage*, vol. 223, p. 117 344, Dec. 2020.
- [2] P. Krishnaswamy et al., “Sparsity enables estimation of both subcortical and cortical activity from MEG and EEG,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, Nov. 2017.
- [3] L. M. Andersen, K. Jerbi, and S. S. Dalal, “Can EEG and MEG detect signals from the human cerebellum?” *NeuroImage*, vol. 215, p. 116 817, Jul. 2020.
- [4] Y. Attal and D. Schwartz, “Assessment of Subcortical Source Localization Using Deep Brain Activity Imaging Model with Minimum Norm Operators: A MEG Study,” *PLoS ONE*, vol. 8, no. 3, G. R. Barnes, Ed., e59856, Mar. 2013.
- [5] F. Lopes da Silva, “EEG and MEG: Relevance to Neuroscience,” *Neuron*, vol. 80, no. 5, pp. 1112–1128, Dec. 2013.
- [6] M. Fahimi Hnazaee et al., “Towards predicting ECoG-BCI performance: Assessing the potential of scalp-EEG*,” *Journal of Neural Engineering*, vol. 19, no. 4, p. 046 045, Aug. 2022.
- [7] V. Planche et al., “Structural progression of Alzheimer’s disease over decades: The MRI staging scheme,” *Brain Communications*, vol. 4, no. 3, fcac109, May 2022.
- [8] N. D. Anderson, “State of the science on mild cognitive impairment (MCI),” *CNS Spectrums*, vol. 24, no. 1, pp. 78–87, Feb. 2019.
- [9] M. Seeber, L.-M. Cantonas, M. Hoevels, T. Sesia, V. Visser-Vandewalle, and C. M. Michel, “Subcortical electrophysiological activity is detectable with high-density EEG source imaging,” *Nature Communications*, vol. 10, no. 1, p. 753, Feb. 2019.
- [10] C. S. Herrmann, “Human EEG responses to 1–100 hz flicker: Resonance phenomena in visual cortex and their potential correlation to cognitive phenomena,” *Experimental Brain Research*, vol. 137, no. 3–4, pp. 346–353, Apr. 2001.
- [11] M. X. Cohen, “Effects of time lag and frequency matching on phase-based connectivity,” *Journal of Neuroscience Methods*, vol. 250, pp. 137–146, Jul. 2015.
- [12] M. Vinck, R. Oostenveld, M. Van Wingerden, F. Battaglia, and C. M. Pennartz, “An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias,” *NeuroImage*, vol. 55, no. 4, pp. 1548–1565, Apr. 2011.
- [13] M. Muthuraman, R. Govindan, G. Deuschl, U. Heute, and J. Raethjen, “Differentiating phase shift and delay in narrow band coherent signals,” *Clinical Neurophysiology*, vol. 119, no. 5, pp. 1062–1070, May 2008.
- [14] E. Khachatryan et al., “Cognitive tasks propagate the neural entrainment in response to a visual 40 Hz stimulation in humans,” *Frontiers in Aging Neuroscience*, vol. 14, p. 1 010 765, Oct. 2022.
- [15] T. Mlinarič et al., “Visual gamma stimulation induces 40 Hz neural oscillations in the human hippocampus and alters phase synchrony and lag,” *Communications Biology*, vol. 8, no. 1, p. 1301, Aug. 2025.
- [16] B. Fischl, “FreeSurfer,” *NeuroImage*, vol. 62, no. 2, pp. 774–781, Aug. 2012.
- [17] F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy, “Brainstorm: A User-Friendly Application for MEG/EEG Analysis,” *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–13, 2011.
- [18] *Statistical Parametric Mapping*. Elsevier, 2007.
- [19] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, “FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data,” *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–9, 2011.
- [20] P. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, Jun. 1967.
- [21] R. W. Thatcher, “Coherence, Phase Differences, Phase Shift, and Phase Lock in EEG/ERP Analyses,” *Developmental Neuropsychology*, vol. 37, no. 6, pp. 476–496, Aug. 2012.
- [22] R. Srinivasan, W. R. Winter, J. Ding, and P. L. Nunez, “EEG and MEG coherence: Measures of functional connectivity at distinct spatial scales of neocortical dynamics,” *Journal of Neuroscience Methods*, vol. 166, no. 1, pp. 41–52, Oct. 2007.
- [23] S. Malekpour, J. A. Gubner, and W. A. Sethares, “Measures of generalized magnitude-squared coherence: Differences and similarities,” *Journal of the Franklin Institute*, vol. 355, no. 5, pp. 2932–2950, Mar. 2018.
- [24] F. Pizzo et al., “Deep brain activities can be detected with magnetoencephalography,” *Nature Communications*, vol. 10, no. 1, p. 971, Feb. 2019.

Comparison of Lightweight EEG Pipelines for Silent Speech Interfaces

E. Nikolopoulos^{1,2}, G. Kontaxakis^{3,4}, E. J. Gómez^{1,2,5}, I. Oropesa^{1,2,5}

¹ Biomedical Engineering and Telemedicine Centre, ETSI Telecomunicación, Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain

² Instituto de Investigación Hospital 12 de Octubre (imas12), Hospital Universitario 12 De Octubre, Madrid, Spain

³ ETSI Telecomunicación, Universidad Politécnica de Madrid (UPM), Madrid, España

⁴ Information Processing and Telecommunications Center, ETSI Telecomunicación, UPM, Madrid, Spain

⁵ Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Madrid, Spain

E-mail: manosnikol@gmail.com

ABSTRACT: This study investigates various machine learning (ML) and deep learning (DL) methodologies for decoding inner speech using non-invasive electroencephalography (EEG). Utilizing a publicly available high-density 128-channel dataset from ten participants, nine distinct computational scenarios were tested, ranging from traditional spatial filtering and statistical ensembles to Riemannian geometry and convolutional neural networks. The research aims to compare and identify the most effective signal processing strategies to overcome the inherently low signal-to-noise ratio of EEG signals generated during inner speech. Results demonstrate that Riemannian manifold-based approaches and compact deep learning architectures significantly outperform traditional Euclidean-based methods, achieving average accuracies across the ten participants of 32.41% and 30.48%, with peak accuracies reaching 44.50% and 34.50% respectively, compared with the 25% chance level of a four-class inner speech task.

INTRODUCTION

A brain-computer interface (BCI) serves as a computational bridge between neural activity and external hardware. These systems function by acquiring brain signals, decoding the underlying intent, and translating that data into digital commands to operate devices such as computers or prosthetic limbs [1,2]. Beyond assistive control, BCIs can also be used to monitor cognitive states such as fatigue and mental workload and support a wide range of applications. Silent Speech Interfaces (SSIs) represent a class of communication systems that enable speech processing in the absence of an intelligible acoustic signal [3]. Unlike traditional speech recognition that relies on microphone input, an SSI captures physical or physiological data

produced during the speech production process, such as brain signals, movement of the articulators or surface muscle activation. By decoding these internal precursors to sound, the system can generate text or synthesized speech, effectively bypassing the requirement for vocalization [4].

The development of SSIs relies on a diverse array of sensing modalities designed to capture articulatory intent without the need for acoustic output. These technologies are broadly categorized into methods that measure physiological activity directly and those that track physical articulatory movements [5]. A critical frontier in this domain is the intersection of SSIs with BCIs, which bypass peripheral articulators entirely to decode the cognitive intent to speak. While invasive neural approaches like electrocorticography capture high-resolution signals with superior signal-to-noise ratio (SNR) directly from the speech-motor cortex [6–9], non-invasive electroencephalography (EEG) remains the preferred practical choice for monitoring scalp-level brain activity due to its safety, affordability, and high temporal resolution [10–13]. EEG-based BCIs capture the cortical electrical activity generated during the cognitive formulation of words, which is a process widely referred to in the literature as 'covert' or 'imagined' speech [14]. By measuring voltage fluctuations at the scalp, researchers can record the high-temporal-resolution neural dynamics associated with language production, completely bypassing the need for physical articulation. This non-invasive approach offers a safe, portable, and highly accessible alternative to surgical neural implants.

The state-of-the-art (SoA) processing pipeline for these EEG-based covert speech decoders heavily relies on robust artifact removal, such as Independent Component Analysis (ICA) [15], followed by advanced feature extraction [14]. Recently, classification strategies have experienced a paradigm shift; contemporary research

increasingly leverages deep learning (DL) architectures, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which autonomously extract complex, non-linear spatiotemporal patterns from raw or minimally preprocessed multichannel EEG data to significantly improve decoding accuracy [14].

Despite these algorithmic advancements, the practical deployment of EEG-based BCI systems for silent speech remains constrained by fundamental physiological and hardware challenges. The inherently low SNR of EEG signals and high inter-subject variability restrict current systems primarily to offline, synchronous environments that utilize highly limited vocabularies, such as isolated vowels, phonemes, or simple directional commands [13]. Furthermore, multi-class classification accuracy significantly deteriorates as the vocabulary size expands, and many experimental protocols still rely on the unnatural mental repetition of prompts to achieve viable decoding rates [14]. Crucially, the implementation of advanced deep learning architectures demands immense computational processing power, which is exceedingly difficult to achieve in real-world, standalone neuroprosthetic devices. This stringent hardware requirement severely restricts the portability of these systems and limits their practical, everyday applications outside of controlled laboratory settings.

The goal of this study is to investigate different methodologies used in machine learning (ML) and DL in the field of EEG-driven SSIs, with a focus on minimizing computational requirements for real-world deployment. Using an open access dataset, we compare the classification performance of several algorithms across different methodological scenarios. More specifically, we perform intrasubject classification, and compare the average accuracy across participants and the maximum accuracy achieved among nine different scenarios, while maintaining low computational requirements to ensure feasibility for real-world applications.

MATERIALS AND METHODS

Dataset description: In this work, we analyzed the publicly available dataset by Nieto et al. [16]. The study provides a comprehensive, open-access EEG resource specifically designed to advance the study of inner speech and BCIs. The dataset contains over nine hours of continuous EEG recordings. It significantly contributes in the neuroengineering research by offering a publicly available, high-density EEG database of inner, pronounced and visualized speech commands that can be used to better understand speech-related brain mechanisms and train robust classification models.

The dataset features ten healthy, right-handed, native Spanish-speaking adults (mean age 34; four females, six males) with no history of neurological or speech disorders and no prior BCI experience. The data acquisition was performed in an electrically shielded room utilizing a high-resolution, 136-channel BioSemi Active Two system. The hardware setup prioritized

dense cortical coverage with 128 active EEG channels, supplemented by 8 external channels for electrooculography (EOG) and electromyography (EMG) to rigorously monitor and control for eye and articulatory movement artifacts. The EEG signals were initially recorded at a high sampling rate of 1024 Hz with 24-bit resolution. To ensure optimal signal quality, researchers fitted each subject with appropriately sized head caps, applied conductive gel, and strictly maintained electrode impedances below 40 ohms prior to all recording sessions.

The experimental protocol, as described on Fig. 1, yielded a total of 5,640 EEG trials across all participants and across the three distinct paradigms: inner speech, pronounced speech, and a visual spatial condition. During the tasks, participants responded to visual cues corresponding to four directional commands ("arriba," "abajo," "derecha," "izquierda").

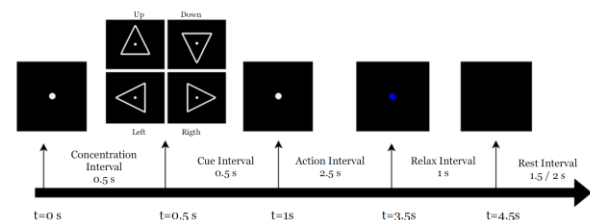


Figure 1: The experimental procedure described in [16]. Each trial spanned 4.5 seconds, starting with a 0.5-second concentration interval where participants focused on a white circle. This was followed by a 0.5-second cue interval displaying a directional triangle (up, down, left, or right). At $t = 1.0$ s, the cue disappeared, initiating a 2.5-second action interval where the subject performed the designated mental or vocal task. At $t = 3.5$ s, the circle turned blue to signal a 1.0-second relax interval, concluding the trial before a variable rest period began. Modified from Nieto et al. [16] under the Creative Commons CC-BY 4.0 license.

In this work, we focused exclusively on the data obtained during the inner speech condition. This subset comprised a total of 2,236 trials, yielding an average of approximately 224 total inner speech trials per participant (averaging 56 trials per class). Inner speech can be defined as an internalized cognitive process in which individuals think in pure meanings, typically accompanied by the auditory imagery of their own inner “voice”. Unlike overt speech, inner speech occurs without physical articulation or acoustic output. Specifically, during these trials, participants were instructed to mentally repeat the cued word using their own voice as a direct command to the computer until the on-screen circle turned blue. To ensure purely covert generation, they were not provided with any rhythmic cues, and they were explicitly directed to remain completely still, avoiding any tongue or mouth movements or focus on articulatory gestures.

Preprocessing of the data: A major advantage of this dataset is that it provides both the continuous raw data and precisely preprocessed EEG epochs. The data processing pipeline, as provided by [16], implemented primarily using the MNE Python library, began with re-referencing the signals using a virtual channel created by averaging the two earlobe electrodes (EXG1 and EXG2)

to eliminate common-mode voltage. The signals were then band-pass filtered between 0.5 and 100 Hz using a zero-phase finite impulse response filter, along with a 50 Hz notch filter to remove power line noise interference. Next, the continuous data was decimated to a 256 Hz sampling rate and epoched into 4.5-second segments corresponding to individual trials. Finally, ICA using the infomax algorithm was applied strictly to the 128 EEG channels to isolate and remove artifacts related to blinks, gaze, and mouth movements, resulting in ready-to-use data tensors tailored for BCI classification and time-frequency analysis [17].

Word Classification: To provide a comprehensive analysis of inner speech classification, the following methodologies represent a diverse range of EEG signal processing strategies. These scenarios span from traditional spatial filtering to advanced adaptive denoising, Riemannian geometry, and deep learning.

Scenario 1: Multi-Band Common Spatial Patterns (CSP). This approach utilizes CSP to find spatial filters that maximize the variance for one class while simultaneously minimizing it for another [18]. The methodology tests three distinct frequency bands, motor-related (8–30 Hz), broadband (1–45 Hz), and gamma (30–80 Hz), to determine which oscillatory rhythms provide the most discriminative information for inner speech. The spatially filtered features are then fed into a Linear Discriminant Analysis (LDA) classifier to perform the final classification of the inner speech targets [19].

Scenario 2: Region of Interest (ROI) Spatial Filtering. This methodology employs a physiologically driven channel reduction strategy, focusing exclusively on 39 specific channels that correspond to the left hemisphere's speech-dominant regions and the midline, such as Broca's area, Wernicke's area, the primary motor cortex, and the supplementary motor area. Consequently, 89 of the 128 total channels are explicitly removed to eliminate irrelevant spatial data. This removal includes the entire right hemisphere (all 32 'B' channels), the majority of the 'A' block (channels A3–A32), as well as specific peripheral 'C' (C3–C11, C28–C32) and 'D' channels (D1–D5, D25–D32). By filtering out this irrelevant spatial data before applying CSP, the pipeline concentrates the classifier's attention on the cortical areas most likely to generate language-related signals during inner speech. After applying CSP to these targeted spatial regions, the extracted features are classified using a Support Vector Machine (SVM) with an RBF kernel to decode the specific inner speech commands [20].

Scenario 3: Statistical Feature Ensemble. This strategy treats EEG signals as a source for high-dimensional statistical modeling by extracting ten specific features per channel, including time-domain measures (mean, skewness, kurtosis, root mean square) and frequency-domain band powers. It uses mutual information to select the top 500 features and employs a soft-voting ensemble of five logistic regression models to improve classification stability and mitigate the risk of overfitting.

Scenario 4: Multi-Wavelet and Autoregressive Modeling. This methodology captures the complexity of neural signals through a combination of temporal and frequency-scale analysis. It extracts autoregressive (AR) coefficients to model temporal dynamics and uses wavelet packet shannon entropy and wavelet variance to quantify the complexity and variability of the signal across different frequency scales. A third-degree polynomial SVM is then used to handle the non-linear boundaries of these features.

Scenario 5: Optimized ROI and Spectral Feature Engineering. This methodology refines the ROI approach by combining dynamic channel subsetting with targeted spectral extraction during the active "thought" window (1.5–3.5s). It extracts statistical features alongside power in the alpha (8–13 Hz) and beta (13–30 Hz) bands from the left hemisphere. A mutual information selector then retains the top 100 features, which are then fed to an SVM.

Scenario 6: Variational Mode Decomposition (VMD). This advanced signal processing methodology uses VMD to adaptively decompose EEG signals into Intrinsic Mode Functions (IMFs). By reconstructing the signal using only relevant middle modes that capture alpha and beta activity, it acts as a non-linear filter to remove low-frequency drift and high-frequency noise. The cleaned signal is then processed for statistical and spectral features before classification via a logistic regression ensemble.

Scenario 7: Riemannian Geometry and Filter-Banks. This methodology approaches EEG classification by treating signal covariance matrices as points on a curved Riemannian manifold. It utilizes a filter-bank approach across three frequency bands, applying principal component analysis (PCA) for dimensionality reduction and Xdawn covariances to enhance the signal-to-noise ratio. The matrices are then mapped into a tangent space, allowing for robust linear classification [21].

Scenario 8: Linear Predictive Coding (LPC) Formant Analysis. Borrowing from speech processing, this methodology treats EEG channels as if they were audio sources capturing the "inner voice". It applies a pre-emphasis filter to flatten the spectrum and then calculates LPC coefficients to model the spectral envelope of the signal. These coefficients, representing "formants" of inner speech, are classified using a gradient-boosted XGBoost model [22].

Scenario 9: Deep Learning with EEGNet. This methodology employs EEGNet [23], a compact CNN designed specifically for EEG data. It uses depthwise and separable convolutions to learn both temporal and spatial filters directly from raw, broadband EEG data (1–40 Hz). This allows the model to automatically discover discriminative patterns across the array without manual feature engineering [24].

To evaluate the robustness of the proposed models, the classification performance was assessed using a 5-fold stratified cross-validation strategy following an intra-subject approach. This approach ensures that each training and testing fold maintains an equal

representation of the four inner speech classes, providing a reliable estimate of the models' generalization capabilities across different data subsets while preventing class imbalance bias [25]. While the stratified cross-validation utilized shuffled trials to maximize training set diversity, the temporal integrity of the data was preserved through prior ICA artifact removal and the use of strictly unsupervised spatial filtering within each fold to prevent data leakage from future signals. Given the finite sample size, the theoretical chance level of 25% is insufficient; therefore, the adjusted statistical significance threshold for non-random classification was calculated to be 29.8% based on the binomial distribution ($\alpha = 0.05$) for an average of 224 trials per subject, following the framework described by [26].

RESULTS

Tab. 1 presents a comparative performance analysis of the abovementioned nine distinct classification scenarios for inner speech, highlighting the average testing and maximum accuracy achieved by each methodology across a 10-subject dataset. The results indicate that Scenario 7 (Riemannian Geometry) is the most effective methodology for this dataset, achieving the highest average accuracy of 32.41% and a peak max accuracy of 44.5%. By treating EEG covariance matrices as points on a curved manifold and utilizing Xdawn covariances, this approach successfully managed the low signal-to-noise ratio inherent in inner speech. It significantly outperformed traditional spatial filtering methods like Scenario 1 (CSP), which struggled to exceed the adjusted 29.8% random chance threshold for this 4-class problem.

Table 1: Performance metrics for inner speech classification across nine experimental scenarios, detailing the grand average accuracy (%), and peak subject performance (%).

Object	Average Accuracy	Max Accuracy
Scenario 1	24.43 (%)	33.83 (%)
Scenario 2	23.3 (%)	28.9 (%)
Scenario 3	26.2 (%)	31.1 (%)
Scenario 4	24.9 (%)	27.5 (%)
Scenario 5	25.3 (%)	28.9 (%)
Scenario 6	24.4 (%)	28.5 (%)
Scenario 7	32.41 (%)	44.5 (%)
Scenario 8	24.5 (%)	30.4 (%)
Scenario 9	30.48 (%)	34.5 (%)

Scenario 9 (EEGNet) emerged as the second-strongest performer with an average accuracy of 30.48%. The success of this DL approach suggests that the native spatial-temporal filters learned by the CNN are better at capturing the complex, non-linear patterns of broadband EEG data than most manual feature engineering methods. While Scenario 3 (Statistical Ensemble) and Scenario 5 (Optimized ROI/Stats) showed slight improvements over the baseline by reaching roughly 25-26% average accuracy, they remained notably behind the Riemannian and Deep Learning pipelines in terms of

consistency and peak performance.

Interestingly, specialized signal processing techniques such as Scenario 6 and Scenario 8 did not yield the expected performance breakthroughs, hovering around 24.4% to 24.5% average accuracy.

DISCUSSION

The experimental results highlight a significant performance gap between manifold-based or automated feature extraction methods and traditional manual feature engineering. Scenario 7, which utilized Riemannian geometry and Xdawn covariances, proved the most robust, suggesting that treating EEG signal covariance as points on a curved manifold better captures the subtle neural dynamics of inner speech than standard spatial filters like CSP. CSP-based methods (Scenarios 1) struggled to maintain performance above the adjusted random chance threshold, likely due to the high spatial attenuation and non-stationarity of the EEG signal across the 128-channel array.

The insufficient accuracies of the specialized signal processing techniques of Scenarios 6 and 8 suggest that while VMD is effective at cleaning signals and LPC is theoretically sound for speech-related brain activity, the high dimensionality of the 128-channel data may require the more robust manifold-based or automated feature extraction found in Scenarios 7 and 9. Ultimately, the transition from Euclidean-based statistics to Riemannian geometry provided the most substantial leap in classification reliability for this inner speech task.

Furthermore, the results of Scenario 9 (EEGNet) indicate that DL models can more effectively discover latent spatial-temporal patterns in broadband EEG data without the need for manual physiological priors. While specialized techniques such as VMD (Scenario 6) and LPC (Scenario 8) are theoretically grounded in speech and signal processing, they did not yield substantial performance gains in this context. This suggests that the high dimensionality of 128-channel EEG requires either the robust geometric frameworks of Riemannian manifold mapping or the hierarchical feature learning of neural networks to reliably differentiate between complex linguistic cognitive intents.

Our best models, Riemannian Geometry (32.41%) and EEGNet (30.48%), successfully replicate the standard ~30% baseline for the highly noisy Nieto et al. dataset [16,27,28], but fall short of the 45–81% accuracies achieved by the current SoA [29,30]. This performance gap highlights a strict computational trade-off. SoA DL methods, such as 3D-CNNs and spatial-temporal cross-attention networks, achieve high accuracy but require massive parameter counts, billions of FLOPs, and intensive multi-GPU training environments [29,31]. Meanwhile, top-performing ML approaches, like ensemble logistic regression or multi-wavelet SVMs, avoid deep learning's training costs but shift the heavy computational burden onto exhaustive, resource-intensive feature extraction and selection processes [30,32]. Ultimately, bridging the gap between our ~32%

peak accuracy and the SoA will require sacrificing computational efficiency to implement either complex spatial neural modeling or heavy feature engineering. Despite the promise of Riemannian mapping and convolutional architectures, key limitations regarding model capacity and generalization remain. The small cohort of ten participants yields a restricted number of trials, which inherently increases the risk of high-capacity models like EEGNet overfitting to this specific public dataset. As demonstrated by recent cross-session BCI competitions [33], complex neural network architectures frequently suffer drastic performance regression on unseen data compared to traditional methods. Therefore, before committing to high-capacity hypotheses for real-world neuroprosthetic deployment, these architectures must be rigorously evaluated on independent, out-of-distribution data to verify true generalization. Additionally, the 32.41% peak average accuracy underscores the need for broader algorithmic exploration. Future research must investigate advanced architectures to achieve the robust accuracy required for functional BCI communication. Most notably, the emerging field of Geometric Deep Learning [34–36] offers a highly promising trajectory by explicitly combining the manifold robustness of Riemannian geometry with the automated feature extraction capabilities of neural networks. Crucially, future explorations must balance high accuracy with the computational efficiency required for real-world deployment. High-density 128-channel setups and complex mathematical operations, such as Riemannian covariance calculations, impose severe burdens on portable neuroprosthetic hardware. A primary limitation of current research is sustaining classification accuracy while drastically reducing this computational overhead. To realize accessible and lightweight EEG-based silent speech interfaces, future work must prioritize intelligent channel reduction, algorithmic optimization, and edge-computing frameworks.

CONCLUSION

This research concludes that Riemannian geometry provides the most substantial leap in classification reliability for EEG-driven silent speech interfaces, achieving a peak accuracy of 44.5%. Deep learning via EEGNet also offers a highly competitive and automated alternative for decoding imagined speech. The findings suggest that future development in non-invasive SSIs should move away from traditional Euclidean-based statistics in favor of geometric manifold learning and spatial-temporal neural architectures to handle the complex computational hurdles of neural language production.

REFERENCES

[1] Cruz MV, Jamal S, Sethuraman SC. A Comprehensive Survey of Brain-Computer Interface Technology in Health care: Research

- Perspectives. *J Med Signals Sens* 2025;15.
- [2] Zhang H, Jiao L, Yang S, Li H, Jiang X, Feng J, et al. Brain-computer interfaces: the innovative key to unlocking neurological conditions. *Int J Surg* 2024;110:5745–62.
- [3] Ferreira D, Silva S, Curado F, Teixeira A. Exploring Silent Speech Interfaces Based on Frequency-Modulated Continuous-Wave Radar. *Sensors* 2022;22.
- [4] Anumanchipalli GK, Chartier J, Chang EF. Speech synthesis from neural decoding of spoken sentences. *Nature* 2019;568:493–8.
- [5] Denby B, Schultz T, Honda K, Hueber T, Gilbert JM, Brumberg JS. Silent speech interfaces. *Speech Commun* 2010;52:270–87.
- [6] Understanding and Decoding Imagined Speech using Electrographic Recordings in Humans Stéphanie MARTIN. n.d.
- [7] SpringerBriefs in Speech technology. n.d.
- [8] Chakrabarti S, Sandberg HM, Brumberg JS, Krusienski DJ. Progress in speech decoding from the electrocorticogram. *Biomed Eng Lett* 2015;5:10–21.
- [9] Luo S, Rabbani Q, Crone NE. Brain-Computer Interface: Applications to Speech Decoding and Synthesis to Augment Communication. *Neurotherapeutics* 2022;19:263–73.
- [10] Vorontsova D, Menshikov I, Zubov A, Orlov K, Rikunov P, Zvereva E, et al. Silent eeg-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification. *Sensors* 2021;21.
- [11] H Chowdhury ME, Sarman R, Bushnaq D, Chabbouh M, Aljindi R, Pedersen S, et al. IJABC: International Journal of Activity and Behavior Computing Summary of the Silent Speech Decoding Challenge Using EEG Data for Arabic Inner Speech Recognition Summary of the Silent Speech Decoding Challenge Using EEG Data for Arabic Inner Speech Recognition IJABC: International Journal of Activity and Behavior Computing. n.d.
- [12] Lopez-Bernal D, Balderas D, Ponce P, Molina A. A State-of-the-Art Review of EEG-Based Imagined Speech Decoding. *Front Hum Neurosci* 2022;16.
- [13] Rahman N, Khan DM, Masroor K, Arshad M, Rafiq A, Fahim SM. Advances in brain-computer interface for decoding speech imagery from EEG signals: a systematic review. *Cogn Neurodyn* 2024;18:3565–83.
- [14] Panachakel JT, Ramakrishnan AG. Decoding Covert Speech From EEG-A Comprehensive Review. *Front Neurosci* 2021;15.
- [15] Chuang CH, Ko LW, Lin YP, Jung TP, Lin CT. Independent component ensemble of EEG for brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2014;22:230–8.
- [16] Nieto N, Peterson V, Rufiner HL, Kamienkowski

- JE, Spies R. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Sci Data* 2022;9.
- [17] Bell AJ, Sejnowski TJ. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput* 1995;7:1129–59.
- [18] Moufassih M, Tarahi O, Hamou S, Agounad S, Idrissi Azami H. Boosting motor imagery brain-computer interface classification using multiband and hybrid feature extraction. *Multimed Tools Appl* 2024;83:49441–72.
- [19] Almufareh MF, Kausar S, Humayun M, Tehsin S, Farooq A. Inner Speech Decoding: A Comprehensive Review. *Wiley Interdiscip Rev Cogn Sci* 2025;16.
- [20] Gonzalez-Lopez JA, Gomez-Alanis A, Martín Doñas JM, Pérez-Córdoba JL, Gomez AM. Silent speech interfaces for speech restoration: A review. *IEEE Access* 2020;8:177995–8021.
- [21] Nguyen CH, Karavas GK, Artemiadis P. Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features. *J Neural Eng* 2018;15.
- [22] Shah R, Riaz A, Quershi S, Zia A, Iqbal U. Enhancing Silent Speech Decoding Using EEG and Gradient Boosting Classifiers: A Comparative Analysis of XGBoost and LightGBM on a Native Arabic Silent Speech Dataset. 2025 International Conference on Activity and Behavior Computing, ABC 2025, Institute of Electrical and Electronics Engineers Inc.; 2025.
- [23] Milyani AH, Attar ET. Deep learning for inner speech recognition: a pilot comparative study of EEGNet and a spectro-temporal Transformer on bimodal EEG-fMRI data. *Front Hum Neurosci* 2025;19.
- [24] Kobayashi N, Nemoto T, Morooka T. High Accuracy Silent Speech BCI Using Compact Deep Learning Model for Edge Computing. International Winter Conference on Brain-Computer Interface, BCI, vol. 2023- February, Institute of Electrical and Electronics Engineers Inc.; 2023.
- [25] Alzahrani S, Banjar H, Mirza R. Systematic Review of EEG-Based Imagined Speech Classification Methods. *Sensors* 2024;24.
- [26] Müller-Putz GR, Scherer R, Pfurtscheller G, Brunner C, Leeb R. Better than Random? A closer look on BCI results. vol. 10. 2008.
- [27] Radwan YA, Ahmed Mohamed E, Metwalli D, Barakat M, Ahmed A, Kiroles AE, et al. Stochasticity as a solution for overfitting—A new model and comparative study on non-invasive EEG prospects. *Front Hum Neurosci* 2025;19.
- [28] Lopez-Bernal D, Balderas D, Ponce P, Molina A. Exploring inter-trial coherence for inner speech classification in EEG-based brain-computer interface. *J Neural Eng* 2024;21.
- [29] Wang Y, Cai H, Zhou L, Yang Z, Wang L, Zhao J, et al. Decoding inner speech via frequency-specific cortical EEG representations. *Inf Process Manag* 2026;63:104676.
- [30] Tasin SM, Chowdhury MEH, Pedersen S, Chabbouh M, Bushnaq D, Aljindi R, et al. Ensemble Machine Learning Model for Inner Speech Recognition: A Subject-Specific Investigation. *IEEE Access* 2026.
- [31] Dong L, Shao H, Zhang L, Li L. STCCA: Spatial–Temporal Coupled Cross-Attention Through Hierarchical Network for EEG-Based Speech Recognition. *Sensors* 2025;25.
- [32] Abdulghani MM, Walters WL, Abed KH. Enhancing the Classification Accuracy of EEG-Informed Inner Speech Decoder Using Multi-Wavelet Feature and Support Vector Machine. *IEEE Access* 2024;12:147929–41.
- [33] Roy RN, Hinss MF, Darnet L, Ladouce S, Jahanpour ES, Somon B, et al. Retrospective on the First Passive Brain-Computer Interface Competition on Cross-Session Workload Estimation. *Frontiers in Neuroergonomics* 2022;3.
- [34] Aristimunha B, Ju C, Collas A, Bouchard F, Mian A, Thirion B, et al. SPD Learn: A Geometric Deep Learning Python Library for Neural Decoding Through Trivialization 2026.
- [35] Paillard J, Hipp JF, Engemann DA. GREEN: A lightweight architecture using learnable wavelets and Riemannian geometry for biomarker exploration with EEG signals. *Patterns* 2025;6.
- [36] Kobler RJ, Hirayama J, Zhao Q, Kawanabe M. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG 2022.

EEG-Based Identification of Individuals With Spinal Cord Injury Using Motor-Related and Visual Evoked Potentials

E. Nikolopoulos¹, K. Kostoglou², C. Fidas^{3,4}, G. Müller-Putz^{2,5}

¹ Biomedical Engineering and Telemedicine Centre, ETSI Telecomunicación, Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain

² Institute of Neural Engineering, Graz University of Technology, Graz, Austria

³ Interactive Technologies Laboratory, Department of Electrical and Computer Engineering, University of Patras, Patras, Greece

⁴ Neuroengineering & Brain-Computer Interfaces Research Infrastructure, University of Patras, 26504 Patras, Greece

⁵ BioTechMed, Graz, Austria

E-mail: manosnikol@gmail.com

ABSTRACT: A growing area of brain computer interface (BCI) research focuses on user identification, where individual neural signatures support secure authentication and personalization. Visual evoked potentials (VEPs) are widely used for their reliability, yet the integration of multiple neural modalities for identification remains underexplored. This study investigated, for the first time, the combined use of movement-related cortical potentials (MRCPs) and VEP features for electroencephalography (EEG)-based identification of individuals with spinal cord injury (SCI). We analyzed a publicly available EEG dataset from ten participants with cervical SCI who attempted five hand movements: pronation, supination, palmar grasp, lateral grasp, and hand open. VEPs were elicited by visual cues instructing the movement attempts, while MRCPs reflected the motor preparation and execution processes. Classification results showed that the “hand open” movement achieved the highest identification accuracy (40.28%, $4.03 \times$ above the 10% chance level). These findings demonstrate that combining MRCP and cue-evoked VEP features offers a promising multimodal biometric approach for identifying individuals with SCI.

INTRODUCTION

A brain-computer interface (BCI) is a technology that facilitates direct communication between the brain and external devices by capturing, interpreting, and converting brain signals into commands for controlling prosthetics, computers, or other assistive technologies [1,2]. Among the brain signals commonly employed in BCIs, event-related potentials (ERPs) are voltage changes generated in brain structures in response to specific sensory, cognitive, or motor events [3]. They are typically measured with electroencephalography (EEG) and time-locked to discrete events to isolate event-related neural activity from background brain signals [4]. Visual evoked potentials (VEPs) are a specific subtype of ERPs that represent the electrical activity generated in the

visual cortex in response to visual stimulation, providing a non-invasive method to assess the functional integrity of the visual pathway from the retina to the occipital cortex [5]. Beyond clinical and neurophysiological applications, VEPs have been explored as biometric signatures for user identification and authentication. Unlike traditional biometrics based on iris patterns, facial features, or fingerprints [6], VEPs offer neural-level distinctiveness, high reliability, and robustness to spoofing due to their stimulus-locked nature. Identification accuracies exceeding 99% have been reported under controlled visual stimulation [7], and recent code-modulated VEP paradigms have achieved authentication accuracies of 92% with F1 scores of 0.89 [8–11].

While VEPs capture individual differences in visual processing, other EEG components can similarly reflect user-specific brain dynamics in non-visual domains. Movement-related cortical potentials (MRCPs) have been extensively used in BCIs to detect and classify movement intentions. MRCPs are slow negative EEG shifts that begin 1-2 seconds before voluntary movement, reflecting sequential activation of motor-preparation regions (Supplementary Motor Area) and execution areas (Primary Motor Cortex) [12]. Crucially, MRCPs are generated not only during actual motor execution but also during movement attempts in paralyzed individuals and during motor imagery, where no overt movement occurs [13–15]. The amplitude and temporal characteristics of MRCPs systematically change based on different movement types, varying with movement complexity, force, and effector involved, making them sensitive and adaptable markers of motor intention and planning [13,15,16]. This makes MRCPs particularly valuable for BCIs aimed at detecting different movement intentions, enabling intuitive control of prosthetics, computers, and rehabilitation devices [17,18]. Moreover, their sensitivity to individual motor control patterns makes MRCPs a promising complementary feature for user characterization beyond visual responses, extending their

potential use from motor decoding to user identification and authentication.

The goal of this study is to explore, for the first time, the combined use of VEPs and MRCPs for user identification, with a particular focus on individuals with spinal cord injury (SCI). Specifically, we aim to determine whether integrating visually evoked and motor-induced neural responses enhances discriminative power and to identify which movement type contributes most effectively to user differentiation.

Unlike most previous EEG-based biometric studies conducted exclusively on healthy participants, this work targets a neurologically distinct population. SCI individuals often exhibit altered sensorimotor cortical activity patterns due to reorganization following injury, offering a unique opportunity to evaluate whether neuro-based authentication methods remain robust across different neural profiles. Investigating such a population not only broadens the potential applicability of EEG-based identification systems but also aligns with the broader goal of developing neurotechnologies that are inclusive and clinically relevant.

The integration of motor-induced and visually evoked neural responses may therefore represent a promising step toward next-generation neuro-based biometric systems, bridging clinical and security-oriented applications.

MATERIALS AND METHODS

Dataset description: In this work, we analyzed the publicly available dataset by Ofner et al. [19]. The study investigated whether attempted arm and hand movements by individuals with cervical SCI could be decoded from low-frequency, time-domain EEG signals, specifically focusing on MRCPs, with the aim to advance BCIs that can control neuroprosthetic devices. Ten individuals with cervical SCI participated in the study, with neurological injury levels ranging from C1 to C7 and American Spinal Injury Association Impairment Scale (AIS) scores from A to D. The participants had an average 49.22 ± 18.44 years and they were right-handed. All participants had restricted hand function, with some retaining minimal movement capability while others had no active hand movement. During the experiments, participants sat in front of a computer screen with their arm resting on a pillow or table to ensure comfort and stability throughout the recording sessions. The experiment employed a paradigm where each trial began with a fixation cross and an audible beep (Fig. 1a). Participants were instructed to focus their gaze on the cross, which was displayed for 5 seconds throughout the trial, to minimize eye movements and were asked to avoid swallowing or blinking during this period. At 2 seconds into each trial, a class cue appeared and remained for 3 seconds, representing one of five movement classes: pronation, supination, palmar grasp, lateral grasp, or hand open, as shown on Fig. 1b.

Depending on their motor ability, participants either attempted or executed the cued movement immediately

upon cue presentation. After each trial, participants were requested to return to the original resting position, followed by a random break lasting 1-3 seconds before the next trial commenced. EEG data was sampled at 256 Hz using 61 electrodes positioned according to the 10-5 system, with four additional electrodes for electrooculogram (EOG) monitoring.

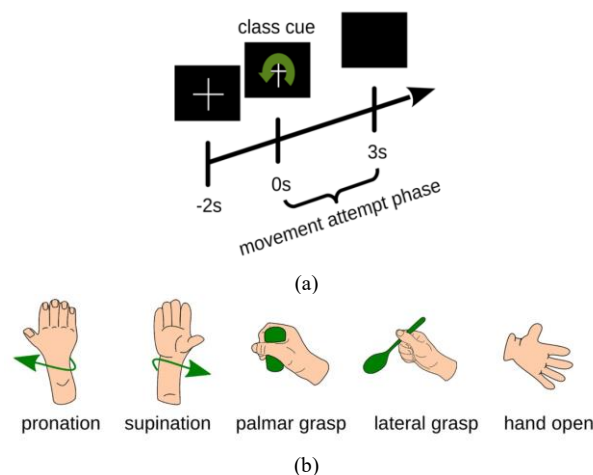


Figure 1: The experimental procedure described in [19]. (a) A fixation cross was presented, and after 2 seconds, a cue appeared and remained on the screen for 3 seconds, during which the participant attempted the corresponding movement. (b) The different attempted movements. Modified with permission from Ofner et al. [19] under the Creative Commons CC-BY license.

Preprocessing of the data: The EEG preprocessing pipeline comprised two sequential stages to optimize signal quality for offline analyses. First, raw data were imported at 256 Hz, filtered with a fourth-order Butterworth bandpass (0.5 – 40 Hz), and zero-phase filtered to avoid phase distortion. In the second stage, the data were loaded into EEGLAB [20]. We applied Independent Component Analysis (ICA), leveraging the assumption of non-Gaussian source distributions, to decompose the signals. We also used ICLabel [21] to automatically classify and remove artifactual components (e.g., eye blinks, muscle activity, etc.). As MRCPs are predominantly contained within the low-frequency range of 0.5–3.5 Hz, the EEG signals were resampled to 16 Hz (applying a low-pass antialiasing filter). Movement-related events were then identified and epochs extracted from -2 to +3 seconds around each cue onset.

User Identification: The classification task was formulated as a multi-class identification problem, performed separately for each movement condition (Fig. 1b). Its objective was to identify individual participants based on their neural signatures during specific motor tasks. In other words, the model aimed to discriminate among subjects using their EEG data, simulating a user identification scenario in which an unknown EEG pattern is matched to a corresponding individual within a predefined database. To achieve this, we employed shrinkage linear discriminant analysis (sLDA), which is

an extension of classical linear discriminant analysis. This sLDA stabilizes covariance matrix estimation by shrinking the sample covariance matrix toward a scaled identity matrix using a data-driven shrinkage coefficient λ . To perform covariance shrinkage within the LDA framework, we utilized the covShrinkage toolbox [22,23], which applies the Ledoit-Wolf closed-form solution to compute the optimal shrinkage coefficient directly from the data. The method was specifically chosen for this identification study because it effectively handles the inherent challenges of EEG classification: the high dimensionality of multi-channel time-domain signals, the limited number of available training trials per participant, and the need for robust classification in noisy biological signals, while maintaining computational efficiency for BCI applications, by selecting automatically the EEG features that hold the most predictive power [24–26].

Classification was performed in a trial-wise manner using overlapping temporal windows centered around the cue onset. Five window lengths (0s, 0.0625s, 0.125s, 0.3125s and 0.625s) were evaluated to examine the influence of temporal integration on classification accuracy, as different window sizes capture distinct aspects of brain dynamics: shorter windows preserve fine temporal detail, while longer windows integrate information over broader time scales and may improve signal-to-noise characteristics [27]. Within each window, amplitude values from all EEG channels were concatenated into a single spatiotemporal feature vector, combining both spatial (multi-channel) and temporal (multi-sample) information. Separate sLDA classifiers were trained and cross-validated for every window size and at each temporal position along the trial, allowing identification of the time point yielding optimal class separation around the cue onset. Performance was quantified through a 10×1-fold cross-validation procedure applied independently within each temporal window, providing reliable estimates of time-resolved discriminative capacity for each participant while preventing overfitting.

RESULTS

The grand average low-frequency EEG time series around the cue onset (−2 to +3 s) is shown in Fig. 2. Following the cue onset ($t = 0$ s), a clear VEP can be observed (around 0.5 s), elicited by the visual cue and superimposed on the ongoing MRCP activity (i.e., negative deflection around 1.5s after cue onset) [19]. Additionally, a smaller positive deflection appears at approximately −1.5 s, corresponding to visual responses elicited by the presentation of the fixation cross at the beginning of each trial.

The time-resolved classification performance is illustrated in Fig. 3, which presents the cross-validated identification accuracy around the cue onset ($t = 0$ s). Five causal EEG window lengths were evaluated (0 s, 0.0625 s, 0.125 s, 0.3125 s, and 0.625 s), shown in distinct colors, while dotted lines indicate the

corresponding mean chance levels [28]. The results demonstrate that identification accuracy varies dynamically over time, increasing after cue presentation and reaching its peak shortly thereafter. The highest accuracy occurs in the early post-cue interval, where VEPs and MRCPs overlap, providing stronger discriminative features. A slight increase in accuracy is also observable at the beginning of the trial, likely due to the fixation cross presentation, which elicits an early VEP contributing additional subject-specific information.

The peak accuracy values for each movement condition are summarized in Tab. 1, which lists the corresponding window length and time point (after cue onset) at which maximum performance was achieved. Across all movements, the 0.3125-second window consistently yielded the best results. Among the five attempted movements, the “hand open” condition achieved the highest identification accuracy of 40.28%, corresponding to 4.03 times the 10% chance level for ten participants.

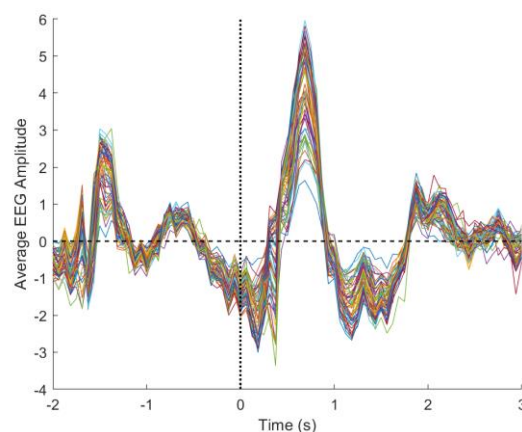


Figure 2: Grand average low-frequency EEG time series across all channels, aligned to the cue onset ($t = 0$ s) and displayed over the [−2, 3] s interval. The plot shows the mean EEG activity used to examine cue-related visual and motor responses across participants.

Table 1: Classification results for all five attempted movements, showing the highest identification accuracy along with the corresponding time point and window length.

Movement	Accuracy (%)	Window Length (s)	Timestamp (s)
Supination	34.02	0.3125	0.6875
Pronation	24.74	0.3125	0.6250
Hand open	40.28	0.3125	0.6875
Palmar grasp	27.07	0.3125	0.7500
Lateral grasp	29.44	0.3125	0.6875

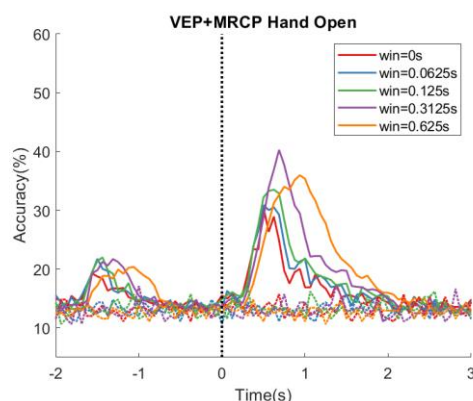


Figure 3: Time-resolved cross-validated classification accuracy for participant discrimination based on combined VEP+MRCP features around the cue onset ($t=0s$). Solid lines represent five causal window sizes: 0s, 0.0625s, 0.125s, 0.3125s and 0.625s. Dotted lines indicate the corresponding estimated mean chance levels for each window size [28]. The vertical dashed line marks the cue onset at $t=0s$. Peak accuracy ($\sim 40.3\%$) was achieved immediately after cue presentation using the 0.3125 s window. Time frame: -2 to 3 seconds relative to cue onset.

DISCUSSION

The results demonstrate that a combination of MRCPs and VEPs can effectively be used for the identification of individuals with SCI. An SCI cohort was specifically selected because their altered sensorimotor cortical activity, resulting from injury-induced neuroplasticity, provides a challenging and clinically relevant model to evaluate whether EEG-based biometric systems remain robust across neurologically distinct profiles. This finding is particularly significant, as it suggests that distinct gesture-related neural patterns, as also observed by Ofner et al [19], can serve as reliable user-specific features. Beyond its clinical relevance for improving the day-to-day life of SCI individuals, this approach also holds promise for enhancing the security and integrity of BCI systems by enabling robust, user-specific authentication and reducing potential misuse or unauthorized access. Also, the consistent superior performance of the 0.3125 window approach across all five movement types suggests that immediate neural responses following cue presentation contain the most discriminative information for user identification. Furthermore, the time-resolved performance, where identification accuracy drops to near chance levels outside the cue window, strongly indicates that the model is capturing true event-related neural dynamics rather than constant session-specific artifacts such as cap positioning.

While the present study demonstrates the feasibility of using combined MRCP and VEP features for user identification in individuals with SCI, several limitations should be acknowledged. First, the relatively small sample size (ten participants) constrains the generalizability of the findings and may limit the robustness of the classification results. As noted in [20], increasing the number of subjects typically enhances model performance in unbiased analyses, as larger datasets capture a broader range of interindividual

variability and reduce overfitting. Similarly, prior work has shown that classification accuracy tends to drop sharply when the number of samples per class falls below 15, emphasizing the need for larger and more diverse participant cohorts in future studies [29].

Another limitation concerns the use of a single classifier (sLDA). While sLDA provides computational efficiency and robustness for small-sample, high-dimensional EEG data, exploring alternative or complementary algorithms could further improve performance. Methods such as Support Vector Machines (SVMs), Common Spatial Pattern (CSP) analysis, or even deep learning approaches (e.g., convolutional neural networks (CNNs)) may more effectively capture nonlinear or spatiotemporal dependencies in EEG signals.

Future research should therefore focus on expanding the dataset, increasing participant diversity, and systematically comparing classification methods to identify the most effective strategies for EEG-based user identification. Additionally, incorporating multi-session recordings could help evaluate the stability and reproducibility of neural signatures over time, an essential step toward real-world deployment of secure and reliable BCI-based biometric systems.

CONCLUSION

This research successfully demonstrates the potential of combining MRCPs and VEPs as a novel biometric identification method for spinal cord injured individuals. This work establishes a promising foundation for developing practical BCI-based identification systems that could significantly impact assistive technology and security applications for individuals with spinal cord injuries.

REFERENCES

- [1] Zhang H, Jiao L, Yang S, Li H, Jiang X, Feng J, et al. Brain-computer interfaces: the innovative key to unlocking neurological conditions. *Int J Surg* 2024;110:5745–62.
- [2] Edelman BJ, Zhang S, Schalk G, Brunner P, Müller-Putz G, Guan C, et al. Non-Invasive Brain-Computer Interfaces: State of the Art and Trends. *IEEE Rev Biomed Eng* 2025;18:26–49.
- [3] Luck SJ. Event-Related Potentials. n.d.
- [4] Odom JV, Bach M, Brigell M, Holder GE, McCulloch DL, Mizota A, et al. ISCEV standard for clinical visual evoked potentials: (2016 update). *Documenta Ophthalmologica* 2016;133:1–9.
- [5] Sur S, Sinha V. Event-related potential: An overview. *Ind Psychiatry J* 2009;18:70.
- [6] Al Rousan M, Intrigila B. A Comparative Analysis of Biometrics Types: Literature Review. *Journal of Computer Science* 2020;16:1778–88.
- [7] Palaniappan R. Method of identifying individuals using VEP signals and neural

- network. IEE Proceedings: Science, Measurement and Technology, vol. 151, 2004, p. 16–20.
- [8] Zhao H, Chen Y, Pei W, Chen H, Wang Y. Towards online applications of EEG biometrics using visual evoked potentials. *Expert Syst Appl* 2021;177.
- [9] Barros Barbosa I, Vilhelmsen K, Van Der Meer A, Van Der Weel R, Theoharis T. EEG Biometrics: On the Use of Occipital Cortex Based Features from Visual Evoked Potentials. n.d.
- [10] Das R, Maiorana E, Campisi P. EEG Biometrics Using Visual Stimuli: A Longitudinal Study. *IEEE Signal Process Lett* 2016;23:341–5.
- [11] Huang Z, Liao Z, Ou G, Chen L, Zhang Y. Authentication using c-VEP evoked in a mild-burdened cognitive task. *Front Hum Neurosci* 2023;17.
- [12] Savić AM, Lontis ER, Mrachacz-Kersting N, Popović MB. Dynamics of movement-related cortical potentials and sensorimotor oscillations during palmar grasp movements. *European Journal of Neuroscience* 2020;51:1962–70.
- [13] Pereira J, Sburlea AI, Müller-Putz GR. EEG patterns of self-paced movement imaginations towards externally-cued and internally-selected targets. *Sci Rep* 2018;8.
- [14] Schwarz A, Escolano C, Montesano L, Müller-Putz GR. Analyzing and Decoding Natural Reach-and-Grasp Actions Using Gel, Water and Dry EEG Systems. *Front Neurosci* 2020;14.
- [15] Li H, Huang G, Lin Q, Zhao JL, Lo WLA, Mao YR, et al. Combining movement-related cortical potentials and event-related desynchronization to study movement preparation and execution. *Front Neurol* 2018;9.
- [16] Schurger A, Hu P Ben, Pak J, Roskies AL. What Is the Readiness Potential? *Trends Cogn Sci* 2021;25:558–70.
- [17] Olsen S, Alder G, Williams M, Chambers S, Jochumsen M, Signal N, et al. Electroencephalographic Recording of the Movement-Related Cortical Potential in Ecologically Valid Movements: A Scoping Review. *Front Neurosci* 2021;15.
- [18] Shakeel A, Navid MS, Anwar MN, Mazhar S, Jochumsen M, Niazi IK. A review of techniques for detection of movement intention using movement-related cortical potentials. *Comput Math Methods Med* 2015;2015.
- [19] Ofner P, Schwarz A, Pereira J, Wyss D, Wildburger R, Müller-Putz GR. Attempted Arm and Hand Movements can be Decoded from Low-Frequency EEG from Persons with Spinal Cord Injury. *Sci Rep* 2019;9.
- [20] Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 2004;134:9–21.
- [21] Pion-Tonachini L, Kreutz-Delgado K, Makeig S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* 2019;198:181–97.
- [22] Crell MR, Kostoglou K, Sterk K, Müller-Putz GR. A novel paradigm for fast training data generation in asynchronous movement-based BCIs. *Front Hum Neurosci* 2025;19.
- [23] Kostoglou K, Muller-Putz GR. Motor-Related EEG Analysis Using a Pole Tracking Approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2024;32:3837–47.
- [24] Ye J, Xiong T, Li Q, Janardan R, Bi J, Cherkassky V, et al. Efficient Model Selection for Regularized Linear Discriminant Analysis. 2006.
- [25] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, et al. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *J Neural Eng* 2018;15.
- [26] Blankertz B, Lemm S, Treder M, Haufe S, Müller KR. Single-trial analysis and classification of ERP components - A tutorial. *Neuroimage* 2011;56:814–25.
- [27] Orkan Olcay B, Karaçalı B. Time-resolved EEG signal analysis for motor imagery activity recognition. *Biomed Signal Process Control* 2023;86.
- [28] Müller-Putz GR, Scherer R, Pfurtscheller G, Brunner C, Leeb R. Better than Random? A closer look on BCI results. vol. 10. 2008.
- [29] Bomatter P, Gouk H. Is Limited Participant Diversity Impeding EEG-based Machine Learning? 2025.

Deep Sleep Classification via EEG Signal Criticality: A Passive BCI Approach for Sleep-Improvement Neurofeedback

Stanisław Narębski¹, Tomasz Komendziński¹, Tomasz M. Rutkowski^{2,3,1}

¹Nicolaus Copernicus University, Toruń, Poland

²ARAYA Inc., Tokyo, Japan

³The University of Tokyo, Tokyo, Japan

E-mail: stanislaw@bci-lab.info & tomek@bci-lab.info

ABSTRACT: Automated sleep staging is a fundamental application of passive Brain-Computer Interfaces (pBCI), decoding spontaneous neural states to enable closed-loop interventions independent of user intent. This study evaluates criticality features derived from Detrended Fluctuation Analysis (DFA) for the specific identification of deep sleep (N3).

We analyzed 347,232 EEG epochs from 290 older women using UMAP manifold learning to visualize state transitions. Subsequently, six classifiers were benchmarked via 10-fold cross-validation, using balanced accuracy to determine the optimal "state-sensing" engine for neurofeedback. Naive Bayes achieved the highest mean balanced accuracy ($87.17\% \pm 0.24\%$), significantly outperforming a fully connected deep neural network (FNN: 81.58%) and Random Forest (80.97%). Linear models (LDA: 57.21%; SVM: 51.01%) performed poorly, indicating that DFA-derived criticality features reside on a distinct, non-linear manifold.

Probabilistic decoding of EEG criticality provides a high-accuracy sensing mechanism for pBCIs. This robust classification pipeline supports the development of state-dependent neurofeedback, such as targeted auditory stimulation, to enhance cognitive recovery.

INTRODUCTION

The conceptualization of Brain-Computer Interfaces (BCI) has traditionally been dominated by "active" paradigms [1, 2], where users exert conscious effort to modulate neural activity for external control. However, as the field shifts toward everyday neurotechnology, the Passive BCI (pBCI) has emerged as a superior framework for state-monitoring and autonomous intervention [3, 4]. Perhaps the most robust, yet underappreciated, "textbook example" of a pBCI is automated sleep staging [5, 6]. Unlike active BCIs that require task-related intent, a sleep-based pBCI monitors the brain's spontaneous, ongoing neural oscillations to infer physiological state without any voluntary participation from the user [7].

By definition, a BCI is classified as "passive" when it satisfies three core criteria: the use of spontaneous brain activity as input, the absence of conscious intent from the user, and the output being a state estimation rather

than a direct command [3]. Sleep staging maps perfectly to this architecture. During sleep, the user is unconscious and unable to voluntarily modulate neural patterns. The system, therefore, must decode "hidden" physiological markers—such as sleep spindles, K-complexes, and slow-wave activity—to classify the internal state (e.g., N3 vs. REM) [8].

In recent years, the utility of sleep staging has evolved from retrospective clinical diagnosis to real-time, "closed-loop" intervention [7]. This transformation is best exemplified by Closed-Loop Auditory Stimulation (CLAS) [9–11]. In these systems, a pBCI architecture acquires EEG signals in real-time, identifies the specific "up-state" of a slow oscillation during deep sleep (N3) [8], and triggers a millisecond-long auditory pulse. This intervention is designed to enhance the amplitude of slow waves, thereby improving memory consolidation and metabolic health [7]. In this context, the staging algorithm is the "decoder" of the BCI, providing the necessary state-inference to trigger a neurofeedback loop entirely below the level of conscious awareness [9–11].

The Critical Brain Hypothesis [12] describes the idea that our nervous system organizes itself to be near a phase transition between order and disorder, a point analogous to the boundary between a solid and a liquid in physical systems. At this critical point, neural networks exhibit a characteristic set of emergent properties: optimal information processing, scale-free architecture (both in time as well as space), high sensitivity to external stimuli and flexibility in transitioning between states. Based on this we can predict that healthy brains will be closest to criticality with disease moving us away from this [13]. Furthermore, criticality has been shown to diminish during waking hours and to be restored during sleep [14].

While neural signals often exhibit complex heterogeneity, DFA provides a robust framework for quantifying long-range temporal correlations (LRTC) by identifying a single scaling exponent that governs signal fluctuations across time scales, which serve as a proxy for signal criticality. Previous studies indicate that disease states, specifically Alzheimer's disease and related dementias, are associated with a breakdown in the brain's capacity to maintain flexible, long-range correlations, often resulting in a shift of the scaling exponent toward values indica-

tive of decreased physiological complexity [15–17]. By quantifying these shifts, we can evaluate the predictive utility of EEG criticality in identifying early-stage cognitive decline.

The current study builds upon the pBCI framework by leveraging EEG signal criticality [15, 18, 19] as a primary feature set for deep sleep classification. Criticality—the study of neural systems operating at a phase transition between order and disorder—offers a unique window into the brain’s self-organization during the descent into deep sleep [20]. By utilizing criticality-derived metrics—specifically channel-wise DFA—as features for machine learning, we propose a high-accuracy, automated pipeline for the detection of N3 sleep stages. We further employ UMAP manifold learning to characterize the geometric structure of the resulting feature space, and benchmark six classifiers spanning linear, probabilistic, ensemble, and deep learning paradigms. This approach serves as the foundational “sensing” component for a subsequent sleep-improvement neurofeedback system, aiming to optimize the timing and efficacy of state-dependent interventions for the enhancement of human recovery and, ultimately, long-term cognitive health [7].

MATERIALS AND METHODS

The research dataset was acquired following proper authorization from the National Sleep Research Resource (NSRR) [21], drawing upon a selected portion [22] of the Study of Osteoporotic Fractures (SOF) records [23]. SOF participants were American women at least 65 old, with precise age unavailable due to de-identification. Overnight polysomnography (PSG) assessments were carried out in subjects’ homes employing Compumedics P-Series equipment, which captured four electroencephalography (EEG) channels from 461 subjects. To establish a more uniform baseline and allow meaningful comparison, the present investigation included only those participants who obtained a Mini-Mental State Examination (MMSE) [24] score exceeding 24 during PSG administration and who also attended a subsequent evaluation five years later ($N = 290$). This cohort was divided into three groups according to their results on the Teng-modified Mini-Mental State Examination (3MS) [25] administered at follow-up. Individuals obtaining a score of 88 or above were designated as cognitively normal, those scoring between 78 – 88 were identified as having Mild Cognitive Impairment (MCI), and those scoring below 78 were classified as having dementia. The polysomnography electrode configuration comprised C3, C4, A1, and A2 channels (representing central and mastoid sites) with a sampling frequency of 128 Hz.

Subsequently, these recordings were divided into 30-second epochs and hand-scored by qualified technicians to classify wakefulness, rapid eye movement (REM) sleep, or non-REM phases (N1, N2, N3 and N4), following Rechtschaffen and Kales scoring standards [26]. For this study, N3 and N4 stages as originally scored in SOF

were merged into a unified N3 stage consistent with contemporary clinical practice [27]. In addition, the awake state was partitioned into pre-sleep - starting from the first epoch until the first recorded sleep stage and post-sleep - the last awake state not followed by any more sleep. For six participants, a singular last sleep epoch was removed from analysis.

The following analysis received ethical approval from the Nicolaus Copernicus University in Toruń, Poland, Faculty of Philosophy Research Ethics Committee (Decision No. 60/2025).

EEG Criticality Estimation with Detrended Fluctuation Analysis (DFA): DFA captures how signal fluctuations scale across time, yielding a single exponent that reflects the degree of long-range temporal correlations (LRTC) in EEG signal. In the context of sleep, this scaling exponent serves as a proxy for criticality: as the brain transitions into N3, theory predicts a characteristic shift toward long-range ordered dynamics, distinct from the activity during other sleep stages. We also aimed to verify whether the differences in criticality remain a reliable indicator of N3 regardless of cognitive status across healthy aging, early-stage cognitive decline and dementia. The DFA procedure for each 30-second EEG epoch was implemented in four primary steps:

1. **Integration:** The original time series $x(i)$ is transformed into an integrated profile $Y(i) = \sum_{k=1}^i [x(k) - \langle x \rangle]$ to convert the noise-like signal into a random walk-like profile, revealing underlying scaling properties.
2. **Segmentation:** The integrated profile is partitioned into $N_n = \text{int}(N/n)$ non-overlapping segments of equal length n .
3. **Local Detrending:** For each segment v , a local polynomial trend $y_{v,n}$ (of order $k = 1$) is calculated via least-squares fitting and subtracted from the integrated profile to remove non-stationary “drifts.”
4. **Fluctuation Function Calculation ($q=2$):** The root-mean-square fluctuation is computed by averaging the variance over all segments:

$$F(n) = \sqrt{\frac{1}{2N_n} \sum_{v=1}^{2N_n} [F^2(v, n)]}. \quad (1)$$

The output is characterized by the scaling exponent α (equivalent to the Hurst exponent H), determined as the slope of $\log F(n)$ versus $\log n$. A value of $\alpha \approx 0.5$ indicates white noise (randomness), while $\alpha \approx 1.0$ suggests $1/f$ noise, characteristic of healthy, self-organized criticality. Significant deviations from these values serve as primary metrics for pathological decline and loss of signal complexity.

Manifold Visualization of Sleep Dynamics: To visualize the manifold structure of the DFA scaling features and assess their distribution across sleep architec-

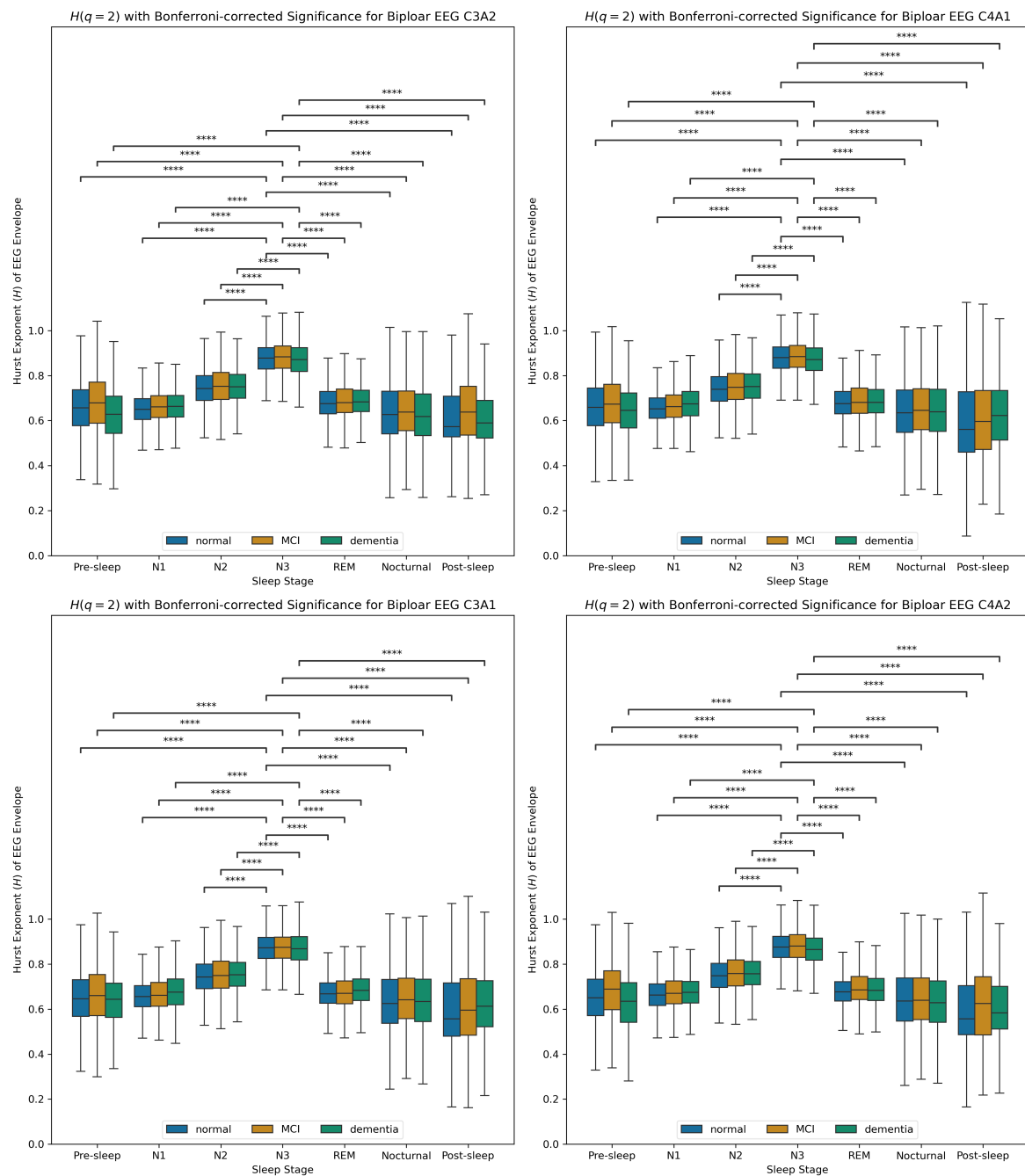


Figure 1: The Hurst exponent distributions ($q = 2$) obtained from Detrended Fluctuation Analysis (DFA) of the EEG amplitude envelope for all sleep stages and cognitive groups. The four panels show values calculated from electrode pairs C3A2, C4A1, C3A1 and C4A2. Above the plots there are Bonferroni-corrected pairwise significance bars showing that the N3 stage exhibits significantly elevated H values relative to all other sleep stages across all channels and cognitive groups (**** : $p \leq 0.0001$).

ture, we applied unsupervised Uniform Manifold Approximation and Projection (UMAP) [28]. Unlike supervised variants that rely on categorical labels, unsupervised UMAP constructs a low-dimensional embedding based solely on the underlying topological structure of the data. This allows for an unbiased observation of how neural scaling properties naturally cluster across different sleep stages. Data for this visualization were derived from the sleep polysomnography subset [22] of the Study of Osteoporotic Fractures (SOF) database [21,

23]. The resulting embedding highlights a clear topological separation of the N3 sleep stage from the continuum of awake, N1, N2, and REM sleep. Mathematically, UMAP optimizes a low-dimensional fuzzy simplicial set to match the high-dimensional topology. In our analysis, the N3 stage—characterized by high-amplitude delta activity and distinct long-range temporal correlations—occupies a restricted, high-density region of the manifold. This isolation reflects the unique "criticality" of deep sleep compared to the more stochastic or frag-

Unsupervised Clustering 3D UMAP Embedding (347232 data points)

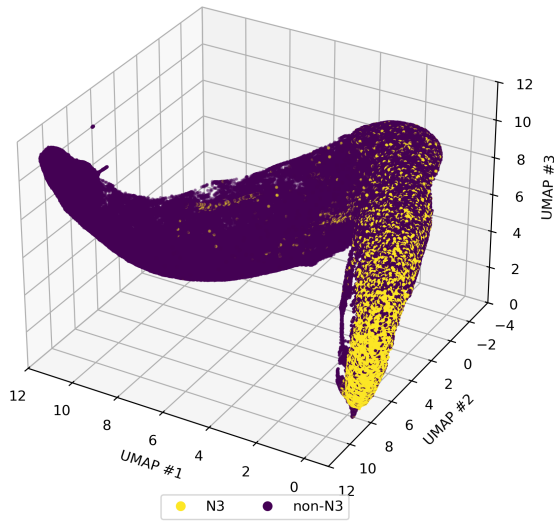


Figure 2: Unsupervised Uniform Manifold Approximation and Projection results for all 347,232 epochs, where yellow represents the N3 stage and purple non-N3 stages. The three-dimensional projection of the high-dimensional feature set created from DFA analysis of four-channel EEG sleep data reveals a distinct cluster of N3 epochs. The non-linear, curved topology of the embedding indicates that the N3 stage is not linearly separable from the rest of the sleep stages.

mented dynamics of lighter sleep stages. By examining these clusters, we can visualize how the breakdown of $q = 2$ scaling exponents specifically impacts the N3 signature in the context of cognitive decline, where the once-distinct N3 cluster typically begins to merge with or shift toward less complex sleep states.

Machine Learning Framework and Model Selection: To evaluate the discriminative power of the DFA scaling exponents across different sleep stages, we implemented a comparative machine learning framework. We selected a diverse ensemble of six classifiers to ensure that our findings were robust to specific algorithmic biases, spanning linear, instance-based, probabilistic, and connectionist architectures:

Linear Models: Linear Discriminant Analysis (LDA) and Linear Support Vector Classification (Linear SVC) were employed to establish a baseline for linear separability in the feature space.

Non-Linear & Ensemble Models: Random Forest (100 estimators, max depth of 10) was used to capture hierarchical feature interactions, while K-Nearest Neighbors (KNN, $k = 5$) assessed local topological clustering.

Probabilistic & Neural Models: Gaussian Naive Bayes was employed to provide a baseline for conditional probability-based classification. Additionally, a four-layer Feedforward Neural Network (FNN) consisting of three 100-neuron hidden layers and a final 50-neuron layer with ReLU activation was

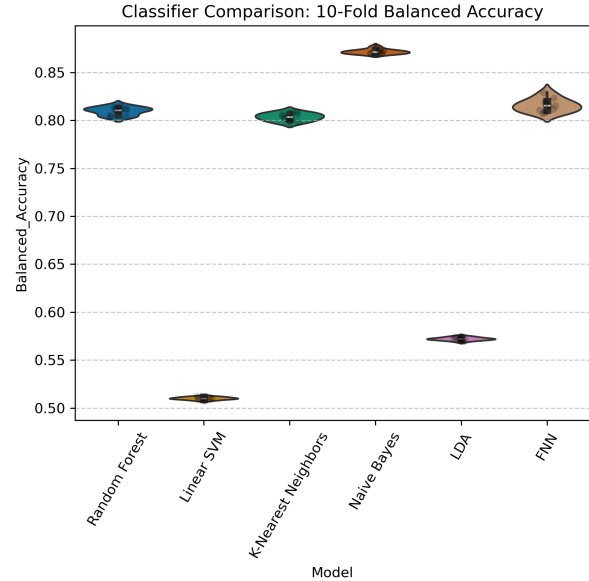


Figure 3: Balanced accuracy results from 10-fold cross-validation for binary classification of N3 versus non-N3 sleep across six evaluated classifiers: Random Forest, Linear SVM, K-Nearest Neighbors, Naive Bayes, Linear Discriminant Analysis (LDA), and a fully connected deep neural network (FNN). The highest score was achieved by Naive Bayes with accuracy of 87.17% ($\pm 0.24\%$). The full list of values can be seen in Table 1.

implemented. This latter architecture was designed to capture high-dimensional, non-linear dependencies, following the successful application of a similar model in prior work [15].

Validation and Performance Metrics: To mitigate the risk of overfitting and ensure the generalizability of the results, we employed a 10-fold Stratified Cross-Validation ($10 \times$ SKF) scheme. Stratification was used to preserve the class distributions (e.g., N3 versus non-N3 sleep) across each fold. Given the potential for class imbalance in sleep stage distribution and clinical diagnostic groups, model performance was quantified using Balanced Accuracy. This metric is defined as the arithmetic mean of class-specific recall:

$$\text{Balanced Accuracy} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right), \quad (2)$$

where TP , FP , TN , and FN represent true positives, false positives, true negatives, and false negatives, respectively. All features were standardized using a StandardScaler within the cross-validation pipeline to ensure zero mean and unit variance, preventing models like SVC and FNN from being biased by differing feature scales.

RESULTS

Analysis of the Hurst exponent from DFA analysis of the EEG envelope ($H, q = 2$) across all 347,232 epochs revealed a consistent and statistically robust elevation of H values during the N3 sleep stage relative to all other

Table 1: Ten-fold cross-validation balanced accuracy results

Model	Mean balanced accuracy
Naive Bayes	0.871667 $\pm$ 0.002397
FNN	0.815764 $\pm$ 0.006825
Random Forest	0.809687 $\pm$ 0.004126
K-Nearest Neighbors	0.803512 $\pm$ 0.003902
LDA	0.572146 $\pm$ 0.001600
Linear SVM	0.510112 $\pm$ 0.001382

recorded states. As illustrated in Figure 1, this pattern was observed uniformly across all four bipolar EEG derivations (C3A2, C4A1, C3A1, and C4A2) and across all three cognitive groups (cognitively normal, MCI, and dementia). Bonferroni-corrected pairwise comparisons confirmed that N3 H values were significantly elevated compared to pre-sleep wakefulness, N1, N2, REM, nocturnal wakefulness, and post-sleep wakefulness (all $p \leq 0.0001$).

Unsupervised UMAP applied to the full 347,232-epoch DFA feature set produced a three-dimensional embedding that revealed a geometrically distinct cluster corresponding to N3 sleep epochs (Figure 2). N3 epochs (shown in yellow) occupied a compact, high-density region sharply separated from the broader manifold of non-N3 sleep stages (shown in purple), which formed a more diffuse, curved continuum.

Six classifiers were evaluated on a binary N3 versus non-N3 classification task using ten-fold stratified cross-validation with balanced accuracy as the primary performance metric. Results are summarized in Table 1 and visualized in Figure 3. Gaussian Naive Bayes achieved the highest mean balanced accuracy of 87.17% ($\pm 0.24\%$), substantially outperforming all other evaluated models. The Feedforward Neural Network (FNN) attained the second-highest performance at 81.58% ($\pm 0.68\%$), followed closely by Random Forest at 80.97% ($\pm 0.41\%$) and K-Nearest Neighbors at 80.35% ($\pm 0.39\%$). These four non-linear and probabilistic classifiers reached higher accuracy relative to the linear models. Linear Discriminant Analysis (LDA) achieved only 57.21% ($\pm 0.16\%$) balanced accuracy, while Linear SVM performed the worst at 51.01% ($\pm 0.14\%$).

DISCUSSION

This study investigated the utility of EEG signal criticality as a feature set for automated deep sleep (N3) classification within a passive BCI (pBCI) framework. The findings demonstrate that criticality-derived features provide a high-accuracy, robust sensing mechanism for N3 detection, achieving a mean balanced accuracy of 87.17% with Naive Bayes classification. The elevated Hurst exponent values observed specifically during N3 sleep, consistently across all four EEG channels and all three cognitive groups, provide strong empirical support for the theoretical prediction that deep sleep serves a criticality-restorative function. Notably, this discriminative signature was preserved across cognitive status groups, sug-

gesting that DFA-derived criticality remains a robust indicator of N3 sleep even in the presence of cognitive decline. While the exclusive inclusion of older women represents a demographic limitation that constrains the generalisability of these findings to the broader population, it simultaneously constitutes a strength in the context of the study's future aims - this cohort closely mirrors the target demographic for home-based sleep intervention and dementia prevention, where older women are both disproportionately represented and at elevated risk. The UMAP embedding provided a critical insight that directly motivated and subsequently validated the classifier selection strategy. The geometrically distinct but topologically curved cluster occupied by N3 epochs in the three-dimensional embedding demonstrates unambiguously that deep sleep is not linearly separable from other stages in the DFA feature space. This curvature implies that the decision boundary required for accurate N3 detection is non-linear, which explains the failure of both linear models evaluated. The poor performance of the Linear SVM (51.01%) and the LDA (57.21%) models are indicative of a mismatch between model architecture and feature geometry.

CONCLUSIONS

The findings of this study reinforce the conceptualization of automated sleep staging as a premier use case for pBCI. By leveraging neural criticality—specifically through channel-wise DFA—we have demonstrated that spontaneous, unconscious oscillations contain high-fidelity signatures of physiological state. The robust performance of non-linear classifiers, such as the Naive Bayes model (87.17% balanced accuracy), validates that these criticality-derived features can reliably decode the "hidden" markers of N3 sleep across diverse cognitive profiles. This architecture satisfies the definitive criteria of a pBCI: it utilizes spontaneous brain activity without conscious intent to produce a state estimation rather than a direct command. Our results show that the geometric complexity of the DFA feature space, as revealed by UMAP manifold learning, necessitates non-linear decoding strategies to accurately capture the brain's transition into deep sleep. Ultimately, this pipeline provides the necessary "sensing" foundation for closed-loop interventions. By accurately identifying the N3 stage in real-time, such a system can trigger neurofeedback or auditory stimulation entirely below the level of conscious awareness. This study thus moves the field closer to a seamless, autonomous neurotechnology capable of monitoring and restoring neural criticality, thereby optimizing human recovery and long-term cognitive health.

REFERENCES

- [1] Wolpaw J, Wolpaw EW, Eds. Brain-computer interfaces: Principles and practice. Oxford University Press: New York, USA (2012).

- [2] Rutkowski T. Robotic and virtual reality BCIs using spatial tactile and auditory oddball paradigms. *Frontiers in Neurorobotics*. 2016;10:20.
- [3] Zander TO, Kothe C. Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general. *Journal of Neural Engineering*. 2011;8(2):025005.
- [4] Rutkowski TM, Zhao Q, Abe MS, et al. Passive BCI for task-load and dementia biomarker elucidation. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE Press: Berlin, Germany, 2019, ThC01.1.
- [5] Rutkowski TM. Sleep monitoring - multivariate and multimodal brain and peripheral body signal processing methods for sleep and consciousness level assessment. In: Abstract Book of the Third APSIPA Workshop on the Frontier in Biomedical Signal Processing and Systems (APSIPA BioSiPS 2015). 2015, 5–6.
- [6] Rutkowski TM. Automatic sleep staging and apnea events classification from EEG and multimodal physiological signals – synchrosquezing transform processing and Riemannian geometry classification approaches. In: The 4th Annual IIS Symposium – Poster Session Abstracts. Tsukuba, Japan, 2016, 6.
- [7] Narebski S, Komendzinski T, Rutkowski TM. Perspective: Home-based sleep intervention for dementia prevention. In: Extended Abstracts of The 8th Annual Conference on Cognitive Computational Neuroscience. Amsterdam, The Netherlands, 2025, A112.
- [8] Iber C, Ancoli-Israel S, Chesson A, Quan SF. The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. American Academy of Sleep Medicine: Westchester, IL (2007).
- [9] Ngo HVV, Martinetz T, Born J, Mölle M. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron*. 2013;78(3):545–553.
- [10] Santostasi G et al. Phase-locked loop for precisely timed acoustic stimulation during sleep. *Journal of neuroscience methods*. 2016;259:101–114.
- [11] Papalambros N et al. Acoustic enhancement of sleep slow oscillations and concomitant memory improvement in older adults. *Frontiers in Human Neuroscience*. 2017;11:109.
- [12] Beggs JM. The criticality hypothesis: How local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2008;366(1864):329–343.
- [13] Massobrio P, De Arcangelis L, Pasquale V, Jensen HJ, Plenz D. Criticality as a signature of healthy neural systems. *Frontiers in systems neuroscience*. 2015;9:22.
- [14] Pearlmutter BA, Houghton CJ. A new hypothesis for sleep: Tuning for criticality. *Neural computation*. 2009;21(6):1622–1641.
- [15] Rutkowski TM, Abe MS, Tokunaga S, Komendzinski T, et al. Dementia digital neuro-biomarker study from theta-band EEG fluctuation analysis in facial and emotional identification short-term memory oddball paradigm. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). IEEE Press: Glasgow, UK, 2022, 4129–4132.
- [16] Rutkowski TM, Abe MS, Komendzinski T, Sugimoto H, Narebski S, et al. Machine learning approach for early onset dementia neurobiomarker using eeg network topology features. *Frontiers in Human Neuroscience*. 2023;17.
- [17] Rutkowski TM, Komendziński T, et al. Mild cognitive impairment prediction and cognitive score regression in the elderly using EEG topological data analysis and machine learning with awareness assessed in affective reminiscent paradigm. *Frontiers in Aging Neuroscience*. 2024;15.
- [18] Linkenkaer-Hansen K, Nikouline VV, Palva JM, Ilmoniemi RJ. Long-range temporal correlations and scaling behavior in human brain oscillations. *Journal of Neuroscience*. 2001;21(4):1370–1377.
- [19] Rutkowski TM, Abe MS, et al. Neurotechnology and AI approach for early dementia onset biomarker from EEG in emotional stimulus evaluation task. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). 2021, 6675–6678.
- [20] Scarpetta S et al. Criticality of neuronal avalanches in human sleep and their relationship with sleep macro- and micro-architecture. *science*, 26 (10), 107840. 2023.
- [21] Zhang GQ et al. The national sleep research resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*. 2018;25(10):1351–1358.
- [22] Spira AP et al. Sleep-disordered breathing and cognition in older women. *Journal of the American Geriatrics Society*. 2008;56(1):45–50.
- [23] Cummings SR et al. Appendicular bone density and age predict hip fracture in women. *JAMA*. 1990;263(5):665–668.
- [24] Folstein MF, Folstein SE, McHugh PR. “MINI-MENTAL STATE”: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*. 1975;12(3):189–198.
- [25] Teng E, Chui H. The modified mini-mental state examination (3MS). *Can J Psychiatry*. 1987;41(2):114–21.
- [26] Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. 1968.
- [27] Iber C. The AASM manual for the scoring of sleep and associated events: Rules, terminology, and technical specification. American Academy of Sleep Medicine. 2007.
- [28] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. 2018.

Execution and Omission Errors in Human–Robot Interaction: An EEG Study

J. Mu^{1,2,*}, M. Wimmer^{3,4,*}, H.S. Wimmer^{5,*}, A.N. Burkitt^{1,2}, G.R. Müller-Putz⁴, D.B. Grayden^{1,2}

¹Department of Biomedical Engineering, University of Melbourne, Melbourne, Australia

²Graeme Clark Institute for Biomedical Engineering, University of Melbourne, Melbourne, Australia

³Know Center Research GmbH, Graz, Austria

⁴Institute of Neural Engineering, Graz University of Technology, Graz, Austria

⁵Institute of Business Informatics and Data Science, FH JOANNEUM, Graz, Austria

*These authors share first authorship

E-mail: jing.mu@unimelb.edu.au, mwimmer@know-center.at

ABSTRACT: Brain-computer interface research has extensively investigated the use of error-related potentials in human-robot interaction. However, while previous research demonstrated the successful decoding of erroneous task execution from the user's electroencephalogram, errors of omission, i.e., the non-execution of an expected activity, remain underexplored. The current work investigates the cortical responses to both execution and omission errors during a robot-assisted task. The grand average event-related potentials ($N = 12$) revealed a significantly attenuated error positivity related to activity omission, likely reflecting reduced confidence in the occurrence of an error. In addition, we could decode both types of errors at a single-trial level with an average balanced accuracy of 80 % in a synchronous classification approach. The results of this work contribute to the development of autonomously corrective systems for human-robot interaction.

INTRODUCTION

Whether to improve, extend, or restore our capabilities, collaborative interaction between humans and robots is increasingly gaining attention from health sciences to industrial settings [1]. Especially for people experiencing limitations in motor function, systems that enable human-robot interaction have considerable potential to provide greater autonomy and thus improved quality of life [2]. However, limitations in the performance of robotic assistance [3], and the intrinsic challenges of adapting such complex systems to changing environments [4], still hinder advancements in the field. To alleviate these shortcomings, human-in-the-loop approaches continue to emerge in which, e.g., robot parameters are adjusted based on the current collaborative performance [5, 6]. However, these approaches require external performance metrics (e.g., ground truth and targets), which are rarely available, particularly in the context of robotic end-effectors designed to restore human capabilities. One approach to circumvent this lack of external metrics for improving performance relies on the users themselves. Specifically, previous studies in the field of brain-

computer interfaces (BCIs) [7, 8] directly leveraged human error-monitoring capabilities to automatically correct errors committed by the robot and to adapt the system itself based on user-derived feedback [9]. To this end, the sequence of error-related negativity (ERN) and error positivity (Pe) — two prominent markers of unconscious and conscious error processing in error-related potentials (ErrPs) [10] arising 100 ms and 300-500 ms after an erroneous event, respectively [11] — can be decoded non-invasively using electroencephalography (EEG). For example, a study by Salazar-Gomez and colleagues employed ErrPs to automatically correct erroneous robot movement during a binary object selection task [12]. In contrast, other work used ErrPs as reward signals in reinforcement-learning to adapt the behavior of a robotic arm or learn to navigate a mobile robot in a foreign environment, both in virtual and real-world settings [13, 14].

Overall, a considerable body of research has investigated ErrPs in the context of human-robot interaction. Nevertheless, these reports almost exclusively document ErrP decoding in relation to erroneous *execution* of an intended task [12, 15–19]. In contrast, reports on *omission* errors — errors relating to the *absence* of a reaction altogether rather than the illegitimate execution of a task [20] — remain scarce [21]. In BCI contexts, however, such aborted responses hold specific relevance for spelling devices [22], motor decoding tasks [23, 24], and other applications. Detailed investigations of neurophysiology and applicability of the different error types in human-robot-interaction remain, to the best of our knowledge, to be demonstrated.

The current study aims to close this gap by exploring cortical responses to errors in execution and errors by omission during human-robot interaction. Specifically, we aim to answer two research questions: (1) Are there differences in the ErrPs elicited by erroneous task execution and omission? (2) Can these error types be distinguished from correct task execution in BCIs? The results of this work should expand the current scope of EEG-based error detection for human-robot interaction and advance the development of adaptive and corrective systems.

MATERIALS AND METHODS

Participants: Twelve volunteers (6 male, 6 female, 28 ± 2 years (mean (M) $\pm$ standard deviation (SD)) participated in this study. All participants were right-handed and 6 were naïve to BCIs. The study was approved by the University of Melbourne Human Research Ethics Committee (Project ID: 20853). Written consent was obtained from all participants before the experiment. All participants received a \$20 AUD voucher after completion.

Data acquisition: We recorded EEG signals with 32 electrodes positioned according to the 10-10 system at AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FCz, FC2, FC6, C3, C1, Cz, C2, C4, CP5, CP1, CPz, CP2, CP6, P7, P3, Pz, P4, P8, PO7, PO3, PO4, PO8, and Oz. Ground and reference electrodes were placed at AFz and on the right ear lobe, respectively. In addition, we acquired 4-channel electroocular (EOG) signals (outer canthi of both eyes, above and below the right eye) and 4-channel electromyographic (EMG) signals (flexor digitorum profundus, referenced to the respective elbow on each arm). All signals were recorded with active electrodes (g.USBamp amplifier with g.Scarabeo electrodes, g.tec medical engineering, Austria) and synchronized with the experimental markers using Simulink (Matlab 2015a, MathWorks Inc., USA). Data were recorded at a sampling rate of 512 Hz.

Experimental setup and design: The experimental task resembled a whack-a-mole game in which participants used a robotic arm (Reachy, Pollen Robotics, France) to press a button indicated to them. The participants sat comfortably in a chair facing a two-button setup, holding a tennis ball in each hand (Fig. 1). Upon illumination of one of the buttons, they were instructed to squeeze the ball in the corresponding hand. The increase in EMG activity (see *EMG data processing*) triggered the robot to start a predefined sequence of movements (i.e., *movement initiated*) to press the button. First, the robot moved toward the selected button and stopped momentarily at the via point above the button (i.e., *above button*). Next, the robot moved downward to press the button (400 ms), which turned off the light and finally returned to the home position (*correct* task execution). In *error* trials, the robot would either move to the wrong button (erroneous task execution; **ErrMov**) or hover above the correct button for 500 ms without pressing the button (omitted task execution; **NoMov**) before returning to the home position. The experimental protocol encompassed 10 blocks of 40 trials (12 *error*: 6 **ErrMov** and 6 **NoMov**, 28 *correct*), with left and right trials balanced and a randomized trial sequence. The participants were not instructed regarding erroneous robot behavior and hence expected a correctly functioning system.

EMG data processing: EMG signals were decoded online using Matlab 2015a with Simulink to initiate actions of the robotic arm. Specifically, the root mean square (RMS) of the EMG signal was continuously computed using a 250 ms sliding window shifted one sample at a

time. When the RMS exceeded a predefined threshold for either arm, the corresponding robotic movement was triggered. Thresholds were manually tuned for each participant prior to the experiment.

EEG data processing: Offline data processing and analysis were performed in Matlab 2015a and Python using the EEGLAB toolbox [25] and MNE-Python [26], respectively. First, power-line noise was removed using a notch filter at 50 Hz (Butterworth, 2nd order, non-causal). Next, data were band-pass filtered between 1 and 30 Hz (Butterworth, 4th order, non-causal) and downsampled to 128 Hz. Bad channels were identified based on kurtosis and joint probability (both ± 5 SDs) [27] and spherically interpolated. We filtered the data in the 1-10 Hz range (MNE-Python default filter settings) to focus on low frequencies relevant to error processing [9]. Subsequently, we performed independent component analysis using the FastICA algorithm [28] to remove ocular artifacts based on visual inspection. Then, we segmented the data into epochs of 4 s (i.e., [-1, 3] s relative to *movement initiated*). Epochs with residual contamination were removed based on amplitude (exceeding ± 100 μ V), kurtosis, and joint probability (both ± 5 SDs). Finally, we extracted segments of 1.75 s, i.e., [-0.5, 1.25] s around both *movement initiated* and *above button*, related to **ErrMov** and **NoMov**, respectively. Epochs of each error type were corrected with the corresponding [-0.25, 0] s baselines.

After data preprocessing, a total of 258 ± 13 *correct* trials, as well as 54 ± 5 (**ErrMov**) and 57 ± 3 (**NoMov**) *error* trials remained per participant (all M $\pm$ SD). On average, we rejected 8% of the recorded trials.

Classification: We implemented the same trial-wise and stimulus-locked classification strategy (*error* vs. *correct*) for **ErrMov** and **NoMov**. Classification features were derived from the time domain signals in the [0.2, 1] s segment after *movement initiated* or *above button*, respectively. Specifically, we computed the average amplitude of overlapping windows with a window length of 100 ms and an overlap of 50 ms (i.e., [0.2, 0.3] s, [0.25, 0.35] s, etc.) [29]. For each participant, we iteratively trained and tested personalized shrinkage linear discriminant analysis (sLDA) classifiers [30] using a stratified 10 times 5-fold cross-validation (CV) scheme with random sampling. The classification threshold was adjusted to the prior probability of the minority class (i.e., *error*) to account for the class imbalance and fixed at 0.15 [31]. To summarize the decoding performance, we computed the true positive rate (TPR; proportion of errors correctly classified), the true negative rate (TNR; proportion of non-errors correctly classified), and the balanced accuracy (BA; the average of TPR and TNR). The reported metrics are the averages of the 50 held-out sets.

Statistics: We performed sample-wise Wilcoxon signed-rank tests (two-sided, $\alpha = 0.05$) to compare event-related potentials (ERPs) of *correct* and *error* events at Cz and controlled the false discovery rate (FDR) with the Benjamini-Hochberg procedure [32]. Similarly, we used Wilcoxon signed-rank tests and subsequent FDR correc-

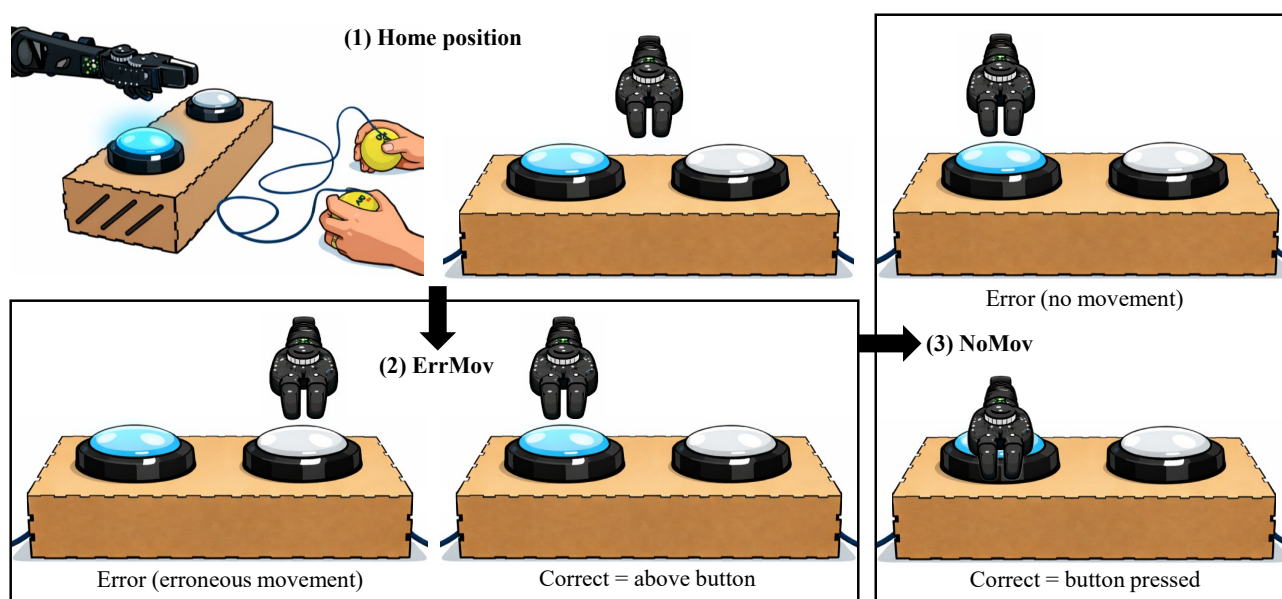


Figure 1: **Experimental setup and paradigm.** (1) The robot rests in the home position and a light instructs the participant to squeeze the left ball. (2) After squeezing (*movement initiated*), the robot moves toward the corresponding button (*above button*). In *error* trials, it instead moves toward the wrong button (**ErrMov**). (3) From the *correct* position (*above button*), the robot either presses the button (*correct*) or omits any movement (*error*; **NoMov**). At the end of the trial, the robot returns to the home position.

tion to compare participant-level ERN and Pe amplitudes. To assess the significance of the classification results, we computed significance levels (SLs) using a shuffling approach. Specifically, we repeated the same 10 times 5-fold CV procedure described in *Classification* 100 times while randomly permuting the class labels in each iteration. The individual SLs were defined as the 95th percentile of the resulting test BA distributions. Additionally, we compared the BAs of **ErrMov** and **NoMov** using a two-sided Wilcoxon signed-rank test ($\alpha = 0.05$).

RESULTS

EEG results: The grand average ERPs ($N = 12$) for both error conditions are shown in Fig. 2. The ErrP for **ErrMov** exhibits the strongest response in Cz and features an initial negative peak 0.45 s ($-4.5 \mu\text{V}$) after *movement initiated*, followed by a positive peak at 0.64 s ($4.5 \mu\text{V}$). Significant differences between erroneous and correct robot movements were observed in the respective segments [0.41, 0.47] s and [0.61, 0.72] s (see Fig. 2a). Similarly, the ErrP to **NoMov** errors comprised negative and a positive components peaking at 0.52 s ($-4.2 \mu\text{V}$) and 0.6 s ($2.8 \mu\text{V}$) after *above button*, with significant activity in the segments [0.5, 0.53] s and [0.59, 0.65] s (all $p < 0.05$). The topographical distributions of the ERPs reveal error-related activity in frontocentral and centroparietal areas of the brain.

Moreover, significant difference arose in participant-level Pe amplitudes in central channels relevant for error processing, i.e., FCz ($p = 0.003$), Cz ($p = 0.005$), and CPz ($p = 0.003$) (see Fig. 3). ERN amplitudes did not significantly differ between error types (FCz: $p = 0.08$; Cz: $p = 0.16$; CPz: $p = 0.2$).

Classification results: The single-trial classification results are summarized in Tab. 1. All participants achieved above-chance decoding performance, irrespective of the error type. Specifically, the grand average BAs ($N = 12$) of 80.6 % in **ErrMov** (TNR = 90.4 %, TPR = 70.8 %) and 79.8 % in **NoMov** (TNR = 89.6 %, TPR = 70.1 %) exceeded the respective SLs (approximately 56%) considerably. No significant BA differences were observed between the two error types ($p = 0.62$).

DISCUSSION

The present work investigated passive BCIs for human-robot interaction. In a game-inspired paradigm, participants initiated a sequence of goal-directed robot movements (i.e., to whack a mole) while experiencing abnormal system behavior typical of real-world robotic applications. Extending prior work on unexpected robot actions [12, 15], we studied neurophysiological responses to erroneous (**ErrMov**) and omitted (**NoMov**) movements, including their use in BCIs.

The responses to both **ErrMov** and **NoMov** comprise a uniform sequence of ERP components in frontocentral and centroparietal brain areas (i.e., ERN and Pe), characteristic of error processing [11]. The latencies observed for both error types were slightly longer than typically reported in related studies (i.e., 200 to 300 ms) [9]. This discrepancy likely resulted from a combination of factors, including hardware limitations, such as delays between user commands and robot activity. Further, contrary to most studies, participants were not confronted with distinct error events (**ErrMov**) but rather with a gradually solidifying error perception, arguably contributing to increased ERP latencies. Indeed, similar findings have been

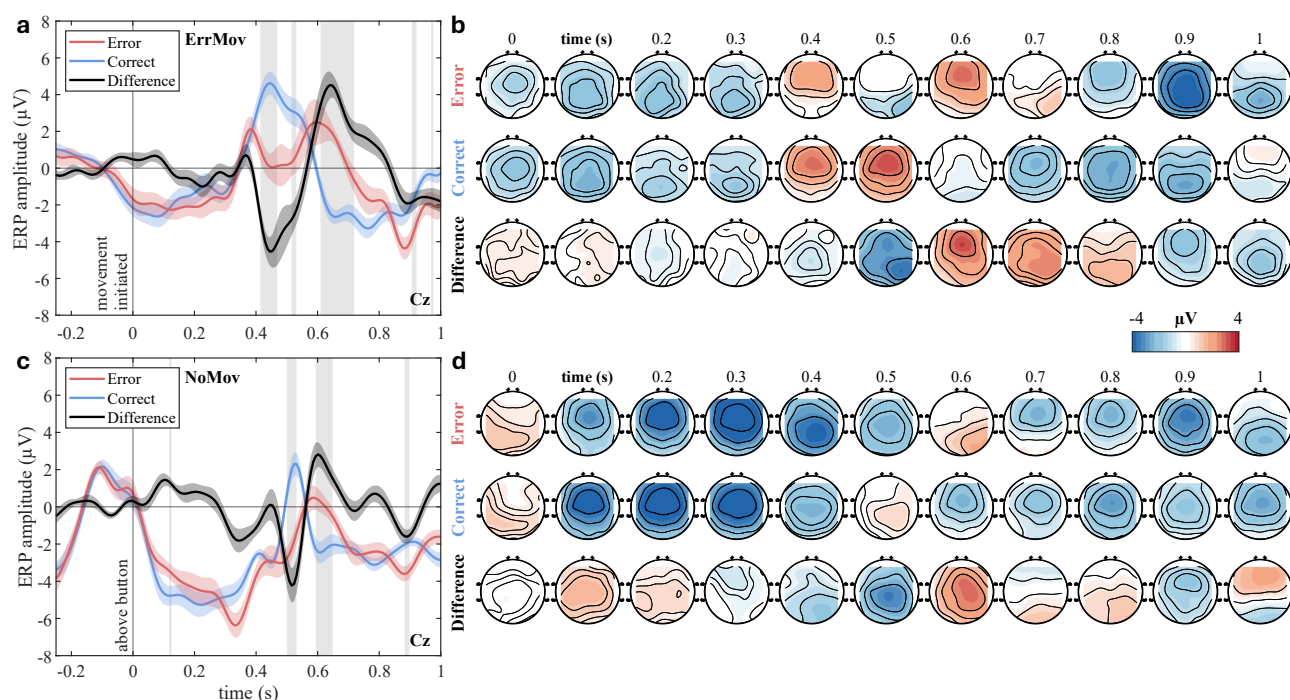


Figure 2: **Grand average ERP results ($N = 12$).** (a) ERPs at Cz for *error* (red), *correct* (blue), and their difference *error minus correct* (black) for **ErrMov**. Shaded areas indicate $\pm$ SE. Gray boxes show significant differences between *error* and *correct* ($p < 0.05$, two-sided Wilcoxon signed-rank, FDR-corrected). Timings are relative to *movement initiated* at time = 0 s. (b) Topographical distribution of the ERPs at the given time points for *error* (top row), *correct* (middle), and *difference* (bottom). (c, d) as in (a, b) for **NoMov**. Timings are relative to *above button* at time = 0 s.

Table 1: **Classification results.** Participant-level results for **ErrMov** and **NoMov**, including TNR, TPR, BA, the significance level (SL), and the corresponding grand average metrics ($M \pm SD$).

Participant	ErrMov				NoMov			
	TNR (%)	TPR (%)	BA (%)	SL (%)	TNR (%)	TPR (%)	BA (%)	SL (%)
P01	91.6	77.2	84.4	54.9	85.3	64.0	74.7	55.8
P02	86.0	58.6	72.3	55.5	85.9	48.0	67.0	56.4
P03	94.8	80.0	87.4	55.6	91.6	79.6	85.6	55.0
P04	87.2	54.5	70.8	55.9	87.5	62.1	74.8	54.5
P05	94.7	76.5	85.6	55.4	89.8	70.7	80.2	55.2
P06	86.7	63.7	75.2	55.6	89.0	71.6	80.3	56.0
P07	89.8	75.9	82.8	55.2	94.3	80.8	87.6	55.6
P08	89.7	66.3	78.0	53.8	93.7	84.8	89.2	55.1
P09	93.2	78.4	85.8	56.6	90.4	66.2	78.3	54.8
P10	87.4	70.8	79.1	55.9	85.4	66.0	75.7	56.7
P11	94.1	82.3	88.2	54.7	91.0	71.6	81.3	55.4
P12	89.8	65.3	77.6	55.3	91.2	75.5	83.3	55.4
$M \pm SD$	90.4 ± 3.2	70.8 ± 9.0	80.6 ± 5.9	55.4 ± 0.7	89.6 ± 3.1	70.1 ± 9.9	79.8 ± 6.3	55.5 ± 0.6

reported for errors in continuous cursor movements [33]. Interestingly, we additionally observed a reduction in peak Pe amplitude in **NoMov** compared to **ErrMov**, while the ERN remained unaffected by the error type. This finding is in line with prior literature that has linked Pe to error awareness [11, 34]. For instance, in studies that explicitly asked participants to signal their committed errors, Pe (but not ERN) amplitudes were attenuated after unreported errors [35]. Similarly, Pe has been associated with the accumulation of evidence that an error has occurred [34]. Consequently, our results suggest that participants were uncertain about the system status when

an action was omitted, whereas movements toward the wrong button constituted an obvious malfunction. The absence of instructions regarding the robot's erroneous behavior may have contributed to these results. Additionally, differences in the physiology might be partially attributed to a smearing effect in the grand average ERPs, resulting from the inherent variability in the onset of the **NoMov** error (*When* did participants realize the error, i.e., that the button press would not happen?). Similar effects have been described earlier [36, 37].

The discrepancy in ERP amplitudes did not translate into a significant difference in classification performance. The

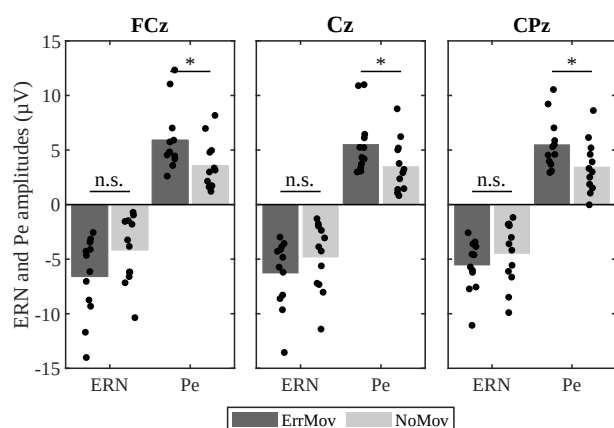


Figure 3: **Comparison of ERN and Pe components.** Dots show the participant-level ERN and Pe amplitudes at FCz (left), Cz (middle), and CPz (right) for **ErrMov** (dark) and **NoMov** (light) and the corresponding group-level averages (bars). Results of the significance tests are indicated (two-sided Wilcoxon signed-rank, FDR-corrected, $\alpha = 0.05$, $*p < 0.01$).

single-trial BAs for both **ErrMov** and **NoMov** scaled at approximately 80 % and were well within the range of typically reported accuracies (i.e., 70 to 80 % for time-locked ErrP classification [9]). This provides evidence that a variety of errors in human-robot interactions, both in unexpected and omitted actions, can be identified through passive BCIs.

CONCLUSION

This work extends the current scope of BCI applications in human-robot interaction by characterizing and decoding EEG responses to erroneously omitted robot activity. We found that both erroneously executed and omitted tasks elicit ErrPs of comparable morphology exhibiting frontocentral and centroparietal activity. Notably, Pe components associated with task omission were significantly attenuated, likely linked to increased uncertainty about the occurrence of an error. Additionally, both error types could be classified on a single-trial level with 80 % BA. This provides evidence that decoding previously underexplored omission errors via EEG can advance autonomously corrective systems in human-robot interaction.

ACKNOWLEDGEMENTS

This work is supported by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (IC170100030).

The Know Center is funded within the Austrian COMET Program - Competence Centers for Excellent technologies - under the auspices of the Austrian Federal Ministry of Innovation, Mobility and Infrastructure, the Federal Ministry of Economy, Energy and Tourism, and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

REFERENCES

- [1] Sheridan TB. Human-robot interaction: Status and challenges. *Human Factors*. 2016;58(4):525–532.
- [2] Han IH et al. Human-robot interaction and social robot: The emerging field of healthcare robotics and current and future perspectives for spinal care. *Neurospine*. 2024;21(3):868–877.
- [3] Farina D et al. Toward higher-performance bionic limbs for wider clinical use. *Nature Biomedical Engineering*. 2023;7(4):473–485.
- [4] Rodríguez-Guerra D, Sorrosal G, Cabanes I, Calleja C. Human-robot interaction review: Challenges and solutions for modern industrial environments. *IEEE Access*. 2021;9:108557–108578.
- [5] Slade P et al. On human-in-the-loop optimization of human-robot interaction. *Nature*. 2024;633(8031):779–788.
- [6] Kokorin K, Zehra SR, Mu J, Yoo P, Grayden DB, John SE. Semi-autonomous continuous robotic arm control using an augmented reality brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024;32:4098–4108.
- [7] Zander TO, Kothe C. Towards passive brain-computer interfaces: Applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering*. 2011;8(2):025005.
- [8] Zhang Y et al. Mind meets robots: A review of EEG-based brain-robot interaction systems. *International Journal of Human-Computer Interaction*. 2025;41(20):12784–12815.
- [9] Chavarriaga R, Sobolewski A, Millán JdR. Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*. 2014;8.
- [10] Schalk G, Wolpaw JR, McFarland DJ, Pfurtscheller G. EEG-based communication: Presence of an error potential. *Clinical Neurophysiology*. 2000;111(12):2138–2144.
- [11] Gehring WJ, Liu Y, Orr JM, Carp J. The error-related negativity (ERN/Ne). *The Oxford handbook of event-related potential components*. 2012;641:231–291.
- [12] Salazar-Gomez A, DelPreto J, Gil S, Guenther F, Rus D. Correcting robot mistakes in real time using EEG signals. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, 6570–6577.
- [13] Akinola I et al. Accelerated robot learning via human brain signals. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020, 3799–3805.
- [14] Iturrate I, Omedes J, Montesano L. Shared control of a robot using EEG-based feedback signals. In: *Proceedings of the 2nd Workshop on Machine Learning for Interactive Systems (MLIS)*. 2013, 45–50.
- [15] Ehrlich SK, Cheng G. A feasibility study for validating robot actions using EEG-based error-related potentials. *International Journal of Social Robotics*. 2019;11(2):271–283.

- [16] Ferrez PW, Millán JdR. Error-related EEG potentials generated during simulated brain–computer interaction. *IEEE Transactions on Biomedical Engineering*. 2008;55(3):923–929.
- [17] Lopes-Dias C, Sburlea AI, Müller-Putz GR. Online asynchronous decoding of error-related potentials during the continuous control of a robot. *Scientific Reports*. 2019;9(1):17596.
- [18] Pulferer HS, Kostoglou K, Müller-Putz GR. Getting off track: Cortical feedback processing network modulated by continuous error signal during target-feedback mismatch. *NeuroImage*. 2023;274:120144.
- [19] Wimmer M, Weidinger N, Veas E, Müller-Putz GR. Neural and pupillometric correlates of error perception in an immersive VR flight simulation. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2023, 1–4.
- [20] Alain C, Richer F, Achim A, Saint Hilaire JM. Human intracerebral potentials associated with target, novel, and omitted auditory stimuli. *Brain Topography*. 1989;1(4):237–245.
- [21] Perri R, Spinelli D, Di Russo F. Missing the target: The neural processing underlying the omission error. *Brain Topography*. 2017;30.
- [22] Dal Seno B, Matteucci M, Mainardi L. Online detection of P300 and error potentials in a BCI speller. *Computational Intelligence and Neuroscience*. 2010;(1):307254.
- [23] Müller-Putz GR et al. Feel your reach: An EEG-based framework to continuously detect goal-directed movements and error processing to gate kinesthetic feedback informed artificial arm control. *Frontiers in Human Neuroscience*. 2022;110.
- [24] Mondini V, Sburlea AI, Müller-Putz GR. Towards unlocking motor control in spinal cord injured by applying an online EEG-based framework to decode motor intention, trajectory and error processing. *Scientific Reports*. 2024;14(1):4714.
- [25] Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*. 2004;134(1):9–21.
- [26] Gramfort A et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroinformatics*. 2013;7:267.
- [27] Delorme A, Sejnowski T, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*. 2007;34(4):1443–1449.
- [28] Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*. 1999;10(3):626–634.
- [29] Yasemin M, Cruz A, Nunes UJ, Pires G. Single trial detection of error-related potentials in brain–machine interfaces: A survey and comparison of methods. *Journal of Neural Engineering*. 2023;20(1):016015.
- [30] Blankertz B, Lemm S, Treder M, Haufe S, Müller KR. Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*. 2011;56(2):814–825.
- [31] Hancock J, Johnson JM, Khoshgoftaar TM. A comparative approach to threshold optimization for classifying imbalanced data. In: 2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC). 2022, 135–142.
- [32] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995;57(1):289–300.
- [33] Omedes J, Iturrate I, Minguez J, Montesano L. Analysis and asynchronous detection of gradually unfolding errors during monitoring tasks. *Journal of Neural Engineering*. 2015;12(5):056001.
- [34] Steinhäuser M, Yeung N. Error awareness as evidence accumulation: Effects of speed-accuracy trade-off on error signaling. *Frontiers in Human Neuroscience*. 2012;6.
- [35] Nieuwenhuis S, Ridderinkhof KR, Blom J, Band GP, Kok A. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*. 2001;38(5):752–760.
- [36] Lopes-Dias C, Sburlea AI, Müller-Putz GR. Masked and unmasked error-related potentials during continuous control and feedback. *Journal of Neural Engineering*. 2018;15(3):036031.
- [37] Wimmer M, Weidinger N, Veas E, Müller-Putz GR. Multimodal decoding of error processing in a virtual reality flight simulation. *Scientific Reports*. 2024;14(1):1–14.

Investigating self-introspection in an auditory ERP-based BCI: A pilot experiment

Man Yang¹, Jordy Thielen¹, Michael Tangermann¹

¹Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

E-mail: man.yang@donders.ru.nl

ABSTRACT: The decoding performance of a brain–computer interface (BCI) depends on both, the quality of the classification model and the signal-to-noise ratio of the brain signals. In this pilot study, we introduce *self-introspection* as a subjective measure of a user’s awareness of their own BCI performance, i.e., their ability to anticipate how well the BCI will work. We investigated this concept in an auditory event-related potential (ERP)-based BCI paradigm across six experimental sessions of a single participant. After each trial, the participant rated the expected BCI feedback using a continuous mechanical slider, before receiving the true BCI feedback. This enabled a direct comparison between subjective self-introspection scores and BCI system scores. Results indicate a positive relationship between these scores across sessions, suggesting that the user had an internal sense of his BCI performance. As the present study included a single participant, the findings should be interpreted as a feasibility demonstration. Nevertheless, they show that subjective system performance can be estimated within an auditory ERP BCI paradigm, providing a potential method for studying user learning and adaptation in BCI use.

INTRODUCTION

Electroencephalography (EEG) is widely used in brain–computer interface (BCI) research due to its non-invasiveness, high temporal resolution, and relatively low cost. EEG-based BCIs translate users’ neural activity into signals that can be decoded by computational models and used to provide feedback.

Among different BCI modalities, auditory BCIs provide an important alternative when visual attention is limited or unavailable, especially in assistive and rehabilitation contexts. For example, auditory BCIs have been explored as communication tools for individuals with severe motor impairments such as amyotrophic lateral sclerosis (ALS), where visual BCIs may be difficult to use due to impaired eye movement or visual function [1]. In addition, auditory paradigms have been investigated for language rehabilitation in patients with aphasia [2]. These applications highlight the potential of auditory BCI paradigms as assistive and rehabilitative technologies.

Many auditory BCIs rely on variants of the oddball paradigm to elicit event-related potentials (ERPs) like

the P300 response. Previous studies demonstrated that spatially distributed sound sources can reliably evoke such responses. A prominent example is the AMUSE paradigm, in which stimuli are presented from multiple loudspeakers at different spatial locations around the participant [3]. This AMUSE paradigm has been used for practical applications, including auditory spelling systems [4] and BCI-supported language rehabilitation for patients with stroke-induced aphasia [2].

Despite these advances, BCI performance often shows variability within and across sessions. Factors influencing performance are often categorized into system-related and user-related factors [5, 6]. While system-related improvements such as signal processing and machine learning have been widely studied, user-related factors remain comparatively underexplored. In particular, it remains unclear how users internally evaluate their performance during a BCI task. One possible aspect of this internal evaluation is the user’s ability to estimate the outcome of a BCI task. In this study, we refer to this ability as *self-introspection*. Understanding this ability may provide insight into user–system interaction and may offer a complementary readout of how well users monitor and adapt their mental strategy during BCI use.

Therefore, the present study investigates self-introspection in an auditory ERP-BCI paradigm adapted from Kojima et al [7]. A single participant performed a language-related auditory BCI task while the BCI decoded the recorded EEG. After each trial, the participant first rated their perceived performance before receiving the BCI feedback. By examining the relationship between subjective self-introspection and system performance, this study aims to investigate whether users can reliably estimate their BCI performance in the absence of explicit feedback and whether this measure is feasible as a future tool for studying learning and adaptation in BCI use.

MATERIALS AND METHODS

Participant: One participant (male, 23 years old) was included in this pilot study. The participant was familiar with general EEG and BCI knowledge, but had no prior knowledge of the specific experimental design used in this study. Written informed consent was obtained before participation. Ethical approval for the experimental

protocol was granted by the local ethics committee of the Faculty of Social Sciences, Radboud University, and all procedures adhered to the approved guidelines. The presented data from a single participant represents an initial pilot stage of a larger study on self-introspection in the context of an auditory BCI.

Materials: EEG data were collected with 32 active electrodes positioned according to the international 10-10 system and sampled at 500 Hz. Data were amplified using a BrainAmp amplifier (Brain Products GmbH). Data acquisition was handled by the BrainAmpSeries Connector application (version 1.18.0), which transmitted the EEG signals via Lab Streaming Layer (LSL) [8], and all data were recorded using LabRecorder (version 1.16.5).

BCI Paradigm and Task: We employed an auditory ERP-based BCI paradigm with six Dutch words as stimuli. Each word was associated with a corresponding Dutch sentence that served as a cue. All audio files were presented at a sampling rate of 44.1 kHz, and exclusively via headphones (Sony MDR-7506).

To simulate spatially separated sound sources, virtual auditory positions were generated using Reaper (version 7.36) in combination with the dearVR MICRO plugin (version 1.0.3). Six fixed virtual positions were defined in the horizontal plane around the participant, covering the full 360° with equal angular spacing (see Figure 1(a)). In the spatial mode, the six words were presented from the six virtual positions. In the mono mode, the six words were presented without spatial separation.

The experiment consisted of a familiarization phase prior to the experimental sessions which included offline and online tasks. The familiarization phase was conducted during the first session, and EEG was not recorded during this phase. It included three tasks.

First, in the position familiarization task, the participant was introduced to the six virtual auditory source positions. The purpose of this task was to help the participant get used to the simulated spatial locations. The participant pressed keys 1 to 6 to play sounds from the corresponding virtual positions. This allowed him to actively explore and perceive the spatial separation between the different sound sources.

Second, in the sentence completion task, the participant learned the mapping between each sentence cue and its corresponding target word. After a sentence cue was presented, the participant was asked to select the correct target word from the six-word set. This ensured that the participant had memorized the cue–target associations before proceeding to the formal experimental sessions. The sentence-word pairs used are listed in Table 1.

Third, in the position identification task, the participant indicated the perceived spatial position of the target word after hearing a sentence cue and a six-word sequence. The purpose was to verify whether the participant had established a reliable association between sentence cues, target words, and their corresponding virtual positions.

After completing the familiarization phase, the participant proceeded to the formal offline and online tasks. The

offline and online tasks were organized into runs, where each run consisted of six trials. In each trial, the participant attended one of the six target words which had been cued by the sentence at the trial start.

During the **offline task**, the participant passively performed the auditory task. The EEG data collected during the offline runs were used to train a block Toeplitz linear discriminant analysis (btLDA) classifier [9], which was used in later online runs to decode the ERP responses and provide feedback.

At the beginning of each trial, the sentence cue was presented. This was followed by a silent interval, with the combined duration of the sentence and silence fixed to 6 s. Then, a fixation cross was displayed on the screen before the stimulation sequence started 2 s later.

The sequence consisted of 15 iterations of the six-word set, resulting in a total of 90 word presentations per trial. Within each iteration, the order of the six words was pseudo-randomized. To avoid temporal predictability, successive words were separated by at least three in-between words. In the spatial condition, each word was assigned to one spatial position within a trial. This relation was changed pseudo-randomly for the next trial.

After the fixation cross disappeared, a 3 s inter-trial pause was presented before the next trial began. The participant completed several offline runs across different sessions. Both the stimulus onset asynchrony (SOA) and the audio presentation mode (spatial or mono) differed between sessions. (see Figure 2).

In the **online task**, each trial included the sentence cue and an auditory stimulus presentation sequence as in the offline task, but in addition also a self-introspection rating phase, and a subsequent feedback display (see Figure 1(b)). Specifically, the fixation cross disappeared 2 s after the onset of the last word stimulus had been played, and a 5 s countdown started. During this period, the participant used a continuous mechanical slider to rate his perceived performance for that trial, the self-introspection score. The subsequent feedback phase lasted 3 s, during which both the self-introspection score and the true system score were displayed on the screen. The next trial started after the feedback phase.

Across the online sessions, task difficulty was manipulated by varying the SOA and the auditory presentation mode. Three SOA levels were tested: 350 ms, 250 ms, and 200 ms. The average duration of individual word stimuli was approximately 280 ms. For SOAs shorter than the word duration, successive auditory stimuli overlapped in time, which increased attentional demands.

To ensure consistency within each session, the six-word set remained unchanged across trials. Between sessions, predefined replacement words could be introduced if noticeable performance imbalances between target words had been observed in the previous session. This procedure prevented persistent ceiling effects while maintaining stable stimulus conditions within a session.

Session Design: The experiment comprised six sessions conducted on different days. The exact number of

Table 1: **Dutch sentence–word pairs.** During a trial, the participant attended one target word, which was associated with and cued by a specific sentence. All stimuli were presented in Dutch; English translations are provided for reference only.

Sentence (Dutch)	Sentence (English)	Target (Dutch)	Target (English)
Ik mix een milkshake in de ____	I mix a milkshake in the ____	Blender	Blender
We eten aardappelen, vlees en ____	We eat potatoes, meat, and ____	Groente	Vegetable
Op de kermis vind ik een grote ____	At the fair, I find a large ____	Knuffel	Stuffed toy
Ik zie mijn reflectie in de ____	I see my reflection in the ____	Spiegel	Mirror
Om zijn nek draagt de zakenman een ____	Around his neck, the businessman wears a ____	Stropdas	Tie
Het veld wordt omgeploegd met een ____	The field is plowed with a ____	Tractor	Tractor

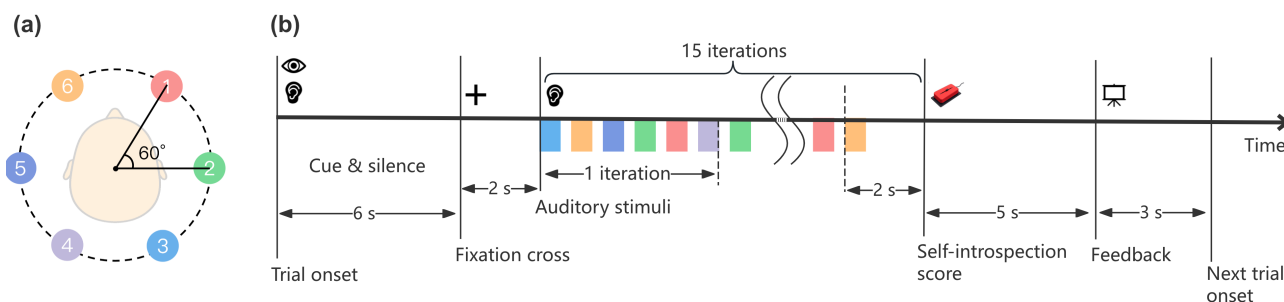


Figure 1: **Experimental setup and online trial timing.** (a) Arrangement of the six virtual auditory speaker locations, equally distributed around the participant in the horizontal plane with an angular separation of 60°. (b) Schematic illustration of one online trial. The structure was identical to the offline trial up to stimulus presentation, followed by a self-introspection rating and a feedback display.

offline and online runs in each session is shown in Figure 2. Rather than predefining a fixed number of runs, the aim was to collect as many trial-wise scores as possible within each session while keeping the total session duration feasible for the participant. Each session comprised offline and online tasks, except for session 5, which included online tasks only. In sessions 1–4 and session 6, EEG data were first collected during offline tasks for classifier training. The trained classifier was subsequently used in the online tasks of the same session. Session 2 contained two online phases. After the initial spatial online runs, the presentation mode was switched to mono. The classifier trained on the spatial data was directly reused for the mono runs, allowing us to assess within-session transfer without additional recalibration. Session 5 consisted of online runs only, for which the classifier trained in the previous session was reused.

Across sessions, task difficulty was manipulated by varying the SOA and the auditory presentation mode. Earlier sessions used longer SOAs and spatial separation, whereas later sessions introduced shorter SOAs and mono presentation to increase task demands. These changes were intentionally exploratory in this pilot study. They served to identify a parameter range that remained challenging while still allowing stable task performance.

The stimulus set remained constant within each session. Sessions 1–3 used the default word set (Blender, Groente, Knuffel, Spiegel, Stropdas, Tractor). In Sessions 4–6, the word “Groente” was replaced by “Stoeprand” to balance performance across target words.

Self-Introspection Score: The self-introspection score was obtained directly from the participant using a continuous mechanical slider interface described above. After each trial, the participant rated how they expected the

BCI system to evaluate their performance of that trial. The slider value was recorded within the range of $[-1, 1]$, enabling direct comparison with the system score.

The participant was instructed that a value of -1 , corresponding to the leftmost slider position, indicated “I think the system will give a very negative score”. The value 0 , positioned in the middle, indicated “I think the system will give a neutral score” and 1 , corresponding to the rightmost slider position, indicated “I think the system will give a very positive score”. Based on their subjective judgment of the current trial, the participant then used the slider to provide their introspection score prior to receiving the system score as feedback.

System Score: For each trial, EEG responses were processed using a btLDA classifier. Each trial consisted of 15 iterations of a six-word stimulus set. For every stimulus, the classifier produced an output, which was transformed into a probability using a sigmoid function.

For each of the six words, probabilities were averaged across the 15 iterations present in a trial. The true *target word* for that trial was defined by the cueing sentence, with each sentence corresponding to exactly one word (see Table 1), irrespective of whether it obtained the highest mean classifier score. Among the remaining five words, the *best non-target* was defined as the non-target word with the highest mean probability. This comparison between true *target* and *best non-target* focuses on the most difficult classification case, as the best non-target represents the strongest competitor to the target.

To quantify how well the system distinguished the target from competing stimuli, the area under the receiver operating characteristic curve (AUC) was computed using the 15 probabilities of the target word and the 15 probabilities of the best non-target word.

	Blender, Groente, Knuffel, Spiegel, Stropdas, Tractor												
session ID	1		2			3		4		5		6	
number of runs	4	6	4	5	3	4	11	4	12	8	4	6	
spatial/mono	▲	▲	▲	▲	●	●	●	●	●	●	●	●	
SOA	350	350	250	250	250	200	200	200	200	200	250	250	
offline/online	off	on	off	on	on	off	on	off	on	on	off	on	
	Blender, Stoeprand, Knuffel, Spiegel, Stropdas, Tractor												

Figure 2: Overview of the experimental protocol. The experiment consisted of six sessions with different presentation modes (spatial or mono), numbers of runs, and SOAs. Triangles (▲) indicate spatial presentation and circles (●) indicate mono presentation. Labels “off” denote offline runs used for classifier training, whereas “on” denote online runs. Sessions 1–2 used SOAs of 350 ms and 250 ms, sessions 3–5 used an SOA of 200 ms, and session 6 used an SOA of 250 ms. The first three sessions used the word set *Blender, Groente, Knuffel, Spiegel, Stropdas, Tractor*, while in the last three sessions *Groente* was replaced by *Stoeprand*.

The resulting AUC value was linearly rescaled to the range $[-1, 1]$. That is, for trial t :

$$\text{SystemScore}_t = 2 \times \text{AUC}_t - 1 \quad (1)$$

With this transformation, -1 represents reversed discrimination, 0 corresponds to chance level, and 1 indicates perfect separation. The relationship between the self-introspection score and the system score is summarized in Figure 3.

EEG Data Processing and Feature Extraction: EEG data were recorded and stored in XDF format. Offline data were used for classifier training, and online data were processed in real time using the trained classifier.

A subset of 32 EEG channels was used for analysis (Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10). The continuous EEG data were first epoched from -0.65 s to 1.35 s relative to stimulus onset. These extended epochs were band-pass filtered between 0.5 and 20 Hz, downsampled to 100 Hz to reduce computational load. Finally, the epochs were cropped to the analysis window from -0.35 s to 1.05 s relative to the stimulus onset.

For feature extraction, the post-stimulus interval was divided into consecutive time windows of 75 ms duration, resulting in 14 temporal segments covering the interval from 0 to 1.05 s. Within each window, EEG samples were averaged for each channel. The resulting values across all channels and time windows were concatenated to form the feature vector for classification with btLDA.

Classifier Updating Strategy: For sessions that included an offline phase (Sessions 1–4 and Session 6), an initial btLDA classifier was trained using the offline data collected at the beginning of the session. In Session 5, no offline data were collected. Instead, the classifier trained at the end of Session 4 was reused as the initial model. To enable adaptation during online operation, a sliding-window update strategy was applied. The classifier was

continuously trained on the most recent six runs ($3,240$ trials) using a first-in-first-out (FIFO) scheme. Each trial was decoded before being added to the training window, ensuring that no trial contributed to its own prediction. At the start of the online task, newly acquired trials were added until six runs were available; thereafter, the window was updated after each trial by removing the oldest and adding the newest. In session 5, initial trials were decoded using the classifier from session 4, and subsequent trials were decoded using models that had been updated with previously completed trials from session 5.

RESULTS

Trial-wise Relationship Between Self-Introspection and System Scores: We examined trial-wise correlations separately within each session (see Figure 4). Session 1 showed a significant positive correlation ($r = .536$, $p < .001$), and session 3 showed a similar significant positive correlation ($r = .483$, $p < .001$). In session 2, correlations were computed separately for spatial and mono trials: both spatial trials ($r = .501$, $p = .005$), and mono trials ($r = .618$, $p = .006$) showed a significant positive correlation. In session 4, the positive association remained but was weaker ($r = .244$, $p = .039$). In session 5, no significant correlation was observed ($r = -.075$, $p = .615$ (n.s.)). In session 6, a significant positive correlation re-emerged ($r = .377$, $p = .023$). Overall, these results show that the strength of the relationship varied across sessions, likely influenced by changes in task configuration.

When pooling trials from all sessions, the correlation remained significantly positive ($r = 0.481$, $p < 0.001$), indicating that the participant’s self-introspection scores generally co-varied with the system scores.

Word-Level Classification Performance and Stimulus Adjustment: To examine class separability across target words, confusion matrices were computed separately for sessions 1–3 and sessions 4–6 (see Figure 5).

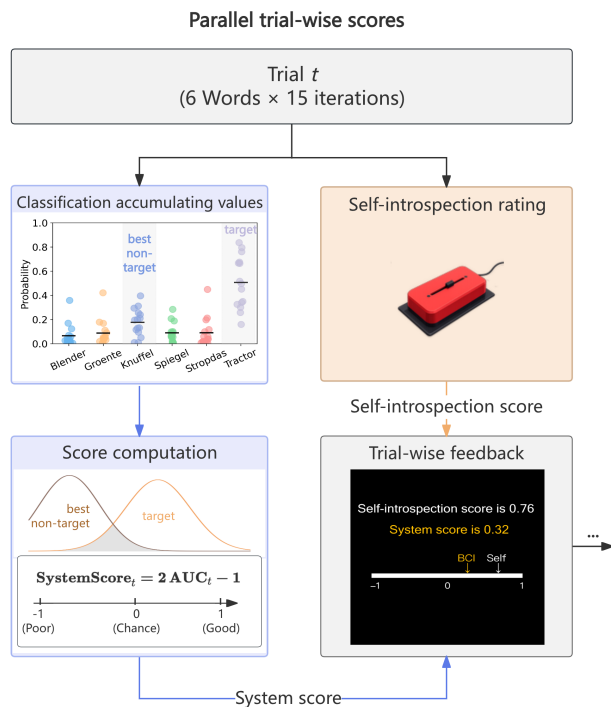


Figure 3: **Trial-wise self-introspection and system scores.** For each trial, classifier outputs were used to compute an AUC value, which was linearly rescaled into the system score. Participants also provided a self-introspection score after stimulus presentation. Both scores were presented as trial-wise feedback.

Figure 5(a) shows the pooled confusion matrix for sessions 1–3. In these sessions, the word *Groente* showed noticeably lower classification performance compared to the other words, as reflected by reduced diagonal counts and increased confusion with competing words. This imbalance was consistent with the participant’s subjective report indicating that this word was more difficult to attend to. Based on both classifier performance and participant feedback, *Groente* was replaced by *Stoeprand* in sessions 4–6. Figure 5(b) presents the pooled confusion matrix for sessions 4–6 after the replacement. Following this adjustment, classification performance across words became more balanced.

DISCUSSION

This pilot study investigated whether a single participant could estimate his own performance in an auditory BCI task, a process we refer to as self-introspection. Across six sessions, we compared subjective self-introspection scores with system scores obtained from a btLDA classifier. The multi-session design allowed us to assess whether this relationship remained consistent across sessions and under different task difficulty levels. The main objective was to evaluate the feasibility of measuring subjective performance awareness within a multi-session auditory BCI paradigm.

Overall, the results suggest that self-introspection can be meaningfully assessed in this paradigm. Across sessions, we observed a consistent positive relationship between

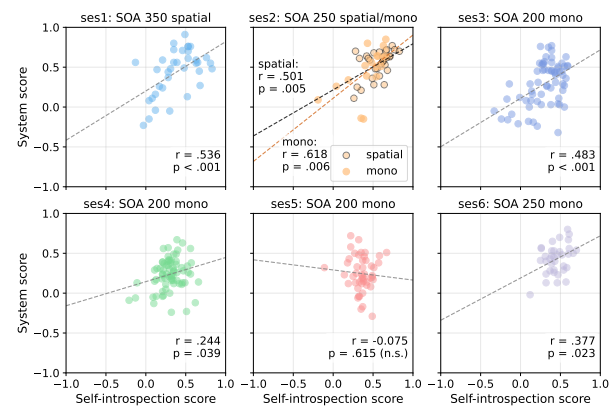


Figure 4: **Session-specific trial-wise correlations between self-introspection scores and system scores.** Each panel represents one session. In session 2, spatial and mono trials are analyzed separately and shown with separate regression lines; spatial data points are marked with black outlines. Correlation coefficients and corresponding significance levels are indicated within each subplot.

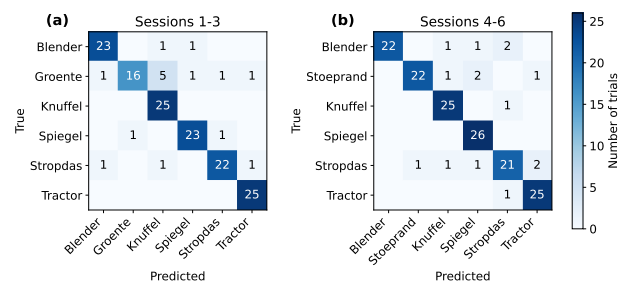


Figure 5: **Confusion matrices pooled across sessions.** (a) Sessions 1–3 using the original word set including *Groente*. (b) Sessions 4–6 after replacing *Groente* with *Stoeprand*. Color intensity reflects the number of trials.

the self-introspection scores provided by the participant and the system scores. Although the strength of this relationship varied across sessions, the overall pooled correlation indicates that the participant was generally able to anticipate how the BCI system would evaluate their performance. These findings provide initial support for the idea that users may possess a degree of internal awareness of their BCI performance.

At the same time, the present study involved only a single participant, and therefore the observed relationship cannot yet be generalized. Consequently, the present findings should be interpreted primarily as evidence that the paradigm is technically and experimentally feasible. Future studies involving a larger participant group will be necessary to determine whether the observed relationship generalizes across individuals.

The pilot phase revealed several aspects that can be improved in future studies. In the present study, SOA and presentation mode (spatial or mono) were varied within sessions to explore task feasibility under different difficulty levels. However, for studies targeting learning effects, a stable task configuration is preferable, as frequent changes may introduce variability that interferes

with learning. Future experiments should therefore keep parameters constant within and ideally across sessions. A key challenge is to select an appropriate initial difficulty level that prevents ceiling effects while still allowing performance improvements over time.

Another consideration concerns the interpretation of the system score derived from the AUC metric. Although it provides a continuous measure of separability, its meaning at the single-trial level is not always straightforward. In Figure 6, each point represents one trial, with the horizontal axis showing the scaled system score. Circles indicate trials with correct decoding (target word has the highest mean classifier score), whereas crosses indicate incorrect decoding.

We observed that some trials were decoded correctly despite AUC values below 0.5 (scaled < 0), while others were incorrect despite AUC values above 0.5 (scaled > 0). This apparent inconsistency arises because this AUC-like metric quantifies the global separability between the target and the strongest competing non-target across repetitions within a trial. It does not directly reflect whether the target achieved the highest *mean* score among all six words in that trial. As a result, the binary decoding outcome and the continuous AUC-based system score are not necessarily aligned. A discrepancy becomes problematic near the decision boundary.

This mismatch may influence how participants interpret the feedback. In particular, if the feedback does not consistently reflect the decoding outcome, it may affect both the participant's ability to anticipate the system's evaluation and the learning process during repeated BCI use.

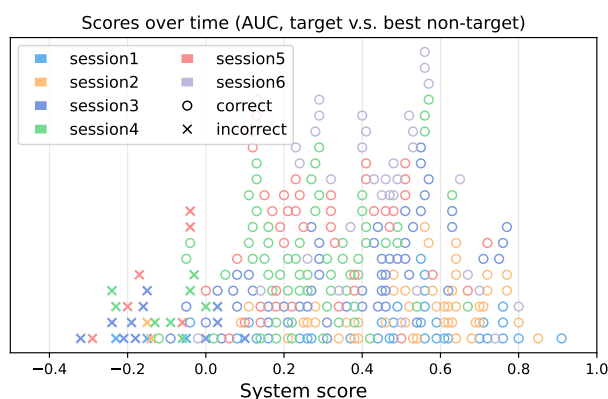


Figure 6: Distribution of trial-wise system scores across sessions. Each point represents one trial. The horizontal axis shows the scaled system score derived from the AUC ($2 \times \text{AUC} - 1$). Circles indicate correctly decoded trials, whereas crosses indicate incorrectly classified trials. Colors correspond to different sessions.

CONCLUSION

The present pilot study demonstrates the feasibility of measuring self-introspection in an auditory BCI task and suggests that a user may possess an internal sense of their BCI performance. While the results are preliminary

due to the single-participant design, the study provides a proof of concept for integrating subjective performance awareness into auditory neurofeedback BCI paradigms. Future work with more participants and improved experimental parameters are required to determine whether self-introspection can reliably predict BCI performance and potentially support adaptive training strategies.

ACKNOWLEDGMENTS

This project has received funding from the European Union's research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101118964, and from the Dutch Research Council (grant 024.005.022).

REFERENCES

- [1] P. Séguin et al., "The challenge of controlling an auditory BCI in the case of severe motor disability," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, p. 9, Jan. 2024.
- [2] M. Musso et al., "Aphasia recovery by language training using a brain-computer interface: A proof-of-concept study," *Brain Communications*, vol. 4, no. 1, fcac008, Jan. 2022.
- [3] M. Schreuder, B. Blankertz, and M. Tangermann, "A New Auditory Multi-Class Brain-Computer Interface Paradigm: Spatial Hearing as an Informative Cue," *PLOS ONE*, vol. 5, no. 4, e9813, Apr. 2010.
- [4] M. Schreuder, T. Rost, and M. Tangermann, "Listen, You are Writing! Speeding up Online Spelling with a Dynamic Auditory BCI," *Frontiers in Neuroscience*, vol. 5, Oct. 14, 2011.
- [5] C. Jeunet, B. N'Kaoua, and F. Lotte, "Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates," *Progress in Brain Research*, vol. 228, pp. 3–35, 2016.
- [6] F. Lotte et al., "A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031 005, Jun. 2018.
- [7] S. Kojima et al., "Influence of pitch modulation on event-related potentials elicited by Dutch word stimuli in a brain-computer interface language rehabilitation task," *Journal of Neural Engineering*, vol. 22, no. 3, p. 036 010, Jun. 1, 2025.
- [8] C. Kothe et al., "The lab streaming layer for synchronized multimodal recording," *Imaging Neuroscience*, vol. 3, IMAG.a.136, 2025, Open Access.
- [9] J. Sosulski and M. Tangermann, "Introducing block-Toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain-computer interfaces," *Journal of Neural Engineering*, vol. 19, no. 6, Nov. 7, 2022.

EEG MARKERS OF COGNITIVE IMPAIRMENT IN CEREBRAL SMALL VESSEL DISEASE AND NEUROFEEDBACK: A NARRATIVE REVIEW

M. Bourdil^{1,2}, H. Sauzéon^{1,2}, F. Lotte¹

¹Inria Center at the University of Bordeaux, F-33400 Talence, France

²Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France

E-mail: manon.bourdil@inria.fr

ABSTRACT: Cerebral small vessel disease (CSVD) is a major cause of vascular cognitive impairment (VCI) and vascular dementia (VaD), yet early detection and monitoring remain challenging. Electroencephalography (EEG) may offer a non-invasive and cost-effective approach to identify functional brain alterations associated with CSVD. This narrative review summarizes current evidence on EEG markers in CSVD patients with mild or major cognitive impairment. Across studies, CSVD was consistently associated with spectral slowing (increased delta and theta power, reduced alpha activity), disrupted functional connectivity (particularly in fronto-parietal networks) and altered event-related potentials. Several EEG measures correlated with cognitive performance and lesion burden, suggesting potential value for disease monitoring. However, heterogeneity in diagnostic criteria and methodological approaches limits comparability across studies. Overall, EEG biomarkers appear promising for early detection and monitoring of CSVD and may provide targets for future neurofeedback interventions aimed at mitigating cognitive decline, which we also discuss.

INTRODUCTION

The term cerebral small vessel disease (CSVD) encompasses a heterogeneous group of pathological processes affecting the small vessels of the brain, including small arteries, capillaries, arterioles, and venules [1]. These pathologies can lead to both ischemic and hemorrhagic consequences, resulting in various parenchymal lesions. The progression of CSVD is highly variable, with many patients remaining asymptomatic for years (covert CSVD). However, the cumulative burden of lesions can eventually lead to cognitive, neuropsychiatric, and mobility disorders [2]. The prevalence of the condition increases with aging and is very common in older individuals [3], making CSVD a growing public health challenge.

Vascular cognitive impairment (VCI) refers to all forms of cognitive impairment due to cerebrovascular disease. According to the Vascular Impairment of Cognition Classification Consensus Study (VICCCS) [4], the classification of VCI severity ranges from mild VCI to vascular dementia (VaD). CSVD is the primary cause of VCI and subcortical ischemic vascular dementia

(SiVaD) [5], one of the four subtypes of VaD with post-stroke dementia, multi-infarct dementia, and mixed dementias. SiVaD is characterized by executive dysfunction, impairing information processing and problem resolution, caused by the disruption of prefrontal-subcortical connections [6].

The lesions related to CSVD are typically detected via magnetic resonance imaging (MRI) or computed tomography (CT) [7]. However, CSVD lesions can induce remote effects which, associated with patient-specific factors (e.g., age, genetic background, cognitive reserve), contribute to the heterogeneous clinical presentation [8]. Electroencephalography (EEG) is a promising alternative as it is non-invasive, relatively inexpensive, portable, and easily repeatable. Therefore, EEG could help identify neuromarkers of early CSVD-related cognitive impairment to improve the early identification and monitoring of the disease. Moreover, EEG-based brain-computer interfaces (BCIs) could support symptom management through neurofeedback techniques [9] that modulate these neuromarkers and potentially reduce associated cognitive deficit.

Few studies have investigated EEG neuromarkers of CSVD-related cognitive impairment, and heterogeneity in VCI definitions complicates interpretation, as many studies do not distinguish VCI subtypes. Babiloni et al. [10] reported reduced resting-state posterior alpha power, widespread increase in delta and theta activity, and prolonged N200 and P300 latencies in VCI patients. Dey et al. [11] found similar results in their systematic review focusing on CSVD patients. They additionally reported reduced coherence of higher-frequency rhythms across fronto-parietal networks, interpreted as markers of disrupted long-range connectivity linking white matter damage to attentional and executive dysfunction. In their systematic review of EEG and MEG biomarkers of VCI, Torres-Simón et al. [12] attempted to categorize the studies they included according to the VCI subtypes defined by the VICCCS classification and found EEG alterations consistent with previous reviews. While these reviews have substantially advanced understanding of EEG alterations in VCI, important gaps remain. First, only Dey et al. focuses exclusively on CSVD as the underlying pathology: Babiloni et al. evaluated major VCI as a broad category, and Torres-Simón et al. could not draw SiVaD-specific conclusions due to diagnostic

heterogeneity. Second, none of the existing reviews has examined the implications of these EEG findings for neurofeedback-based therapeutic interventions, despite growing interest in non-pharmacological brain stimulation approaches for cognitive impairment.

The aim of this narrative review is to summarize and update the findings from the existing reviews with a focus on CSVD patients with mild or major cognitive impairment. We also discuss the methodological heterogeneity across studies and the challenges associated with translating these findings into tools for the diagnosis or monitoring of covert CSVD. Finally, we outlined the open questions that remain to be addressed for the development of effective neurofeedback interventions.

MATERIALS AND METHODS

A search of the literature was conducted using PubMed, Scopus and PsycInfo databases with the following query in title and abstract: ("small vessel disease" OR "small-vessel disease" OR "subcortical vascular dementia" OR "subcortical ischemic vascular dementia" OR "subcortical vascular cognitive impairment" OR "subcortical vascular lesion" OR "subcortical cerebrovascular lesion" OR "subcortical vascular damage" OR "subcortical cerebrovascular damage") AND (EEG OR EEGs OR electroencephalogra* OR electro-encephalogra* OR "neural oscillation" OR "brain oscillation" OR ERP OR ERPs OR "event related potential" OR "event-related potential" OR P300 OR N200 OR N400 OR "spectral power" OR "power spectrum" OR "frequency band" OR "functional connectivity") AND (cognit* OR "cognitive dysfunction" OR "cognitive changes" OR "cognitive impairment" OR "cognitive disorder" OR "cognitive dysfunction" OR "cognitive decline" OR "mild cognitive impairment" OR attention OR "executive function" OR "memory").

The database search identified 160 references. Many articles identified using the term "functional connectivity" were excluded because they focused exclusively on MRI analyses rather than EEG data. After screening titles and abstracts according to the inclusion criteria, 18 articles were retained for further evaluation. Because several relevant EEG studies did not explicitly use terms related to CSVD in their titles, abstracts, or keywords, additional studies were identified through manual screening of the reference lists of selected articles, particularly from review papers by Dey et al. [11] and Torres-Simón et al. [12]. In total, 27 articles were included in the present review.

Studies were included in this review if they reported analyses of EEG data in patients with cognitive impairment associated with CSVD, or subcortical vascular lesions such as WMH and lacunes confirmed by brain imaging. Only peer-reviewed papers published in English were considered, with no restriction on the date of publication. Book chapters and case studies were excluded as well as studies including patients with

comorbid neurological or psychiatric disorders (e.g., Parkinson's disease or depression).

RESULTS

The main findings indicate that patients with VCI associated with subcortical lesions show a disruption in spectral activity, functional connectivity and event-related potentials (ERP) compared to healthy controls (HC). Some studies found a correlation between these results and neuropsychological scores or lesion burden. Multiple studies also compared EEG markers between VCI patients and Alzheimer's disease (AD) patients. The results of this review are detailed below and summarized in Table 1.

Spectral analysis: Multiple studies show that CSVD patients present a slowing of brain activity compared to HC with some differences compared to AD patients. Main results included a mean frequency in the CSVD group lower than in the HC group but higher than in the AD group [13], [14], slowing of theta/alpha transition frequency [15], [16], reduced alpha peak frequency [16], [17], increased power in delta and theta bands, decreased alpha power and decreased alpha/slow wave power ratio [13], [14], [15], [16], [18], [19], [20], [21], [22], [23]. Specifically, low-resolution brain electromagnetic tomography (LORETA) studies show that SiVaD patients exhibit largely abnormal distributed theta sources compared to AD patients [18] and present a significant decrease in parieto-occipital alpha 1 (8 - 10.5 Hz) and beta 1 (13 - 22 Hz) relative power current density and a decrease in frontal delta relative power current density compared with patients with mild non-vascular cognitive impairment [20].

Connectivity analysis: CSVD damages long-range white matter tracts, especially frontosubcortical circuits. Using EEG synchronization likelihood, Babiloni et al. reported reduced fronto-parietal and interhemispheric coupling in SiVaD and AD patients compared to HC, with a stronger fronto-parietal alpha 1 reduction in VaD patients [24]. Moretti et al. also found reduced fronto-parietal and fronto-temporal coherence in patients with mild cognitive impairment due to subcortical cerebrovascular disease, with increased low-frequency and decreased high-frequency coherence linked to subcortical damage [25], [26]. Studies using phase lag index (PLI) also showed reduced alpha-band connectivity in CSVD patients with VCI compared to CSVD patients without VCI, and altered directional connectivity patterns compared to HC [27], [28]. Additionally, studies showed disrupted long-range networks in the alpha band for patients with WMH, particularly in long-range connections to the frontal lobe [22] and abnormal changes in EEG microstates compared to HC [29].

Event-related potential analysis: Studies using auditory tasks using an oddball paradigm in SiVaD patients report a smaller amplitude in N100 [30], a smaller amplitude and longer latency in P300 [30], [31], a longer latency in N200 [32] and a smaller amplitude

and shorter latency in frontocentral mismatch negativity [33]. P300 prolonged latency during auditory tasks also seems to appear at an earlier stage in CSVD compared to AD [31]. Regarding studies using a visual oddball paradigm, SiVaD patients exhibit a smaller N75 to P100 amplitude [34] and a smaller amplitude and longer latency in P300 [35]. In a visual memory workload task, patients also had lower P450 amplitude than AD patients and longer latencies over anterior regions and shorter latencies over posterior regions compared to HC [36].

Correlation with neuropsychological scores or lesion burden: Some studies report correlations between these EEG abnormalities and the severity of dementia measured by neuropsychological scores such as the Mini-Mental State Examination (MMSE) [37], or with lesion burden scales such as the Fazekas score [38]. Zhao et al. found that abnormal alpha band oscillations were correlated with Fazekas score and neuropsychological tests [17]. Gawel et al. showed that

the alpha/slow-wave power ratio differentiates mild from moderate dementia in CSVD patients [13], [14] while Moretti et al. proposed the theta/alpha1 ratio as an indicator of subcortical damage [15]. Guo et al. also reported that Montreal Cognitive Assessment (MoCA) scores [39] were negatively correlated with delta power and positively with alpha power and with Fazekas score. Moreover, reduced global average power spectral density value and low education were risk factors for cognitive impairment independently of Fazekas score [21]. Sheorajpanday et al. also identified (delta+theta)/(alpha+beta) ratio as an independent predictor of VCI and found that pairwise derived brain symmetry index correlated with neuropsychological tests [23]. Connectivity measures also relate to disease progression, such as alpha-band PLI positively correlated with MMSE and coherence changes linked to WMH severity and poorer cognitive performance [26], [27].

Table 1: Summary of studies assessing EEG markers of VCI in patients with vascular subcortical lesions.

EEG feature	Main findings for VCI patients	Related references
Spectral	VCI vs HC: Decreased alpha and beta power Widespread increase in delta and theta power Slowing of theta/alpha transition frequency Lower mean frequency Reduced alpha peak frequency	[13]*, [14]*, [15] ^o , [16], [17]* ^o , [19]*, [20], [21]* ^o , [22] ^o , [23]*
	VCI vs AD: Abnormally distributed theta sources Higher alpha/slow wave power ratios and mean frequency Higher delta and theta power	[14]*, [16], [18]
Connectivity	VCI vs HC: Reduction of fronto-parietal and interhemispheric synchronization likelihood Increase of coherence in low frequency and decrease of coherence in high frequency in right fronto-temporal and temporo-parietal neural networks. Abnormal changes in microstates Lower phase-lag index in alpha band and abnormal directional pattern	[22] ^o , [24], [25], [26]* ^o , [27]*, [28], [29]*
	VCI vs AD: Lower interhemispheric alpha 1 synchronization likelihood More homogeneous decrease of synchronization likelihood over the whole scalp	[24]
Event-related potentials	VCI vs HC: Longer latency and/or smaller amplitude	[30]*, [31], [32]*, [33], [35]*, [36]
	VCI vs AD: Lower P450 amplitude, lower N75-P100 amplitude Prolonged P300 latency earlier in the evolution of the disease	[31], [34], [36],

VCI : vascular cognitive impairment, with or without dementia depending on studies ; AD : Alzheimer's disease ; HC : healthy controls. References in bold indicate that the authors found a correlation between EEG markers and neuropsychological score (indicated by the "*" symbol) and/or the severity of cerebrovascular lesions (indicated by the "°" symbol).

DISCUSSION

Heterogeneity in diagnostic criteria and terminology: This review included studies on CSVD patients as well as those with subcortical vascular cognitive impairment or dementia, since CSVD is the main pathology in these conditions. However, inconsistent terminology across studies represents a limitation to this review: some describe CSVD lesions without naming them, while others report VCI without specifying subcortical origin. Consequently, relevant studies may have been missed despite the search strategy. Most articles diagnosed SiVaD using NINDS-

AIREN criteria [40], though some Chinese studies used national guidelines without official English translations. Some studies also assessed the Hachinski Ischemic Score [41] as additional criteria, particularly when differentiating with AD patients. Diagnoses typically required cognitive impairment or dementia supplemented by MRI or CT examination to ensure the subcortical vascular involvement and exclude patients with cortical or non-vascular lesions. However, methodological details varied: some studies mentioned STRIVE guideline [7] or rated WMH severity with Fazekas or ARWMC scales [42], but most only indicated the presence of lacunes or leukoaraiosis

without quantifying the severity of lesions. Variability in lesion burden, non-linear VCI progression, and limitations of MMSE for diagnosis of mild cognitive impairment [43] complicate comparisons across studies. These limitations are consistent with those reported by Torres-Simón et al. [12], who similarly noted that diagnostic heterogeneity prevented SiVaD-specific conclusions.

Use of EEG for diagnosis of asymptomatic CSVD: A recent article proposed an algorithm to diagnose CSVD patients with EEG and report a 96% accuracy in differentiating patients with or without VCI from HC [44]. However, the article does not specify whether the cross-validation procedure ensured strict subject separation between training and test sets, creating a substantial risk of intra-subject information leakage which can artificially inflate classification performance [45]. Other studies highlight potential EEG markers in asymptomatic CSVD: lower amplitude and longer latency of P300 [35], [31], increased theta power and disrupted alpha network [22], altered theta/alpha transition frequency [15], and abnormal EEG activity in regions involved in the default mode network [20] could be markers for cerebrovascular damage at an early stage of the disease, prior to clinical presentation. However, more longitudinal studies are necessary to assess if EEG markers could help predict the development of VCI in CSVD patients. Another limitation is the fact that the studies included in this review collected EEG data only during resting state conditions. The analysis of brain oscillation during cognitive tasks could be promising to better detect CSVD patients from healthy individuals and to monitor the progression of the disease.

Development of EEG-based neurofeedback therapies: To the best of our knowledge, only one study investigated neurofeedback for cognitive impairment in CSVD [46]. The authors report that an infralow-wave neurofeedback training increased alpha rhythms and improved patients' executive function up to 6–8 weeks post-training. However, this article did not provide detailed information about the neurofeedback protocol used during the experiment, complicating the replication of its results and the assessment of their validity [47]. More generally, there is a limited number of studies on EEG neurofeedback in patients with mild cognitive impairment or dementia [48]. Potential interventions for CSVD could target the alpha frequency band by increasing alpha power spectral density or alpha peak frequency as previous studies have reported that alpha-band training can improve cognitive performance [49]. Theta neurofeedback protocols also seem promising in improving executive function in older adults [50] and may warrant further investigation in patients with CSVD. A second neurofeedback approach could focus on coherence between brain regions in order to improve functional connectivity, but this type of protocol remains relatively underexplored. Similarly, ERP-based neurofeedback has primarily been applied to attentional training in individual with attention deficit hyperactivity disorder (ADHD) [51] but its use for broader cognitive

training is limited [52]. Therefore, given the limited evidence, more studies are necessary to assess whether neurofeedback can improve cognitive performance, particularly in CSVD patients. In addition to selecting appropriate EEG markers, several practical factors should be considered when designing neurofeedback protocols [53]. As these interventions primarily target older adults, the number and duration of sessions should be optimized to maintain engagement and minimize dropout, a common challenge in this population. Moreover, the type of feedback provided during training should be adapted to older adults, as intuitive and easily interpretable feedback may enhance both engagement and performance.

CONCLUSION

Subcortical lesions associated with CSVD often eventually lead to VCI or dementia and cause several EEG abnormalities. The most consistent finding is spectral slowing, characterized by increased delta and theta power and reduced alpha power. Other important results include disrupted functional connectivity, especially in connections involving frontal regions, and altered event-related potentials with reduced amplitude and prolonged latencies. Importantly, these changes can appear early, before cognitive impairment is evident. Some studies also report significant correlations between EEG markers and cognitive scores or lesion severity, suggesting that they may reflect disease progression. However, differences in diagnostic criteria and in the assessment of lesion burden and cognitive impairment limit the generalization of these findings. Moreover, most of the studies used resting-state EEG, leaving a gap in the literature regarding task-related activity. These findings also suggest potential for neurofeedback interventions. Future work will aim to identify EEG markers in patients with CSVD during cognitive tasks to guide the development of neurofeedback protocols and test whether modulating these signals improves cognition.

ACKNOWLEDGMENTS

This work was supported by a grant overseen by the French National Research Agency (ANR) as part of the “France 2030” Program, ANR-23-IAHU-0001 (Vascular Brain Health Institute – VBHI).

REFERENCES

- [1] Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *Lancet Neurol*. 2010
- [2] Markus HS, Joutel A. The pathogenesis of cerebral small vessel disease and vascular cognitive impairment. *Physiol. Rev*. 2025
- [3] Smith EE, Saposnik G, Biessels GJ, Doubal FN, Fornage M, Gorelick PB, et al.. Prevention of Stroke in Patients With Silent Cerebrovascular Disease: A

- Scientific Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke*. 2017
- [4] Skrobot OA, Black SE, Chen C, DeCarli C, Erkinjuntti T, Ford GA, et al.. Progress toward standardized diagnosis of vascular cognitive impairment: Guidelines from the Vascular Impairment of Cognition Classification Consensus Study. *Alzheimers Dement*. 2018
 - [5] Zanon Zotin MC, Sveikata L, Viswanathan A, Yilmaz P. Cerebral small vessel disease and vascular cognitive impairment: from diagnosis to management. *Curr. Opin. Neurol*. 2021
 - [6] Wallin A, Román GC, Esiri M, Kettunen P, Svensson J, Paraskevas GP, et al.. Update on Vascular Cognitive Impairment Associated with Subcortical Small-Vessel Disease. *J. Alzheimers Dis*. 2018
 - [7] Duering M, Biessels GJ, Brodtmann A, Chen C, Cordonnier C, De Leeuw F-E, et al.. Neuroimaging standards for research into small vessel disease—advances since 2013. *Lancet Neurol*. 2023
 - [8] Ter Telgte A, Van Leijsen EMC, Wiegertjes K, Klijn CJM, Tuladhar AM, De Leeuw F-E. Cerebral small vessel disease: from a focal to a global perspective. *Nat. Rev. Neurol*. 2018
 - [9] Sitaram R, Ros T, Stoeckel L, Haller S, Scharnowski F, Lewis-Peacock J, et al.. Closed-loop brain training: the science of neurofeedback. *Nat. Rev. Neurosci*. 2017
 - [10] Babiloni C, Arakaki X, Bonanni L, Bujan A, Carrillo MC, Del Percio C, et al.. EEG measures for clinical research in major vascular cognitive impairment: recommendations by an expert panel. *Neurobiol. Aging*. 2021
 - [11] Dey AK, Stamenova V, Turner G, Black SE, Levine B. Pathoconnectomics of cognitive impairment in small vessel disease: A systematic review. *Alzheimers Dement*. 2016
 - [12] Torres-Simón L, Doval S, Nebreda A, Llinas SJ, Marsh EB, Maestú F. Understanding brain function in vascular cognitive impairment and dementia with EEG and MEG: A systematic review. *NeuroImage Clin*. 2022
 - [13] Gawel M, Zalewska E, Szmids-Szałkowska E, Kowalski J. The value of quantitative EEG in differential diagnosis of Alzheimer's disease and subcortical vascular dementia. *J. Neurol. Sci*. 2009
 - [14] Gawel M, Zalewska E, Szmids-Szałkowska E, Kowalski J. Does EEG (visual and quantitative) reflect mental impairment in subcortical vascular dementia? *J. Neurol. Sci*. 2007
 - [15] Moretti DV, Miniussi C, Frisoni G, Zanetti O, Binetti G, Geroldi C, et al.. Vascular damage and EEG markers in subjects with mild cognitive impairment. *Clin. Neurophysiol*. 2007
 - [16] Moretti DV, Babiloni C, Binetti G, Cassetta E, Dal Forno G, Ferreris F, et al.. Individual analysis of EEG frequency and band power in mild Alzheimer's disease. *Clin. Neurophysiol*. 2004
 - [17] Zhao M, Hu G, Lu Y, Yang Q, Chen X, Wang D, et al.. Association between electroencephalogram alpha-band oscillations and executive and processing functions in patients with cerebral small vessel diseases. *J. Cereb. Blood Flow Metab*. 2024
 - [18] Babiloni C, Binetti G, Cassetta E, Cerboneschi D, Dal Forno G, Del Percio C, et al.. Mapping distributed sources of cortical rhythms in mild Alzheimer's disease. A multicentric EEG study. *NeuroImage*. 2004
 - [19] Van Straaten EC, De Haan W, De Waal H, Scheltens P, Van Der Flier WM, Barkhof F, et al.. Disturbed oscillatory brain dynamics in subcortical ischemic vascular dementia. *BMC Neurosci*. 2012
 - [20] Sheorajpanday RVA, Marien P, Weeren AJTM, Nagels G, Saerens J, Van Putten MJAM, et al.. EEG in Silent Small Vessel Disease: sLORETA Mapping Reveals Cortical Sources of Vascular Cognitive Impairment No Dementia in the Default Mode Network. *J. Clin. Neurophysiol*. 2013
 - [21] Guo X, Liu Z, Yuan W, Wei A, Luo G. Electroencephalographic power spectrum changes in cerebral small vessel disease combined with cognitive dysfunction and its relationship with neutrophil/lymphocyte ratio and its clinical value – a pilot study. *Front. Neurol*. 2024
 - [22] Quandt F, Fischer F, Schröder J, Heinze M, Lettow I, Frey BM, et al.. Higher white matter hyperintensity lesion load is associated with reduced long-range functional connectivity. *Brain Commun*. 2020
 - [23] Sheorajpanday RVA, Marien P, Nagels G, Weeren AJTM, Saerens J, Van Putten MJAM, et al.. Subcortical Vascular Cognitive Impairment, No Dementia: EEG Global Power Independently Predicts Vascular Impairment and Brain Symmetry Index Reflects Severity of Cognitive Decline. *J. Clin. Neurophysiol*. 2014
 - [24] Babiloni C, Ferri R, Moretti DV, Strambi A, Binetti G, Dal Forno G, et al.. Abnormal fronto-parietal coupling of brain rhythms in mild Alzheimer's disease: a multicentric EEG study. *Eur. J. Neurosci*. 2004
 - [25] Moretti DV, Frisoni GB, Pievani M, Rosini S, Geroldi C, Binetti G, et al.. Cerebrovascular disease and hippocampal atrophy are differently linked to functional coupling of brain areas: an EEG coherence study in MCI subjects. *J. Alzheimers Dis. JAD*. 2008
 - [26] Moretti DV, Pievani M, Geroldi C, Binetti G, Zanetti O, Cotelli M, et al.. Increasing Hippocampal Atrophy and Cerebrovascular Damage Is Differently Associated With Functional Cortical Coupling in MCI Patients. *Alzheimer Dis. Assoc. Disord*. 2009
 - [27] Liu D. Relationship between phase lag index measured by electroencephalography and cognitive dysfunction in patients with cerebral small vessel disease. *Am. J. Transl. Res*. 2025
 - [28] Straaten ECW van, Haan J den, Waal H de, Flier WM van der, Barkhof F, Prins ND, et al.. Disturbed

- phase relations in white matter hyperintensity based vascular dementia: An EEG directed connectivity study. *Clin. Neurophysiol.* 2015
- [29] Jin X, Zhou Y, Zhang D, Chen X, Ruan J, Fang D. Analysis of brain dynamics changes in patients with cerebral small vessel disease-related cognitive impairment based on electroencephalogram microstates. *Eur. J. Med. Res.* 2025
- [30] Yamaguchi S, Tsuchiya H, Yamagata S, Toyoda G, Kobayashi S. Event-related brain potentials in response to novel sounds in dementia. *Clin. Neurophysiol.* 2000
- [31] Muscoso EG, Costanzo E, Daniele O, Maugeri D, Natale E, Caravaglios G. Auditory event-related potentials in subcortical vascular cognitive impairment and in Alzheimer's disease. *J. Neural Transm.* 2006
- [32] Van Harten B, Laman DM, Van Duijn H, Knol DL, Stam CJ, Scheltens P, et al.. The Auditory Oddball Paradigm in Patients with Vascular Cognitive Impairment: A Prolonged Latency of the N2 Complex. *Dement. Geriatr. Cogn. Disord.* 2006
- [33] Jiang S, Yan C, Qiao Z, Yao H, Jiang S, Qiu X, et al.. Mismatch negativity as a potential neurobiological marker of early-stage Alzheimer disease and vascular dementia. *Neurosci. Lett.* 2017
- [34] Rosengarten B, Paulsen S, Molnar S, Kaschel R, Gallhofer B, Kaps M. Activation-flow coupling differentiates between vascular and Alzheimer type of dementia. *J. Neurol. Sci.* 2007
- [35] Nie S, Shen C, Guo Y, Hou X, Hong Y, Xu S, et al.. Preliminary Findings on Visual Event-Related Potential P3 in Asymptomatic Patients with Cerebral Small Vessel Disease. *Neuropsychiatr. Dis. Treat.* 2021
- [36] Beuzeron-Mangina H, Mangina CA. Excessive compensatory recruitment as a compulsory neurophysiological mechanism in Very Early Alzheimer's Disease as compared to Mild Vascular Dementia and to age-matched normal controls. *Int. J. Psychophysiol.* 2009
- [37] Folstein MF, Folstein SE, McHugh PR. « Minimal mental state ». A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 1975
- [38] Fazekas F, Chawluk JB, Alavi A, Hurtig HI, Zimmerman RA. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *AJR Am. J. Roentgenol.* 1987
- [39] Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al.. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *J. Am. Geriatr. Soc.* 2005
- [40] Román GC, Tatemichi TK, Erkinjuntti T, Cummings JL, Masdeu JC, Garcia JH, et al.. Vascular dementia: Diagnostic criteria for research studies: Report of the NINDS-AIREN International Workshop. *Neurology.* 1993
- [41] Rosen WG, Terry RD, Fuld PA, Katzman R, Peck A. Pathological verification of ischemic score in differentiation of dementias. *Ann. Neurol.* 1980
- [42] Wahlund LO, Barkhof F, Fazekas F, Bronge L, Augustin M, Sjögren M, et al.. A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke.* 2001
- [43] Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J. Psychiatr. Res.* 2009
- [44] Guo X, Shen L, Li L, Chang S, Liu Z, Shi Y. Cog-EEG: Early Detection of Cerebrovascular Patients With Cognitive Impairment Using EEG. *IEEE Internet Things J.* 2025
- [45] Brookshire G, Kasper J, Blauch NM, Wu YC, Glatt R, Merrill DA, et al.. Data leakage in deep learning studies of translational EEG. *Front. Neurosci.* 2024
- [46] Dobrynina LA, Novikova ES, Dobrushina OR, Gnedovskaya EV, Korepina OS, Byrochkina AA. A Neural Training Method with Biological Feedback in the Treatment of Cognitive Impairments in Patients with Early Cerebral Small Vessel Disease. *Neurosci. Behav. Physiol.* 2025
- [47] Ros T, Enriquez-Geppert S, Zotev V, Young KD, Wood G, Whitfield-Gabrieli S, et al.. Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). *Brain.* 2020
- [48] Tazaki M. A review: effects of neurofeedback on patients with mild cognitive impairment (MCI), and Alzheimer's disease (AD). *Front. Hum. Neurosci.* 2024
- [49] Angelakis E, Stathopoulou S, Frymiare JL, Green DL, Lubar JF, Kounios J. EEG Neurofeedback: A Brief Overview and an Example of Peak Alpha Frequency Training for Cognitive Enhancement in the Elderly. *Clin. Neuropsychol.* 2007
- [50] Viviani G, Vallesi A. EEG-neurofeedback and executive function enhancement in healthy adults: A systematic review. *Psychophysiology.* 2021
- [51] Fouillen M, Maby E, Le Carrer L, Herbillon V, Mattout J. ERP-based BCI training for children with ADHD: motivations and trial design, in *Proc. 7th Graz Brain-Computer Interface Conf., Graz, Austria, 2017,*
- [52] Ni J, Giles J, Arvaneh M. Feasibility and Acceptability of P300-Based Brain-Computer Interface Neurofeedback Training for Cognitive Enhancement, in *Proc. IEEE Int. Conf. MetroXRaine, 2025*
- [53] Enriquez-Geppert S, Huster RJ, Herrmann CS. EEG-Neurofeedback as a Tool to Modulate Cognition and Behavior: A Review Tutorial. *Front. Hum. Neurosci.* 2017

BENCHMARKING POSITIONAL ENCODING STRATEGIES FOR TRANSFORMER-BASED EEG FOUNDATION MODELS

Ayşe Betül Yüce¹, Sebastian Stober¹

¹Department of Computer Science, Otto von Guericke University, Magdeburg, Germany

E-mail: ayse.yuece@ovgu.de

ABSTRACT: Electroencephalography (EEG) is a widely used non-invasive technique for measuring brain activity in brain-computer interface (BCI) applications. Supervised EEG decoding models often struggle to generalize across tasks, subjects, and datasets, motivating transformer-based EEG foundation models trained with self-supervised learning. Since transformers are permutation-invariant, they require explicit positional information. Unlike textual tokens, EEG electrodes are spatially distributed across the scalp, raising the question of how electrode positions should be encoded in transformer-based EEG models. In this study, we benchmark five positional encoding strategies within the CBraMod backbone and evaluate them under linear probing and fine-tuning protocols on motor imagery classification and emotion recognition. Our results show that no single strategy consistently outperforms across tasks. Spherical Positional Encoding (SPE) yields strong representations for motor imagery but underperforms on emotion recognition, while Asymmetric Conditional Positional Encoding (ACPE) demonstrates more consistent performance across tasks. These findings suggest that the optimal positional encoding strategy is task-dependent, with no universal solution across EEG decoding scenarios.

INTRODUCTION

Electroencephalography (EEG) is a brain activity recording technique that provides high temporal resolution. Due to its non-invasive nature, it is widely preferred in both clinical[1, 2] and brain-computer interface (BCI) applications[3, 4]. However, EEG signals present substantial challenges. They exhibit a low signal-to-noise ratio and are highly susceptible to artifacts, session variability, and inter-subject differences. Furthermore, their high temporal resolution results in complex, high-dimensional data that require careful modeling. Early BCI systems relied on supervised learning with handcrafted features, which often generalize poorly across tasks, subjects, and datasets. These limitations motivate the development of scalable representation learning frameworks capable of learning transferable neural representations.

Inspired by advances in large-scale modeling in natural language processing[5] and computer vision[6], recent studies have introduced transformer-based founda-

tion models for EEG [7–11]. These models leverage self-supervised pretraining to learn general neural representations that can be adapted to downstream tasks. Since transformers operate through self-attention and are inherently permutation-invariant, they require explicit positional information to encode structural dependencies. Consequently, positional encoding (PE) becomes a critical architectural component. EEG electrodes follow anatomically meaningful spatial configurations that vary across datasets due to differences in channel number and electrode placement.

Recent transformer-based EEG foundation models adopt diverse mechanisms to incorporate positional information while accounting for the spatial-temporal structure of EEG signals. BIOT [8] addressed cross-dataset variability using channel-specific embeddings combined with relative temporal positional encoding. Subsequently, Wang J et al. [9] proposed an asymmetric conditional positional encoding (ACPE) module based on depthwise two-dimensional convolutions, where positional representations are dynamically generated from the spatio-temporal neighborhoods of EEG patches. By employing asymmetric kernels with a larger receptive field along the spatial dimension and a smaller one along the temporal dimension, their design captures long-range inter-channel dependencies while preserving short-range temporal context, enabling adaptation to varying EEG formats. Moving toward geometry-aware strategies, Ouahidi YE et al. [12] leveraged Cartesian (x, y, z) electrode coordinates, augmented with noise to improve robustness to head-size and placement variations, and extended them with temporal indices to form a four-dimensional (x, y, z, t) representation. The resulting 4D representation was projected into a multi-frequency Fourier space and combined with learned embeddings. Some researchers learned spatial and temporal embeddings during training and used them as positional encoding [10, 13]. These diverse implementations highlight the lack of systematic evaluation of positional encoding strategies in EEG foundation models.

As part of this benchmark, we introduce Spherical Positional Encoding (SPE), which leverages the spherical geometry of electrode placement. For each electrode, azimuth and inclination angles are computed and transformed into fixed-length representations using sinusoidal functions at multiple frequencies, and the final embedding is obtained by concatenating these angu-

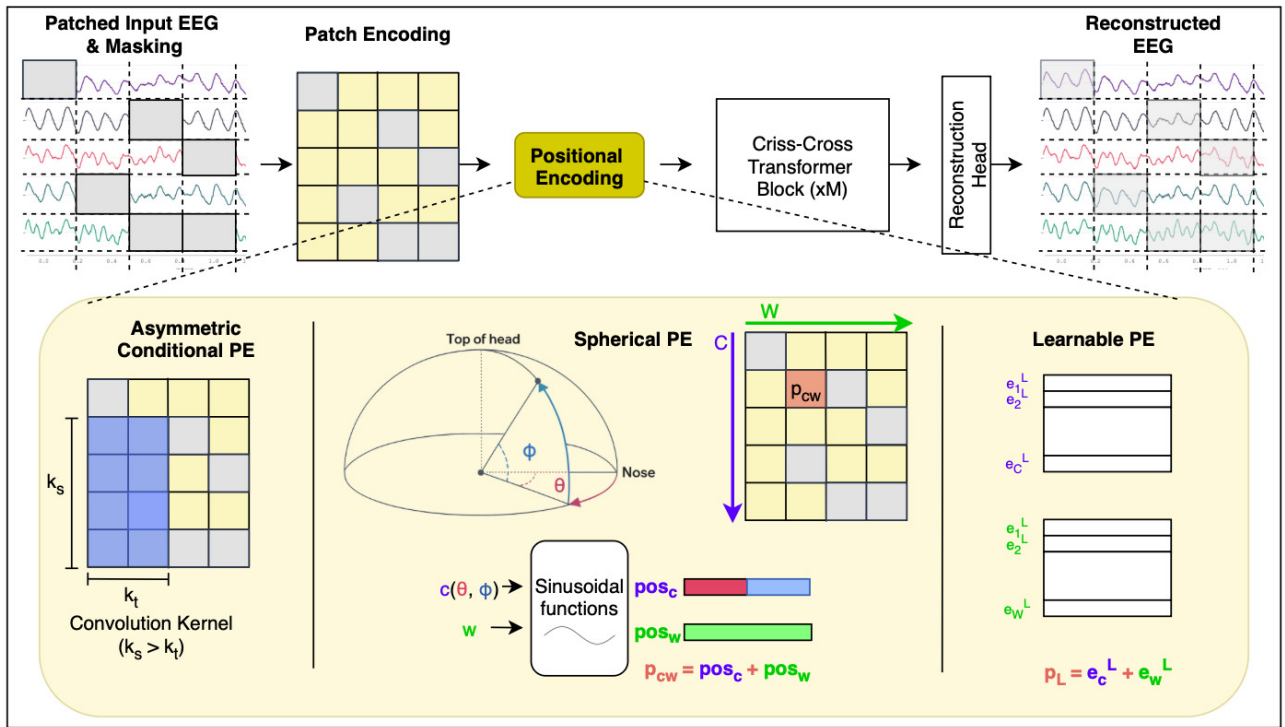


Figure 1: Overview of the experimental setup. The CBraMod backbone is retained with its original transformer blocks, while the positional encoding module is replaced with three primary variants under investigation: asymmetric conditional PE, Spherical PE, and learnable PE. Additionally, we extend SPE with a projection module and include a no positional encoding baseline.

lar encodings. By encoding angular relationships rather than absolute Cartesian positions, SPE preserves intrinsic scalp topology while remaining invariant to global scaling and head-size differences, without introducing additional learnable parameters.

In this work, we systematically evaluate positional encoding strategies within the CBraMod foundation model, which employs a criss-cross transformer mechanism to capture temporal and spatial dependencies between EEG patches. As illustrated in Fig. 1, we compare three primary strategies: ACPE, SPE, and learnable spatio-temporal PE. Additionally, we extend SPE with a projection module and include a no positional encoding baseline to further isolate the contribution of positional information. By keeping the backbone architecture fixed and modifying only the positional encoding component, we isolate the impact of positional design on representation learning and downstream performance. The contributions of this work are summarized as follows:

- We conduct a **controlled benchmark of positional encoding strategies** in a transformer-based EEG foundation model by keeping the backbone architecture fixed and modifying only the positional encoding component.
- We propose **Spherical Positional Encoding (SPE)**, which encodes electrode azimuth and inclination angles using multi-frequency sinusoidal mappings to incorporate spherical scalp geometry into transformer-based EEG models.

MATERIALS AND METHODS

In this study, we conduct a controlled comparison of positional encoding strategies within the CBraMod [9] backbone. The original positional encoding module is replaced while all other architectural components, training protocols, and hyperparameters are kept identical, allowing us to isolate the specific contribution of positional encoding to representation learning and downstream generalization. To further assess robustness and adaptability, we evaluate downstream performance across multiple datasets with differing electrode montages and channel counts, enabling analysis under heterogeneous spatial configurations.

Datasets: For pretraining the EEG foundation model, we used the Healthy Brain Network EEG dataset (HBN-EEG) [14–16], which comprises recordings from more than 3,000 participants aged 5–21 years. EEG data were acquired using a 128-channel GSN-HydroCel montage. The dataset includes 11 releases from different participants. Recordings were collected during both passive paradigms (e.g., resting state, surround suppression, movie watching) and active tasks (e.g., contrast change detection, sequence learning, symbol search). We used the publicly available downsampled (100 Hz) and band-pass filtered (0.5–50 Hz) version of the dataset. The dataset already includes 60 Hz notch filtering in its released preprocessing pipeline. Releases 1–9 were used for pretraining and Release 10 for pretraining validation. Preprocessing steps included removing the ref-

reference channel, automatically detecting and interpolating bad channels [17, 18], and applying average re-referencing. All preprocessing was performed using MNE-Python [19]. The pretrained models are evaluated on two downstream tasks: motor imagery classification and emotion recognition.

In motor imagery classification, we use the PhysioNet Motor Imagery (MI) dataset [20]. The dataset comprises EEG recordings from 109 subjects performing various motor and imagery tasks, acquired using 64 electrodes arranged according to the international 10–10 system at a sampling rate of 160 Hz. Only imagined movement trials are used, comprising four classes: left fist, right fist, both fists, and both feet.

For the emotion recognition task, we use the Fine-grained Affective Computing EEG Dataset (FACED) [21], in which EEG recordings were collected from 123 participants watching 28 video clips covering nine emotional categories: anger, disgust, fear, sadness, amusement, inspiration, joy, tenderness, and neutral. Recordings were acquired using 32 electrodes following the international 10–20 system. We use the preprocessed version of the dataset, in which electrode ordering and naming are standardized across recording cohorts, the sampling rate is 250 Hz, and a band-pass filter of 0.05–47 Hz was applied. To standardize input lengths and control the number of transformer tokens, longer recordings were segmented into fixed-length 10-second epochs, while shorter trials, such as the original 4-second PhysioNet-MI trials, were used directly. The differences in frequency ranges across datasets arise from the publicly available preprocessed versions used in this study. No additional band-pass filtering was applied beyond the provided preprocessing pipelines. All datasets were resampled to 200 Hz for consistency with the backbone configuration. Subjects in the downstream datasets were split in a subject-independent manner, with 70% assigned for training, 15% for validation, and 15% for testing. A random seed of 42 was used for dataset splitting.

Input data consist of multichannel EEG recordings $X \in \mathbb{R}^{C \times T}$, where C denotes the number of electrodes and T represents the temporal samples. During pretraining and fine-tuning, each EEG epoch is further divided into non-overlapping 1-second windows, which are treated as input patches to the model. After patch extraction, each EEG sample is represented as $x \in \mathbb{R}^{C \times w \times t}$, where t denotes the number of samples per patch and $w = \frac{T}{t}$ the number of patches per channel. Unlike the original CBraMod implementation, which normalizes signals by a fixed scalar of 100 after removing samples that exceed $100\mu V$, we apply channel-wise max-absolute normalization to scale each channel to the range $[-1, 1]$, as our pretraining dataset differs and comprises heterogeneous recordings with varying amplitude distributions across channels and subjects.

$$\tilde{X}_{c,T} = \frac{X_{c,T}}{\max(\max_T |X_{c,T}|, \varepsilon)}$$

where ε is a small constant to prevent division by zero. This formulation preserves both zero-crossing structure and signal polarity. While normalization may affect positional encoding methods differently, the same preprocessing pipeline is used across all evaluated methods. For self-supervised pretraining, we adopt the random masking strategy of CBraMod, in which a predefined ratio of patch tokens is randomly selected and replaced with learnable mask token prior to being processed by the transformer backbone.

CBraMod Architecture Overview: CBraMod is a self-supervised EEG foundation model designed to capture spatio-temporal dependencies via a criss-cross attention mechanism. The model first segments the EEG signal into non-overlapping patches. During pretraining, a random subset of patches is replaced with a learnable mask tokens. For each patch, temporal embeddings are extracted via convolutional layers and spectral embeddings are derived through a Fast Fourier Transform followed by a fully connected layer, and the two are fused via element-wise addition to obtain the final patch embedding. Positional encoding is then computed from the patch embeddings and added to form the input sequence to the transformer. In our experiments, the positional encoding module is systematically replaced to evaluate different strategies while keeping the rest of the architecture fixed.

The resulting sequence is processed by a criss-cross transformer backbone, in which attention is factorized into two operations: spatial attention, which models inter-channel relationships across electrodes, and temporal attention, which captures dependencies across time windows. This design enables the model to jointly learn spatial and temporal structure while maintaining computational efficiency. The model is optimized via masked patch reconstruction, where the mean squared error (MSE) between the original and reconstructed masked patches is used as the training objective:

$$L = \|\hat{X}_M - \tilde{X}_M\|_2^2$$

where $\tilde{X}_M$ and $\hat{X}_M$ denote the original normalized patch and reconstructed masked patch, respectively.

Positional Encoding Variants: ACPE. The Asymmetric Conditional Positional Encoding proposed in [9] dynamically generates positional encodings conditioned on patch embeddings via a depthwise two-dimensional convolution with an asymmetric kernel of size (k_s, k_t) , where $k_s > k_t$. The larger spatial kernel k_s is designed to capture longer-range inter-channel positional relationships, while the smaller temporal kernel k_t focuses on encoding local temporal positional information. The positional encoding is added to the original patch embedding before entering the transformer backbone.

SPE. While ACPE derives positional information implicitly from patch content, we propose Spherical Positional Encoding (SPE), which incorporates explicit anatomical priors by encoding the physical configuration of electrodes on the scalp. For each electrode, azimuth θ

and inclination ϕ angles are obtained from its canonical spatial location. Each angular component is mapped through multi-frequency sine and cosine functions to obtain bounded and continuous representations that preserve periodicity:

$$pos_c = [\sin(\omega_i\theta), \cos(\omega_i\theta), \sin(\omega_i\phi), \cos(\omega_i\phi)]$$

where ω_i denotes the i -th frequency of the sinusoidal basis. By using multiple frequencies, the encoding captures angular relationships at different spatial scales. The resulting features are concatenated to form the channel positional embedding $pos_c = [P_{azimuth}, P_{inclination}]$, providing a continuous representation of global scalp topology that is invariant to head size scaling.

For temporal structure, standard sinusoidal positional encoding [22] is applied over patch indices to obtain patch positional encoding pos_w . The final SPE is obtained by summation:

$$P_{cw} = pos_w + pos_c$$

This formulation jointly encodes spatial topology and temporal ordering without introducing additional learnable parameters.

As an extension, we additionally evaluate a projected variant (**SPE+Proj**), in which the channel and patch positional embeddings are each passed through separate linear projections before summation:

$$P_{cw} = W_w pos_w + W_c pos_c$$

where W_w and W_c are learnable projection matrices. This variant explores whether allowing the model to adaptively weight the fixed geometric and temporal encodings improves downstream performance.

Learnable PE. We also evaluate fully learnable positional encoding, in which both channel and patch positional embeddings are learned directly from data and summed to form the final positional encoding:

$$P_L = e_w^L + e_c^L$$

where $e_c^L \in \mathbb{R}^{C \times d}$ and $e_w^L \in \mathbb{R}^{W \times d}$ are learnable embedding matrices for channel and patch indices, respectively. The learned channel embeddings are montage-specific and cannot be directly transferred across datasets with different electrode configurations. Since the pretraining and downstream datasets in our setup use different montages, both channel and patch embeddings are reinitialized and learned from scratch during fine-tuning.

Implementation Details: We follow the same pretraining and fine-tuning procedure as CBraMod. All models are trained on 4× H100 80GB GPUs with 2× 24-core EPYC 9254 CPUs at 2.90 GHz and 1.5 TB RAM. Both pretraining and fine-tuning use 1-second non-overlapping patches as input. The transformer backbone consists of 12 layers with 8 attention heads, an embedding dimension of $d = 200$, and a feedforward dimension of 800. During pretraining, models are trained for 40 epochs with

a batch size of 32 and a learning rate of 10^{-5} scheduled with cosine annealing and the AdamW optimizer. Additionally, 50% of patch tokens are randomly selected and masked.

During fine-tuning, models are trained for 50 epochs with a batch size of 64 using the AdamW optimizer and label smoothing of 0.1. Weight decay during fine-tuning is set to 0.001. Model checkpoints are saved based on validation Cohen’s Kappa, with higher Kappa indicating improved performance. All fine-tuning experiments are repeated with 5 different random seeds, and downstream performance is reported in terms of balanced accuracy, Cohen’s Kappa, and weighted F1 score. For PhysioNet-MI fine-tuning, a three-layer classification head was used as it yielded better performance. For all other experiments, a single-layer head was used, as SPE exhibited training instability with a deeper classifier.

RESULTS

We evaluate the five positional encoding strategies under both linear probing and fine-tuning protocols on motor imagery classification and emotion recognition to assess how positional encoding design affects representation quality and downstream performance. Note that the results reported here are not directly comparable to those in the original CBraMod paper, as we use a different pre-training dataset and normalization strategy.

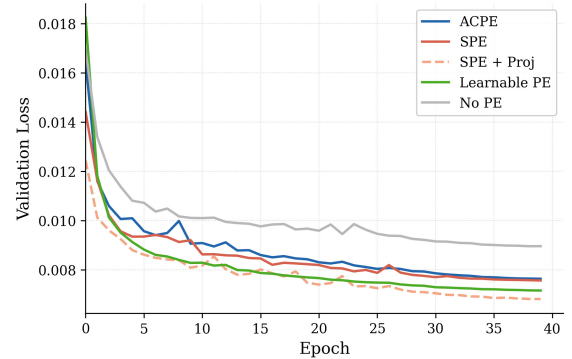


Figure 2: Pretraining masked reconstruction loss curves for CBraMod trained with five positional encoding strategies

As shown in Figure 2, the absence of positional encoding results in the highest pretraining validation loss, demonstrating that positional information contributes to masked patch reconstruction. Among the remaining strategies, SPE converges faster than ACPE in early epochs but reaches a similar final loss. Learnable PE and SPE+Proj achieve the lowest and comparable validation losses.

Tab. 1 reports linear probe performance, where all backbone weights are frozen and only a single-layer classification head is trained to evaluate the quality of pretrained representations. Since learnable PE requires channel embeddings to be reinitialized due to montage mismatch between pretraining and downstream datasets, it is excluded from linear probe evaluation. SPE achieves the best per-

Table 1: Linear probe performance comparison of each positional encoding strategy on motor imagery (PhysioNet MI) and emotion recognition (FACED) tasks.

Model	PhysioNet-MI (4-class)			FACED (9-class)		
	Bal. Acc.	Cohen’s Kappa	Weighted F1	Bal. Acc.	Cohen’s Kappa	Weighted F1
NoPE	0.4757 $\pm$ 0.0111	0.3010 $\pm$ 0.0148	0.4731 $\pm$ 0.0102	0.4098 $\pm$ 0.0082	0.3344 $\pm$ 0.0088	0.4056 $\pm$ 0.0078
ACPE	0.4993 $\pm$ 0.0065	0.3324 $\pm$ 0.0086	0.4940 $\pm$ 0.0069	0.4212 $\pm$ 0.0053	0.3461 $\pm$ 0.0062	0.4180 $\pm$ 0.0065
SPE	0.5105 $\pm$ 0.0057	0.3477 $\pm$ 0.0077	0.5090 $\pm$ 0.0047	0.3878 $\pm$ 0.0061	0.3064 $\pm$ 0.0075	0.3766 $\pm$ 0.0084
SPE+Proj.	0.4835 $\pm$ 0.0111	0.3116 $\pm$ 0.0149	0.4841 $\pm$ 0.0107	0.3466 $\pm$ 0.0043	0.2625 $\pm$ 0.0032	0.3421 $\pm$ 0.0010

Table 2: Fine-tuning performance comparison of each positional encoding strategy on motor imagery (PhysioNet MI) and emotion recognition (FACED) tasks.

Model	PhysioNet-MI (4-class)			FACED (9-class)		
	Bal. Acc.	Cohen’s Kappa	Weighted F1	Bal. Acc.	Cohen’s Kappa	Weighted F1
NoPE	0.5971 $\pm$ 0.0034	0.4628 $\pm$ 0.0045	0.5985 $\pm$ 0.0034	0.4975 $\pm$ 0.0116	0.4306 $\pm$ 0.0124	0.4929 $\pm$ 0.0108
ACPE	0.6171 $\pm$ 0.0115	0.4892 $\pm$ 0.0153	0.6175 $\pm$ 0.0116	0.5013 $\pm$ 0.0139	0.4335 $\pm$ 0.0147	0.4930 $\pm$ 0.0116
SPE	0.6009 $\pm$ 0.0107	0.4676 $\pm$ 0.0143	0.6001 $\pm$ 0.0109	0.4820 $\pm$ 0.0070	0.4136 $\pm$ 0.0083	0.4796 $\pm$ 0.0081
SPE+Proj.	0.5983 $\pm$ 0.0149	0.4642 $\pm$ 0.0199	0.5992 $\pm$ 0.0143	0.4952 $\pm$ 0.0113	0.4289 $\pm$ 0.0120	0.4895 $\pm$ 0.0090
Learnable PE	0.6001 $\pm$ 0.0147	0.4666 $\pm$ 0.0197	0.6008 $\pm$ 0.0034	0.5124 $\pm$ 0.0177	0.4467 $\pm$ 0.0184	0.5078 $\pm$ 0.0170

formance on PhysioNet-MI, while ACPE yields the best results on FACED.

On the other hand, Tab. 2 shows fine-tuning performance, where all model weights are updated. For learnable PE, both channel and patch positional embeddings are reinitialized and learned from scratch during fine-tuning. ACPE achieves the best performance on PhysioNet-MI, while Learnable PE yields the best results on FACED.

DISCUSSION

No single positional encoding strategy achieves consistently superior performance across both tasks and evaluation protocols, suggesting that the optimal PE design is task- and dataset-dependent. Although the observed performance differences within each evaluation protocol are relatively small, all experiments differ only in the positional encoding module while keeping the remaining architecture fixed. Therefore, even modest improvements indicate measurable effects of positional information on learned EEG representations.

NoPE is consistently the weakest strategy in linear probing, confirming that positional information contributes meaningfully to representation quality during pretraining. However, NoPE remains competitive with several PE strategies after fine-tuning, indicating that the transformer backbone can partially recover positional structure through end-to-end optimization on the downstream task.

SPE produces competitive representations in the linear probe protocol for motor imagery, suggesting that explicit scalp topology encoding can be beneficial when spatial electrode layout is relevant to the task. One possible explanation is that motor imagery signals may exhibit more spatially distinguishable activation patterns, making the spherical geometry prior a useful inductive bias during pretraining. In contrast, SPE underperforms on FACED under both protocols, which may indicate that emotion-related EEG representations benefit less from a strict geometric prior. However, these observations remain hypothesis-driven and require further investigation.

A practical advantage of SPE is that it requires no learned parameters and is precomputed from electrode angles, making it applicable to any montage without retraining.

SPE+Proj does not consistently improve over plain SPE despite introducing learnable projections. The learned projections likely corrupt the structured geometric prior rather than enhancing it.

ACPE either achieves the best performance or remains competitive across both tasks and protocols. By conditioning positional encodings on local patch content through asymmetric depthwise convolutions, ACPE captures both spatial and temporal neighborhood structure without relying on explicit electrode coordinates. This content-conditioned formulation appears more generalizable across diverse EEG tasks. However, since positional information is derived implicitly from patch representations rather than absolute electrode positions, ACPE is sensitive to channel ordering and performance may degrade if channels are permuted.

Learnable PE requires channel embeddings to be reinitialized during fine-tuning when the downstream montage differs from pretraining, which prevents direct transfer of spatial representations. Despite this, learnable PE achieves the best fine-tuning performance on FACED, suggesting that when the model is free to learn task-specific spatial structure from scratch on the target dataset, it adapts more effectively to the distributed spatial patterns characteristic of emotion recognition.

CONCLUSION

In this paper, we benchmarked the effect of positional encoding design on transformer-based EEG foundation models by systematically replacing the positional encoding module in the CBraMod backbone and evaluating five strategies under linear probing and fine-tuning protocols on motor imagery and emotion recognition tasks. Our results confirm that positional information is beneficial for pretraining, as NoPE yields the highest masked reconstruction loss and consistently weak linear probe performance on downstream tasks.

No single positional encoding strategy achieves consistently superior performance across both tasks, with results varying based on task and dataset characteristics. SPE yields strong representations for motor imagery, where spatially distinguishable patterns may make the geometric prior beneficial, but is less effective for emotion recognition. Further experiments are required to better understand the interaction between task characteristics and positional encoding design. Additionally, SPE is precomputed, parameter-free, and montage-agnostic, making it a lightweight option when spatial structure is relevant. ACPE, by learning local neighborhood patterns from patch content, demonstrates more consistent performance across both tasks. When the pretraining and downstream montages differ, relearning positional embeddings from scratch on the target domain can outperform fixed strategies.

Evaluations are limited to a single backbone architecture and two downstream datasets; results may differ with alternative architectures or larger and more diverse evaluation benchmarks. Future work could explore PE strategies across more diverse EEG tasks and electrode montages, as well as larger pretraining datasets to better assess the generalizability of topology-aware encoding approaches.

ACKNOWLEDGEMENTS

This work was supported by the AI Co-Working Lab research initiative, funded by the Investitionsbank Sachsen-Anhalt through the Research and Innovation Program (EFRE – European Regional Development Fund).

REFERENCES

- [1] Fernandez J, Innocenti B, López B. EEG classification for neurological disorders using frequency band deciles. *Scientific Reports*. 2025;15:45142.
- [2] Ávila CG, Bott FS, Tiemann L, Hohn VD, May ES, Nickel MM, et al. DISCOVER-EEG: an open, fully automated EEG pipeline for biomarker discovery in clinical neuroscience. *Scientific Data*. 2023;10:613.
- [3] Guerrero-Mendez CD, Blanco-Diaz CF, Ruiz-Olaya AF, López-Delis A, Jaramillo-Isaza S, Milanezi Andrade R, et al. EEG motor imagery classification using deep learning approaches in naïve BCI users. *Biomedical Physics & Engineering Express*. 2023;9(4).
- [4] Torres EPP, Torres EA, Hernández-Álvarez M, Yoo SG. EEG-based BCI emotion recognition: A survey. *Sensors*. 2020;20(18):5083.
- [5] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proc. NAACL*. 2019, 4171–4186.
- [6] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*. 2021.
- [7] Kostas D, Aroca-Ouellette S, Rudzicz F. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*. 2021;15.
- [8] Yang C, Westover M, Sun J. BIOT: Biosignal transformer for cross-data learning in the wild. In: *NeurIPS*. 2023, 78240–78260.
- [9] Wang J, Zhao S, Luo Z, Zhou Y, Jiang H, Li S, et al. CBraMod: A criss-cross brain foundation model for EEG decoding. In: *ICLR*. 2025.
- [10] Jiang W, Zhao L, Lu BL. Large brain model for learning generic representations with tremendous EEG data in BCI. In: *ICLR*. 2024.
- [11] Cui W, Jeong W, Thölke P, Medani T, Jerbi K, Joshi AA, et al. Neuro-GPT: Towards a foundation model for EEG. In: *IEEE ISBI*. 2024, 1–5.
- [12] Ouahidi YE, Lys J, Thölke P, Farrugia N, Pasdeloup B, Gripon V, et al. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*. 2025.
- [13] Klein T, Minakowski P, Sager S. Flexible patched brain transformer model for EEG decoding. *Scientific Reports*. 2025;15:10935.
- [14] Shirazi SY, Franco A, Hoffmann MS, Esper NB, Truong D, Delorme A, et al. HBN-EEG: The FAIR implementation of the Healthy Brain Network (HBN) electroencephalography dataset. *bioRxiv*. 2024.
- [15] Alexander LM, Escalera J, Ai L, Andreotti C, Febre K, Mangone A, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*. 2017;4(1):170181.
- [16] Langer N, Ho EJ, Alexander LM, Xu HY, Jozanovic RK, Henin S, et al. A resource for assessing information processing in the developing brain using EEG and eye tracking. *Scientific Data*. 2017;4(1):170040.
- [17] Bigdely-Shamlo N, Mullen T, Kothe C, Su KM, Robbins KA. The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*. 2015;9.
- [18] Perrin F, Pernier J, Bertrand O, Echallier JF. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*. 1989;72(2):184–187.
- [19] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013;7.
- [20] Schalk G. EEG motor movement/imagery dataset. *PhysioNet*. 2009.
- [21] Chen J, Wang X, Huang C, Hu X, Shen X, Zhang D. A large finer-grained affective computing EEG dataset. *Scientific Data*. 2023;10(1):740.
- [22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Proc. NeurIPS*. 2017, 6000–6010.

Gamification Improves Neurofeedback Performance and Experience among Users with ADHD: Findings from NeuroWizard II

H. Kohnen^{1,3}, P. Mavromoustakos-Blom¹, M. Alimardani²

¹Dept. of Intelligent Systems, Tilburg University, Tilburg, Netherlands

²Dept. of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

³Dept. of Psychology, University of Bamberg, Bamberg, Germany

E-mail: henriette.kohnen@uni-bamberg.de

ABSTRACT: This study explored the effect of gamification on neurofeedback (NF) training in individuals with Attention-Deficit/Hyperactivity Disorder (ADHD). In a within-subjects design, 16 participants with ADHD completed both a gamified and a non-gamified NF task delivered via a Brain-Computer Interface (BCI). Task performance and user experience were evaluated across conditions. The gamified condition led to a higher task accuracy and better user experience, characterized by greater immersion and lower negative affect. However, some participants also reported increased mental fatigue. Notably, most participants expressed a preference for the gamified version, describing it as more motivating and engaging. These findings suggest that carefully applied gamification can enhance both the effectiveness and enjoyment of NF training for individuals with ADHD, paving the way for the development of more user-centered therapeutic tools.

INTRODUCTION

Attention-Deficit/Hyperactivity Disorder (ADHD) is one of our modern society's most prevalent neurodevelopmental disorders, causing approximately 8-12% of the children worldwide to struggle with impulsive and disorganized behavior, physical and mental restlessness, and difficulties to keep focused attention [1]. Those behavioral symptoms relate to distinct neurophysiological characteristics such as elevated theta and reduced alpha and beta activity in the electroencephalographic (EEG) profiles of individuals with ADHD [2, 3] and can persist into adulthood for up to 65% of the diagnosed children [1]. ADHD can thereby not only impact the individual's education, professional career, and personal relationships but also increases their vulnerability to other psychiatric conditions and comorbidities like depression, anxiety, or bipolar disorder and substance abuse [4]. Though currently not curable, a range of therapeutic interventions have been shown to improve symptoms of ADHD and the quality of life of affected individuals [1]. Alongside medication-based interventions, non-pharmacological treatment options for example using neurofeedback (NF) delivered via a Brain-Computer Interface (BCI) have gained popularity in more recent times, to evade the drawbacks of a continuous use of med-

ication [5, 6].

During NF training, the user learns to actively regulate their brain activity through a variety of different mental strategies, for which they receive real-time feedback from a closed-loop BCI system [7]. However, given that attention, concentration, and motivation can influence BCI task performance [8], a significant limitation of standard NF and BCI training protocols lies in their reliance on repetitive and monotonous tasks; a flaw in protocol design that is thought to be partially responsible for BCI inefficiency [9]. BCI inefficiency describes the phenomenon whereby 15-30% of users are unable to successfully control a BCI [10], a problem that is crucially linked to the success of NF training as a treatment method [11, 12]. To address this issue, several studies within the broader BCI literature have proposed the use of gamification as a strategy to enhance user motivation and engagement during training [13]. Gamification refers to the integration of game design elements into non-game contexts to foster increased user participation and intrinsic motivation [14]. A recent systematic review identified 14 commonly used game elements in motor imagery-based BCI training systems and found evidence that gamification generally improves user experience and BCI training outcomes [15].

However, most of the current evidence on the impact of gamification in NF training is based on studies involving healthy individuals. Another recent systematic review [16] highlighted the limited and inconsistent research specifically investigating the effect of gamified NF training in populations with ADHD. While some studies suggest that gamification may enhance motivation and engagement in individuals with ADHD [17], others caution that gamified elements may introduce distractions or lead to frustration [5, 18, 19]. Importantly, out of the 219 studies screened by Kohnen et al. [16], only 14 empirical studies specifically examined gamified BCI applications for NF training in individuals with ADHD. Many of the excluded studies failed to validate their systems with formally diagnosed participants, thus violating the principles of user-centered design essential for the development of effective BCI systems [20].

Therefore, the state-of-the-art literature does not provide

reliable conclusions or guidelines for developing NF applications that offer users with ADHD an optimal experience while improving their performance [16]. The heterogeneity in game design and outcome measures used, alongside the definite lack of studies directly comparing gamified to non-gamified approaches, highlights a critical gap in the literature. To address this gap, the current study sought to explore whether, and how, gamification should be applied to standard NF systems used by individuals with ADHD. We aimed to generate empirical insights that could inform the design of more effective and user-centered NF applications tailored to this target population.

Building on the work of Vrins et al. [17], we developed *NeuroWizard II*, a NF game designed to support attention training in individuals with ADHD using a time-modulated visually evoked potential (tVEP) BCI¹. Compared to the earlier version presented in Vrins et al. [17], *NeuroWizard II* emphasizes a more immersive form of gamification beyond surface-level elements, by embedding visual feedback directly into the game environment, incorporating a more elaborate narrative, minimizing distractions, and eliminating calibration interruptions between BCI blocks. To evaluate the effect of gamification, we compared *NeuroWizard II* to a non-gamified version of the same BCI task. Data was collected from 16 individuals formally diagnosed with ADHD, focusing on user experience, user preference, and task performance. Our research questions were formulated as follows:

RQ1: *Does gamification of NF training improve BCI task accuracy for users with ADHD?*

RQ2: *Does gamification of NF training improve interaction experience for users with ADHD?*

RQ3: *Which one do users with ADHD prefer: A gamified NF training application or a non-gamified standard version of the same training?*

Based on the existing literature [15, 17], we hypothesized that both the BCI task performance and user experience would be better in the *NeuroWizard II* game compared to the standard non-gamified task and that the game would be preferred by most users.

BCI SYSTEM DEVELOPMENT

tVEP Protocol: Both versions of the NF task were developed using the Unity game engine version 6000.0.37.f1 [22] in combination with the eight-channel Unicorn Hybrid Black EEG headset (g.tec Medical Engineering, Austria). The Unicorn Hybrid Black provides an integrated Unity plugin for BCI game development with a complete pipeline for EEG-signal preprocessing and clas-

sification, enabling the creation of immersive 2D or 3D environments controlled by EEG signals. In the present study, a tVEP-based BCI was implemented to translate neural responses into in-game commands.

While traditionally NF applications for users with ADHD rely on the targeted modulation of endogenous brain signals like sensorimotor rhythms (SMR), slow cortical potentials (SCP) or theta-beta ratios (TBR) [23], we decided to use a VEP-based design. VEPs are directly elicited by visual attention, which could make the BCI more intuitive to control while reducing ambiguity in signal interpretation [16, 21].

Game Design in Unity: Participants played as a wizarding apprentice completing a final challenge to join the council. The game comprised three BCI-based quests (see Fig. 1 for design example), themed by seasons: brewing a mushroom potion (autumn), navigating a forest via light-guided paths (winter), and fighting demons (summer). Cut-scenes between BCI blocks introduced in-game characters and objectives and advanced the narrative. Each BCI block included eight four-class selections, with up to three attempts per target guided by scene-specific cues. Target positions were pseudo-randomized, ensuring that no item appeared in the same location in consecutive selections, thereby preventing accidental double selections.

With this design, creating game-immersed visual feedback entailed the incorporation of the game elements *narrative*, *reward* (one collectible item per quest), *score* (implicitly through a progress bar for each scene), and *feedback* (see a full list of game elements commonly applied to BCIs in [15]). While feedback is inherent to NF systems, it was explicitly gamified by embedding the VEP stimuli seamlessly into the game environment and providing visual feedback through environmental changes that reflected user performance. Explicit visual feedback was provided through green or red colored rings appearing around the VEP stimuli that the user had to focus on, indicating correct or incorrect selections. Game elements were chosen to enhance motivation and engagement while minimizing the risk of distraction or stress. Hence, other commonly used elements such as *levels* or *timers* were deliberately excluded [15, 16].



Figure 1: Block 3 of the Gamified Task: Summer-Themed quest

Standard BCI Task: The standard BCI condition was

¹Note that the application developed here is referred to as a NF application following Fouillen et al. [21], though it is not a NF application in the traditional sense [7] and might also be described as a BCI-based attention-training task with NF-like feedback from a closed-loop EEG setup. Importantly, the present study is of a strictly design-focused nature and does not aim to validate a tVEP based paradigm as a self-regulation strategy or investigate the intervention's impact on ADHD symptoms.

designed to closely mirror the structure of the gamified task while omitting the immersive game elements. As in the gamified task, the standard task included three BCI blocks, each requiring participants to complete eight four-class selections, with up to three attempts allowed per selection. However, unlike the immersive design of the gamified version, stimuli here were displayed as four black circles flashing on a plain blue background using the same design for all three BCI blocks (see Figure 2). Target items were indicated by a red arrow above the target VEP stimulus. The target item order was again pseudo-randomized, with precautions implemented to prevent double selections. Between the three BCI blocks, short instructional texts were presented to guide participants through the task. No progress bar or reward system was included. The only feedback came from colored rings around the stimuli, with green rings indicating correct selections and red marking incorrect ones. Overall, both conditions used identical BCI mechanisms, differing only in presentation: the gamified version embedded the task in a narrative-based interactive game environment, whereas the standard version used a minimalistic interface. Videos showcasing both tasks can be found alongside the code for development and analysis at <https://github.com/khnhenriette/NFforADHD>.

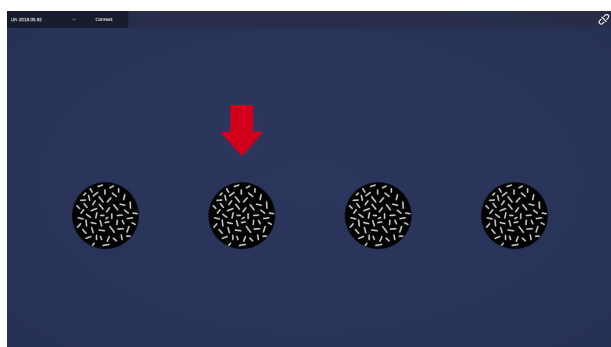


Figure 2: Design of the Standard Task

VALIDATION STUDY

Participants: The study included 17 participants living in the Netherlands or Germany with a formal ADHD diagnosis by a licensed professional. One participant was discarded due to technical issues. The final sample consisted of 16 participants between 19 and 40 years of age ($M = 25.8$; $SD = 6.2$). Nine participants self-identified as male, six as female, and one as non-binary. The study was approved by the ethics committee of Tilburg University's School of Humanities and Digital Sciences and received ethical clearance under the identification code REDC 2024.07a.

Procedure: We used a within-subjects design with randomized condition order to control for novelty, learning, and exhaustion effects. The whole experiment lasted around 60 minutes and took place in a soundproofed room, where participants were seated approximately 50

cm away from a screen adjusted to eye level. After providing informed consent and completing a pre-questionnaire, participants were fitted with an EEG cap and completed both the gamified and non-gamified BCI conditions in randomized order. Following each condition, they filled out the Game Experience Questionnaire (GEQ) [24]. After completing both conditions, participants indicated their preferred condition, provided a brief justification, and were debriefed.

Measures: BCI task accuracy was computed as the proportion of correctly selected items out of total selections, resulting in six accuracy measures across three BCI blocks in two conditions, as well as overall accuracy per condition. Differences between conditions were analyzed at both the aggregate and block level to assess potential learning or exhaustion effects. User experience was assessed using the Core and Post-game modules of the GEQ, with subscales aggregated according to Ijsselstein et al. [24]. Preference data was collected via a multiple choice selection indicating a preference for the gamified, non-gamified, both or neither conditions and open-text justifications.

Data Analysis: A Bayesian 2x3 repeated measures ANOVA was run on the factors *condition* and *block number* to investigate the possible differences in BCI task accuracy between the conditions. For all statistical tests, it was made sure the respective assumptions were met using Q-Q plots and the Shapiro-Wilk test [25]. For the analysis of the GEQ data, a Bayesian paired t-test was run, if the component-wise differences between the conditions met the normality assumptions. If not, a Bayesian Wilcoxon signed-rank test was used instead. The subjective preference indications were analyzed using a Bayesian Multinomial test, while the justifications were simply considered as supporting data and to provide guidelines for possible future research directions.

Bayesian Hypothesis Testing: Instead of traditional frequentist tests, such as t-tests or ANOVAs with p-values, this study employed their Bayesian equivalents. Bayesian hypothesis testing is based on the Bayes Factor (BF), which quantifies the relative evidence for the alternative versus the null hypothesis and allows for graded evidence rather than binary decisions [26, 27]. A BF of 1 indicates no evidence for either hypothesis, values greater than 1 favor the alternative hypothesis, whereas values below 1 favor the null hypothesis, with larger deviations from 1 reflecting stronger evidence [27].

RESULTS

BCI Task Accuracy: BCI task performance was consistently above chance level with the per-block accuracies ranging between 42.86% and 100% ($M = 74.76$; $SD = 16.25$) for the gamified condition and between 33.33% and 100% ($M = 65.85$; $SD = 13.82$) for the standard condition. The overall BCI accuracy was higher in the gamified condition than in the standard condition. The Bayesian repeated measures ANOVA (Table 1) revealed

strong evidence for an effect of condition ($BF = 16.11$), and very strong evidence for a block and condition interaction ($BF = 70.91$), while there was no main effect of BCI block number alone ($BF = 0.87$).

Table 1: Results of the Bayesian 2x3 repeated measures ANOVA examining the differences in BCI accuracy between conditions and BCI blocks. Strong evidences are indicated in bold.

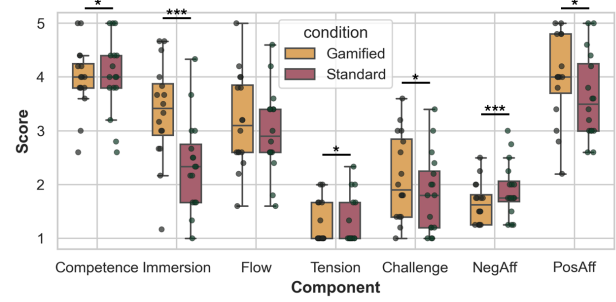
Models	BF
block	0.87
condition	16.11
block + condition	17.90
block + condition + block:condition	70.91

Post-hoc Bayesian paired t-tests (Table 2) indicated no strong evidence for a difference between conditions in block 1 ($BF = 0.28$). However, there was very strong evidence for higher accuracies in the gamified condition in block 2 ($BF = 38.19$) and moderate evidence for higher accuracies in the gamified condition in block 3 ($BF = 3.57$). Within the gamified condition, moderate evidence was found for higher accuracies in block 2 ($BF = 5.72$) and block 3 ($BF = 4.98$) compared to block 1, while no difference was found between blocks 2 and 3. In contrast, no block differences were observed within the standard condition.

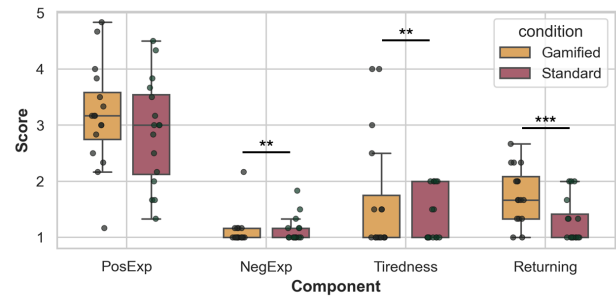
Table 2: Results of the post-hoc Bayesian paired t-tests comparing accuracies between conditions and BCI blocks. Strong and moderate evidences are indicated in bold.

Post-hoc comparisons	Relationship tested	BF
between conditions	gamified 1 > standard 1	0.28
	gamified 2 > standard 2	38.19
	gamified 3 > standard 3	3.57
between block numbers	gamified 1 < gamified 2	5.72
	gamified 1 < gamified 3	4.98
	gamified 2 < gamified 3	0.33
	standard 1 < standard 2	0.26
	standard 1 < standard 3	0.35
	standard 2 < standard 3	0.32

Game Experience Questionnaire: The GEQ Core module (Figure 3a) showed extreme evidence for higher immersion in the gamified condition ($BF = 331.27$) and strong evidence for higher negative affect in the standard condition ($BF = 23.03$). All other Core module components showed only anecdotal evidence: challenge ($BF = 1.56$) and positive affect ($BF = 1.74$) were rated slightly higher in the gamified condition, competence ($BF = 1.42$) and tension/annoyance ($BF = 1.70$) were slightly higher in the standard condition, and flow ratings showed no difference ($BF = 0.44$). In the Post-game module (Figure 3b), positive experience showed no evidence in favor of either the null or the alternative model ($BF = 1.01$). Negative experience was higher in the standard condition ($BF = 3.41$), and tiredness ($BF = 8.20$) and returning to reality ($BF = 30.37$) were higher in the gamified condition, with



(a) Core Module of GEQ



(b) Post-Game Module of GEQ

Figure 3: Boxplots of GEQ responses for each condition, overlaid with individual data points. * BF Indicates Anecdotal Evidence for a Difference. ** BF Indicates Moderate Evidence for a Difference. *** BF Indicates Strong Evidence for a Difference.

moderate and strong evidence, respectively.

Subjective User Preference: Participants showed an overwhelming preference for the gamified BCI task ($BF = 18,469.21$). 14 participants (88%) favored it, while one preferred the non-gamified version and one liked both equally. Participants frequently reported the storyline and visuals to increase enjoyment, engagement, motivation, and provide a sense of accomplishment. Opinions were mixed regarding which version better supported focus, and some participants described higher pressure or additional workload in the gamified condition.

DISCUSSION

The aim of this study was to investigate the effect of gamification on BCI-based NF training for individuals with ADHD. Although numerous NF games have been developed, they are most often validated in healthy populations, leaving their efficacy in individuals with ADHD largely unexplored [16]. This study hypothesized that gamification would enhance both BCI task performance and user experience, and that a gamified NF training application would be preferred over a standard one by the majority of participants in this cohort. The results support this hypothesis; BCI task accuracy was higher in the gamified than in the standard condition. The Bayesian ANOVA showed strong evidence for an effect of con-

dition, with post-hoc tests indicating higher accuracy in the gamified condition, particularly in blocks 2 and 3 following repeated task exposure. Prior research has reported mixed effects of gamification on BCI task performance [15, 17–19], likely depending on the specific implementation and the presence of additional in-game distractions. In the present study, the gamified BCI application was carefully designed to avoid unnecessary distractions for the participants, and the results suggest that thoughtfully implemented gamification can indeed enhance BCI task performance. Notably, the very strong block-condition interaction indicates that the benefits of gamification emerge over time, particularly as users' motivation or engagement with the task naturally declines. Within the gamified condition, BCI accuracy was higher in blocks 2 and 3 than in block 1, whereas no such progressions were observed in the standard condition. This pattern could reflect increased engagement, motivation, or learning in the gamified task, potentially counteracting detrimental effects of boredom and mind-wandering that can arise in later stages of BCI training, which is typical among individuals with ADHD [5]. Overall, these findings align with prior work highlighting motivation as a key psychological factor for BCI performance [8].

The results of the two GEQ modules were also in line with the hypothesis and largely replicate prior findings [17]. The results of the Core module mostly replicate previous trends, showing higher immersion and lower negative affect in the gamified BCI application. Compared to the older version of the *NeuroWizard* game reported by Vrins et al. [17], the current study provided stronger evidence for increased immersion and reduced negative affect in the gamified condition. In contrast, challenge showed anecdotal evidence in the opposite direction, with slightly higher values in the gamified condition, which might reflect increased perceived cognitive demand in gamified tasks [19]. Additionally, small anecdotal effects for lower perceived competence in the gamified condition were observed in the present study, which may similarly reflect slightly higher perceived difficulty in the gamified task, even though the tension/annoyance ratings were lower than in the standard condition. Given the careful gamification design including scores and rewards, but no timer or levels, it is unsurprising that evidence for tension, challenge, and competence remained overall weak. The strong increase in immersion suggests the BCI task was successfully integrated into a coherent game experience, improving interest and experiential richness consistent with the narrative-driven design [24]. Conversely, higher negative affect in the standard condition reflects higher levels of boredom, monotony, and mind-wandering as captured by the GEQ items [24], and aligns with the lower BCI accuracy observed for the standard task [5].

As no prior work has reported Post-game GEQ results within the same context, direct comparisons are not possible. Nevertheless, the higher negative experience in the standard condition is consistent with the results of

the Core module. Further, greater immersion would make disengagement harder [13, 19], possibly explaining higher difficulty in returning to reality. Higher tiredness in the gamified condition is also in line with previous studies suggesting gamified tasks can feel more demanding [19]. Task duration likely contributed as well, since the gamified version took longer to complete due to narrative cut-scenes, despite identical underlying NF components. Overall, the results of the two GEQ modules could be taken together to suggest a better user experience for the gamified version of the NF training. The components showing higher values for the gamified condition are mostly related to a better experience (e.g., immersion, positive affect), while the components exhibiting higher values for the standard condition are pointing towards a slightly worse experience (e.g., tension, negative affect, negative experience).

In BCI research, subjective experience does not always reflect task performance, although correlations can occur [28, 29]. In the present study, GEQ ratings broadly aligned with BCI accuracy, and individual preferences showed some similar patterns. Interestingly, the participant who preferred the standard condition also demonstrated higher accuracy in that condition, whereas the participant who reported equal preference for both conditions showed one of the smallest accuracy differences between the two conditions. In conclusion, the findings of this study suggest that careful gamification of NF training tasks can improve user experience and not only preserve but also enhance BCI task performance among individuals with ADHD, offering valuable design guidance for the development of future gamified NF applications.

Limitations and Future Work: A primary limitation of this study is the small, demographically restricted sample ($n = 16$), consisting mainly of university students. Although slightly larger than samples in prior work [17], it is not representative of the heterogeneous ADHD population [1], limiting generalisability. Future studies should recruit larger, more diverse samples and examine the influence of factors such as age, gender, BCI experience, medication, and gaming background. Second, the gamified scenes were presented in a fixed order, wherefore observed effects may in parts reflect scene-specific design characteristics (e.g., intuitiveness or target detectability). Randomizing game scene order would help control for such potential confounds. Third, multiple game elements were combined in the current design, making it difficult to determine which features contributed to performance improvements and which may have introduced distraction. Future work should systematically vary individual game elements while maintaining a user-centered, low-distraction design to optimize BCI performance and user experience.

Beyond this paradigm, alternative VEP-based BCI approaches such as Steady-State VEPs or code-modulated VEPs, could enable more flexible designs with more classes [30]. It would also be valuable to compare gamified and non-gamified versions of established NF proto-

cols (SCP, SMR, TBR) which not only avoid visual fatigue associated with prolonged use of VEP-based BCIs [31], but importantly, have already demonstrated clinical efficacy for ADHD [23]. Nevertheless, the simplicity and inherent reliance on user attention make visual stimuli a compelling option for future applications. As this study is strictly exploratory and design-focused, longitudinal follow-up studies should investigate the impact of VEP-based training-applications like *NeuroWizard II* on ADHD symptoms. Finally, different game themes should be explored, as user preferences varied even within this small sample, highlighting the importance of adaptable, user-centered design.

CONCLUSION

This study addresses a critical research gap in designing NF training applications for individuals with ADHD, with potential implications for therapeutic interventions. The findings show that carefully implemented gamification can enhance both BCI task performance and user experience, and that most participants prefer a gamified over a non-gamified version. Importantly, these effects were supported by both subjective measures (GEQ and user preferences) and objective accuracy metrics. Together with the identified limitations and proposed directions, the results provide valuable guidance for optimizing therapeutic NF interventions for ADHD. Moreover, the gamified application developed in this study, which is openly available on GitHub, offers a foundation for further development and evaluation of engaging, effective NF training games.

REFERENCES

- [1] Luo Y, Weibman D, Halperin JM, Li X. A review of heterogeneity in attention deficit/hyperactivity disorder (ADHD). *Frontiers in human neuroscience*. 2019;13:42.
- [2] Michele F di, Prichet L, John ER, Chabot RJ. The neurophysiology of attention-deficit/hyperactivity disorder. *International Journal of Psychophysiology*. 2005;58(1):81–93.
- [3] Zafar MB, Shah KA, Malik HA. Prospects of sustainable ADHD treatment through Brain-Computer Interface systems. In: 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT). IEEE: Karachi, Pakistan, Apr. 2017, 1–6.
- [4] Klassen LJ, Katzman MA, Chokka P. Adult adhd and its comorbidities, with a focus on bipolar disorder. *Journal of affective disorders*. 2010;124(1-2):1–8.
- [5] Cervantes JA, López S, Cervantes S, Hernández A, Duarte H. Social Robots and Brain-Computer Interface Video Games for Dealing with Attention Deficit Hyperactivity Disorder: A Systematic Review. *Brain Sciences*. 2023;13(8):1172.
- [6] Lim CG et al. A randomized controlled trial of a brain-computer interface based attention training program for ADHD. *PLOS ONE*. 2019;14(5):e0216225.
- [7] Sitaram R et al. Closed-loop brain training: The science of neurofeedback. *Nature Reviews Neuroscience*. 2017;18(2):86–100.
- [8] Kleih SC, Kübler A. Psychological factors influencing brain-computer interface (bci) performance. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics. 2015, 3192–3196.
- [9] Jeunet C, Jahanpour E, Lotte F. Why standard brain-computer interface (bci) training protocols should be changed: An experimental study. *Journal of neural engineering*. 2016;13(3):036024.
- [10] Blankertz B et al. Predicting bci performance to study bci illiteracy. *BMC Neurosci*. 2009;10(Suppl 1):P84.
- [11] Alkoby O, Abu-Rmileh A, Shriki O, Todder D. Can we predict who will respond to neurofeedback? a review of the inefficacy problem and existing predictors for successful eeg neurofeedback learning. *Neuroscience*. 2018;378:155–164.
- [12] Szewczyk RL, Ratomaska M, Jaśkiewicz M. The neglected problem of the neurofeedback learning (in) ability. In: International Scientific Conference BCI 2018 Opole. 2018, 45–58.
- [13] Gao C, Xia M, Zhang Z, Han Y, Gu Y. Improving the Brain-Computer Interface Learning Process with Gamification in Motor Imagery: A Review. In: Gamification - Analysis, Design, Development and Ludification. IntechOpen, Feb. 2023.
- [14] Castro-Cros M de et al. Effects of gamification in bci functional rehabilitation. *Frontiers in neuroscience*. 2020;14:882.
- [15] Atilla F, Postma M, Alimardani M. Gamification of motor imagery brain-computer interface training protocols: A systematic review. *Computers in Human Behavior Reports*. 2024;100508.
- [16] Kohnen H, Mavromoustakos-Blom P, Alimardani M. Is gamification of neurofeedback training effective for adhd treatment? a systematic review. In: 2025 13th International Conference on Brain-Computer Interface (BCI). 2025, 1–9.
- [17] Vrins A, Romani M, Mavromoustakos-Blom P, Alimardani M. Neurowizard: Investigating the effect of gamification on bci performance among adult users with adhd. In: International Conference on Intelligent Human Computer Interaction. 2024, 181–192.
- [18] Ali A, Puthusserypady S. A 3D learning playground for potential attention training in ADHD: A brain computer interface approach. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE: Milan, Aug. 2015, 67–70.
- [19] Vermeir JF, White MJ, Johnson D, Crombez G, Van Ryckeghem DML. The Effects of Gamification on Computerized Cognitive Training: Systematic Review and Meta-Analysis. *JMIR Serious Games*. 2020;8(3):e18644.
- [20] Schreuder M et al. User-centered design in brain-computer interfaces—a case study. *Artificial intelligence in medicine*. 2013;59(2):71–80.
- [21] Fouillen M, Maby E, Le Carrer L, Herbillon V, Mattout J. Erp-based bci training for children with ADHD: Motivations and trial design. In: GBCIC. 2017.
- [22] Unity Technologies. *Unity documentation*. Unity Technologies, 2025. [Online]. Available: <https://docs.unity3d.com/Manual/index.html>.
- [23] Enriquez-Geppert S, Smit D, Pimenta MG, Arns M. Neurofeedback as a treatment intervention in adhd: Current evidence and practice. *Current psychiatry reports*. 2019;21(6):46.
- [24] IJsselstein WA, De Kort YA, Poels K. The game experience questionnaire. 2013.
- [25] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52(3-4):591–611.
- [26] Wagenmakers EJ, Grünwald P, Steyvers M. Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*. 2006;50(2):149–166.
- [27] Wagenmakers EJ et al. Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*. 2018;25:35–57.
- [28] Rimbart S, Gayraud N, Bougrain L, Clerc M, Fleck S. Can a subjective questionnaire be used as brain-computer interface performance predictor? *Frontiers in human neuroscience*. 2019;12:529.
- [29] Vasiljevic GAM, Miranda LC de. Comparing users' performance and game experience between a competitive and collaborative brain-computer interface. *Behaviour & Information Technology*. 2024;43(1):40–59.
- [30] Bin G, Gao X, Wang Y, Hong B, Gao S. Vep-based brain-computer interfaces: Time, frequency, and code modulations [research frontier]. *IEEE Computational Intelligence Magazine*. 2009;4(4):22–26.
- [31] Xie J, Xu G, Wang J, Li M, Han C, Jia Y. Effects of mental load and fatigue on steady-state evoked potential based brain computer interface tasks: A comparison of periodic flickering and motion-reversal based visual attention. *PloS one*. 2016;11(9):e0163426.

Towards plug & play c-VEP-based BCIs: a pilot study on zero-calibration adaptive decoding using deep learning

Eduardo Santamaría-Vázquez^{1,2,3}, Víctor Martínez-Cagigal^{1,2,3}, Ana Martín-Fernández^{1,2}, and Roberto Hornero^{1,2,3}

¹Biomedical Engineering Group, University of Valladolid, Spain

²Centro de Investigación Biomedica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Valladolid, Spain

³Valladolid Health Research Institute (IBioVALL), Valladolid, Spain

E-mail: eduardo.santamaria.vazquez@uva.es

ABSTRACT: Brain-computer interfaces (BCI) based on code-modulated visual evoked potential (c-VEP) provide fast and robust control, but most high-performance systems still require user-specific calibration before use. This work evaluates a zero-calibration decoding strategy for a 16-command c-VEP speller encoded with binary *m*-sequences, aiming to reduce preparation time while preserving the possibility of online subject personalization. The proposed model is based on the original EEG-Inception implementation, with task-specific adaptations. We also developed a novel adaptive strategy that simulates online operation, in which high-confidence predictions are converted into pseudo-labels to enable conservative unsupervised adaptation of the batch-normalization layers and the classification head. We evaluated the performance of our approach on 15 subjects using a leave-one-subject-out protocol. Results showed that the zero-calibration approach without adaptation achieved a mean command accuracy of 84.79% after 10 stimulation cycles (5.25s). When two outlier subjects were excluded, this value increased to 95.91%. The adaptive strategy further improved performance, particularly during the early stimulation cycles, with a mean improvement of 5.5%. These findings support the feasibility of subject-independent c-VEP decoding using deep learning and suggest that conservative test-time adaptation can improve robustness in practical plug-and-play spelling scenarios.

INTRODUCTION

Brain-computer interfaces (BCIs) based on code-modulated visual evoked potentials (c-VEP) are among the most accurate and fastest non-invasive paradigms for communication and control [1]. In these systems, each command is encoded by a pseudo-random visual sequence, and the attended target is inferred from the corresponding electroencephalographic (EEG) response. In comparison with more classical evoked-potential paradigms, c-VEPs provide a favorable compromise between number of commands, decoding speed, and robustness, which has enabled spelling applications with very high information transfer rates and reliable operation in both healthy and motor-impaired users [2–5].

However, despite these advances, most high-performance c-VEP systems still depend on subject-specific calibration sessions to estimate user-adapted templates or train subject-dependent decoders [1, 6].

This calibration requirement remains one of the main barriers to the practical deployment of c-VEP-based BCIs. From a translational perspective, the need to record labeled data before each use increases setup time, reduces usability, and limits the development of plug-and-play systems. Moreover, calibration is especially problematic in applications where rapid access, repeated sessions, or user fatigue are relevant concerns. Consequently, reducing the amount of subject-specific data required by c-VEP decoders has become an important research objective in recent years [6–9]. Existing strategies include transfer learning across subjects, multi-session training, semi-supervised updates, and fully unsupervised calibration procedures [6, 10–12]. Although these studies have shown that calibration can be substantially reduced, achieving robust subject-independent performance in unseen users remains challenging because of the marked inter-subject variability of VEP responses.

In parallel with these advances, alternative decoding frameworks have expanded the methodological landscape of c-VEP BCIs. Besides classical template matching, event-wise response modeling approaches have shown that the stimulation pattern can be reconstructed directly from short EEG epochs and later recombined to recover the attended command [13]. This bit-wise reconstruction (BWR) approach decouples event classification from the commands codebook, facilitating transfer across stimulation schemes, and naturally lending itself to modern deep-learning architectures. For instance, our previous work showed that EEG-Inception can accurately reconstruct binary and non-binary stimulation events from EEG under subject-specific calibration, offering a competitive alternative to conventional c-VEP decoding pipelines [14]. In principle, this framework is also well suited to zero-calibration scenarios, as models can be trained across subjects and transferred directly to unseen users. However, studies exploring this possibility are still scarce, and the performance reported so far remains well below that obtained with calibration [15].

The goal of this study is twofold: (1) To evaluate the operational feasibility of a zero-calibration c-VEP speller driven by a cross-subject deep-learning model; and (2) To design, develop, and validate a novel adaptive strategy capable of enhancing decoding performance in real-time by progressively integrating new online data through an unsupervised mechanism. To this end, we combine a leave-one-subject-out (LOSO) training strategy with conservative unsupervised test-time adaptation driven by high-confidence pseudo-labels, enabling the model to be applied directly to unseen users and updated online without supervised calibration. Beyond overall spelling performance, the study also examines inter-subject variability and the influence of outlier users, with the aim of better characterizing the practical limitations of zero-calibration c-VEP decoding.

MATERIALS AND METHODS

Dataset and experimental setting:

The experiments were conducted using the open-source non-binary c-VEP database presented by Martinez-Cagigal *et al.* [16]. This dataset includes several stimulation conditions based on p -ary maximal-length sequences and a 4×4 speller with 16 commands (letters A–P). In this work, only the binary condition was considered, corresponding to the black–white stimulation setting. Although the original database included 16 healthy participants (28.8 ± 5.0 years), the analyses reported here were conducted on 15 subjects, as one participant was excluded due to the use of a different sampling rate during the experiments. Each subject completed 30 calibration trials of 10 stimulation cycles each with an m-sequence of 63 bits and a monitor refresh rate of 120 Hz. The online spelling phase comprised 32 trials per subject, with 10 stimulation cycles per trial. Specifically, participants were tasked with selecting each of the 16 commands in alphabetical order across two full iterations [16]. EEG was recorded from 16 electrodes placed according to the international 10–10 system: F3, Fz, F4, C3, Cz, C4, CPz, P3, Pz, P4, PO7, POz, PO8, Oz, I1, and I2. The ground electrode was located at AFz and the reference at the earlobe. All experiments were performed with the open source "P-ary c-VEP Speller" application of MEDUSA [17]. More information about the experimental protocol can be found in [18].

Preprocessing and feature extraction:

The preprocessing parameters are described in our previous work [14]. EEG signals were filtered using a 7th-order Butterworth band-pass filter between 1 and 60 Hz and a 7th-order notch filter between 49 and 51 Hz. Signal epochs were extracted from 0 to 500 ms after each stimulation event and resampled to 200 Hz, yielding 100 temporal samples per observation. Feature extraction followed a bit-wise reconstruction approach, where each stimulation event was segmented into an EEG epoch and labeled according to its corresponding binary stimulation value. Consequently, each stimulation cycle yielded 63

epochs. In total, the calibration and test datasets contained 18,900 ($63 \times 10 \times 30$) and 20,160 ($63 \times 10 \times 32$) observations per subject, respectively. The pipeline used z-normalization within the analysis window and generated event-wise samples that were later recombined at decoding time to recover the attended command.

Deep learning model:

The classifier was a modified EEG-Inception architecture specifically adapted for bit-wise c-VEP decoding [19]. The backbone started with a temporal inception block composed of three parallel convolutional branches with kernel sizes of 50, 25, and 15 samples, respectively, each containing 12 filters to capture temporal patterns at different scales. This stage was followed by a depthwise spatial filtering block with a multiplier of 2, which learned channel combinations while collapsing the spatial dimension. A second temporal stage then applied a dilated inception block with four parallel branches, all using kernel size 5 but with dilation rates of 1, 5, 10, and 15, respectively, again with 12 filters per branch. The resulting feature maps were progressively compressed by the output block, and finally mapped to the binary classification head. Batch normalization was applied after every convolutional layer throughout the network to stabilize internal activations and accelerate convergence. The network used exponential linear unit (ELU) activations and standard dropout with rate 0.2. Overall, this design preserves the main rationale of EEG-Inception [19] (i.e., multi-scale temporal feature extraction, efficient spatial filtering, and progressive feature compression) while adapting it to the subject-independent, event-wise decoding problem considered here.

Zero-calibration approach:

The proposed framework follows a strictly subject-independent training strategy based on leave-one-subject-out (LOSO) cross-validation. For each fold, one participant was held out for testing and the model was trained exclusively with the preprocessed calibration data from the remaining subjects (i.e., 30 trials/subject). Therefore, no labeled EEG data from the test user were available during model training. Consequently, the evaluated system simulates a zero-calibration approach from the user perspective, since decoding starts without any supervised subject-specific adjustment. The training stage for each LOSO fold used a learning rate of 0.01, batch size of 256, validation split of 0.1, early stopping with patience of 5 epochs, and a maximum of 100 epochs.

During testing, the network estimated the binary label of each stimulation event independently [14]. These event-wise predictions were then recombined to reconstruct the stimulation sequence associated with each trial. Command decoding was finally performed by comparing the reconstructed sequence with the 16 binary m-sequences assigned to the speller commands and selecting the command with maximum Pearson correlation. This procedure produced a command estimate after each stimulation cycle, thus allowing performance to be analyzed as a function of stimulation duration.

Unsupervised online adaptation:

In addition to the zero-calibration baseline, we implemented an unsupervised adaptive strategy that simulates progressive online personalization. Starting from the cross-subject model, adaptation was applied trial by trial in causal order. After each prediction, the confidence margin, defined as the difference between the highest and second-highest command-correlation scores at the last available stimulation cycle, was computed. When this margin exceeded a threshold of 0.1, the trial was deemed reliable and used to update the model via pseudo-labeling: the predicted command was mapped back to its event-wise binary sequence, and the corresponding EEG epochs were added to an adaptation buffer.

Model updating was deliberately conservative. Only the batch-normalization layers and the classification head were fine-tuned, while the convolutional backbone remained frozen, thereby mitigating overfitting and catastrophic forgetting while allowing the model to recalibrate its feature normalization and decision boundaries to the new user. Adaptation used a learning rate of 0.001, batch size of 256, validation split of 0.2, early stopping with patience of 3 epochs, minimum improvement of 0.001, and a maximum of 10 epochs per update

RESULTS

We compared our zero-calibration approach with and without adaptation. Performance was assessed in terms of command accuracy as a function of stimulation cycle, which provides a direct measure of the speller's speed-accuracy trade-off.

Table 1 summarizes the results obtained with the subject-independent model without adaptation. Mean accuracy increased progressively from 52.71% at cycle 1 to 84.79% at cycle 10. However, performance was strongly affected by two outlier subjects (S10 and S11), whose accuracies remained close to chance level across the whole session. When these subjects were excluded, mean accuracy increased from 59.38% at cycle 1 to 95.91% at cycle 10, with a clear reduction in inter-subject variability.

Table 2 shows the results obtained with the unsupervised adaptive strategy. In this case, mean accuracy improved from 63.12% at cycle 1 to 87.50% at cycle 10 across all subjects. After excluding the two outliers, performance increased from 72.36% to 98.08%. Overall, the adaptive approach yielded consistently higher accuracies than the non-adaptive baseline across all stimulation cycles.

The magnitude of this improvement is displayed in Figure 1. Across all subjects, the adaptive strategy provided a mean gain of 10.42% at cycle 1, with the benefit gradually decreasing as more stimulation cycles were accumulated, reaching 2.71% at cycle 10. On average, the adaptive strategy improved accuracy by 5.50% across all cycles.

Analysis of outlier subjects:

Given the marked inter-subject variability observed in our experiments, a complementary analysis was conducted to

better characterize the two outlier subjects (S10 and S11). Interestingly, these subjects did not show atypical performance in our previous within-subject frameworks, where their accuracies fell within the normal range [14, 20]. In contrast, in the present cross-subject setting they exhibited substantially lower accuracies than the rest of the cohort under both the baseline and adaptive conditions, remaining close to chance level even after 10 stimulation cycles. Their inclusion therefore had a pronounced effect on the overall mean performance and variability estimates, warranting a more detailed inspection of their electrophysiological patterns.

As shown in Tables 1 and 2, the baseline model reached only 18.75% and 6.25% accuracy at cycle 10 for S10 and S11, respectively. With adaptation, S10 improved modestly to 31.25%, whereas S11 remained at 6.25%. This indicates that the adaptive strategy was not sufficient to recover normal performance in these cases, although it provided a limited benefit for one of the two subjects.

To further explore this finding, Fig. 2 compares the average evoked-response template at POZ and the spatial distribution of template root mean square (RMS) between outlier and non-outlier subjects. The non-outlier group show clearer evoked-response morphology and a more marked posterior RMS maximum over occipital electrodes, whereas the outlier group tends to exhibit lower amplitudes and less distinctive spatial pattern. Although these observations should be interpreted with caution, they are compatible with the possibility that the outlier subjects presented less typical c-VEP responses and electrophysiological profiles, which may have limited the effectiveness of cross-subject transfer.

DISCUSSION

This study evaluated a zero-calibration binary c-VEP decoding framework based on subject-independent deep learning and assessed whether a novel unsupervised online adaptation strategy could improve performance during use. Overall, the approach was feasible for most subjects, with the baseline model reaching a mean accuracy of 84.79% after 10 stimulation cycles and the adaptive strategy increasing this value to 87.50%. When the two outlier subjects were excluded, performance reached 95.91% and 98.08%, respectively. The main benefit of adaptation was observed during the early stimulation cycles, with mean gains of 10.42% and 8.12% at cycles 1 and 2, indicating that its primary effect was to accelerate convergence toward a correct decision rather than to substantially improve final performance. Taken together, these results suggest that our decoding strategy may provide a useful basis for calibration-free c-VEP spelling, while also highlighting the presence of inter-subject variability and the influence of a small number of difficult cases.

A useful reference for contextualizing these results is our previous within-subject study on the same database [14]. In that calibrated setting, the binary condition achieved

Table 1: Accuracy per cycle (%) — Zero-Calibration (no adaptation)

Cycle	1	2	3	4	5	6	7	8	9	10
Stim. dur. (s)	0.53	1.05	1.58	2.10	2.62	3.15	3.68	4.20	4.73	5.25
S0	46.88	62.50	71.88	78.12	81.25	81.25	81.25	84.38	84.38	87.50
S1	56.25	75.00	81.25	87.50	96.88	96.88	100.00	100.00	100.00	100.00
S2	87.50	90.62	93.75	96.88	100.00	100.00	100.00	100.00	100.00	100.00
S3	68.75	87.50	93.75	93.75	96.88	96.88	96.88	100.00	100.00	96.88
S4	37.50	43.75	56.25	53.12	59.38	71.88	71.88	81.25	84.38	84.38
S5	62.50	81.25	81.25	84.38	90.62	90.62	93.75	93.75	100.00	100.00
S6	71.88	84.38	87.50	90.62	90.62	90.62	90.62	93.75	93.75	90.62
S7	59.38	65.62	75.00	81.25	87.50	87.50	96.88	96.88	100.00	93.75
S8	71.88	90.62	93.75	96.88	96.88	96.88	96.88	96.88	96.88	96.88
S9	62.50	81.25	93.75	100.00	100.00	100.00	100.00	100.00	100.00	100.00
S10 [†]	12.50	12.50	12.50	18.75	18.75	15.62	12.50	15.62	18.75	18.75
S11 [†]	6.25	9.38	9.38	6.25	9.38	12.50	6.25	9.38	6.25	6.25
S12	71.88	90.62	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
S13	28.12	40.62	53.12	59.38	65.62	65.62	71.88	75.00	87.50	96.88
S14	46.88	56.25	78.12	84.38	87.50	93.75	96.88	96.88	100.00	100.00
Mean (all)	52.71	64.79	72.08	75.42	78.75	80.00	81.04	82.92	84.79	84.79
Std (all)	23.14	27.33	28.28	29.02	28.96	28.66	30.65	29.62	30.02	29.85
Mean (w/o outliers)	59.38	73.08	81.49	85.10	88.70	90.14	92.07	93.75	95.91	95.91
Std (w/o outliers)	16.19	17.70	14.63	14.67	13.04	11.04	10.33	8.27	6.30	5.32

[†]Outlier subjects (S10, S11).

Table 2: Accuracy per cycle (%) — Zero-Calibration with adaptive strategy

Cycle	1	2	3	4	5	6	7	8	9	10
Stim. dur. (s)	0.53	1.05	1.58	2.10	2.62	3.15	3.68	4.20	4.73	5.25
S0	65.62	87.50	87.50	90.62	96.88	93.75	96.88	96.88	96.88	96.88
S1	71.88	87.50	87.50	90.62	96.88	100.00	100.00	100.00	100.00	100.00
S2	96.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
S3	87.50	93.75	96.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00
S4	40.62	46.88	68.75	81.25	78.12	81.25	87.50	93.75	96.88	96.88
S5	78.12	87.50	90.62	93.75	96.88	96.88	96.88	96.88	100.00	100.00
S6	87.50	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
S7	75.00	81.25	90.62	90.62	93.75	96.88	96.88	96.88	100.00	100.00
S8	93.75	96.88	96.88	96.88	96.88	96.88	96.88	96.88	96.88	96.88
S9	71.88	84.38	87.50	100.00	100.00	100.00	100.00	96.88	100.00	100.00
S10 [†]	3.12	6.25	18.75	15.62	12.50	12.50	25.00	28.12	25.00	31.25
S11 [†]	3.12	3.12	15.62	9.38	9.38	6.25	12.50	9.38	6.25	6.25
S12	84.38	96.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
S13	25.00	34.38	46.88	65.62	78.12	81.25	78.12	78.12	81.25	84.38
S14	62.50	87.50	96.88	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Mean (all)	63.12	72.92	78.96	82.29	83.96	84.38	86.04	86.25	86.88	87.50
Std (all)	30.94	33.36	28.73	29.85	30.52	31.12	28.07	28.17	29.53	28.59
Mean (w/o outliers)	72.36	83.41	88.46	93.03	95.19	95.91	96.39	96.63	97.84	98.08
Std (w/o outliers)	20.60	20.07	15.12	10.06	7.82	6.80	6.49	5.91	5.16	4.33

[†]Outlier subjects (S10, S11).

99.8% accuracy after 10 cycles, whereas the adaptive zero-calibration framework evaluated here reached 87.50%, increasing to 98.08% when the two outlier subjects were excluded. Thus, although the adaptive subject-independent approach did not reach the performance of the calibrated within-subject framework in the full cohort, it approached that reference level for most subjects while eliminating the calibration phase.

Two subjects showed markedly lower accuracies than the rest of the cohort despite not being atypical in our previous within-subject framework [14, 20], suggesting reduced cross-subject transfer in these cases. Although their descriptive characterization was inconclusive, some features pointed to a potentially different neurophysiological profile and less reliable c-VEP response. The limited effect of adaptation in these subjects is also consistent with the fact that this strategy depends on sufficiently reliable initial predictions. Overall, these results suggest

that conservative unsupervised adaptation may be useful as a refinement mechanism, but not always sufficient to compensate for poor initial transfer.

Relative to the state of the art, the present work contributes to the ongoing effort to reduce calibration requirements in c-VEP BCIs. Previous studies, particularly by Thielen *et al.* [12], have shown that calibration-free or zero-training c-VEP decoding is feasible using reconvolution and canonical correlation analysis (CCA) formulations. Other studies have also reported adaptive or training-reduction strategies in c-VEP systems, highlighting both the potential and the difficulty of maintaining robust performance when subject-specific calibration is limited [7, 10]. Although direct comparisons across studies are difficult due to differences in datasets and decoding approaches, the present work presents, to the best of our knowledge, the first unified framework combining subject-independent deep-learning-based c-VEP de-

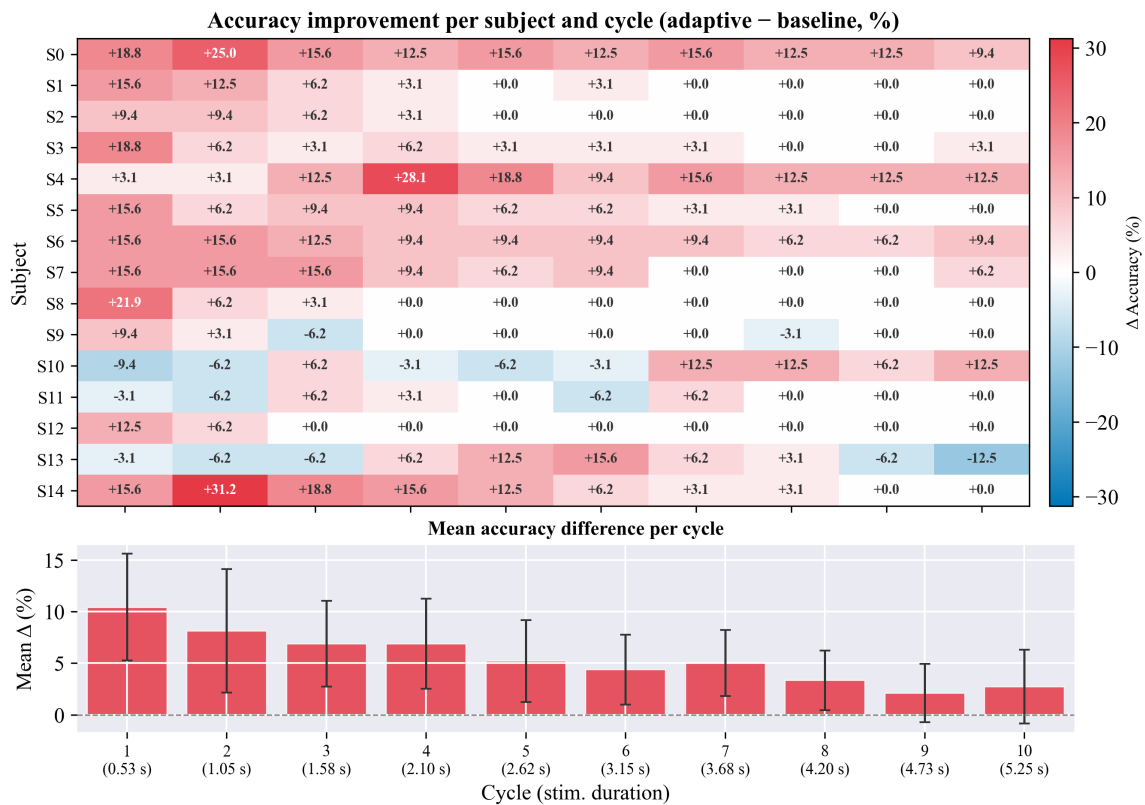


Figure 1: Accuracy improvement per subject and stimulation cycle for the adaptive strategy relative to the zero-calibration baseline. The upper panel shows the subject-wise difference in command accuracy (adaptive – baseline, percentage points) across stimulation cycles, whereas the lower panel shows the corresponding mean improvement across subjects with standard deviation.

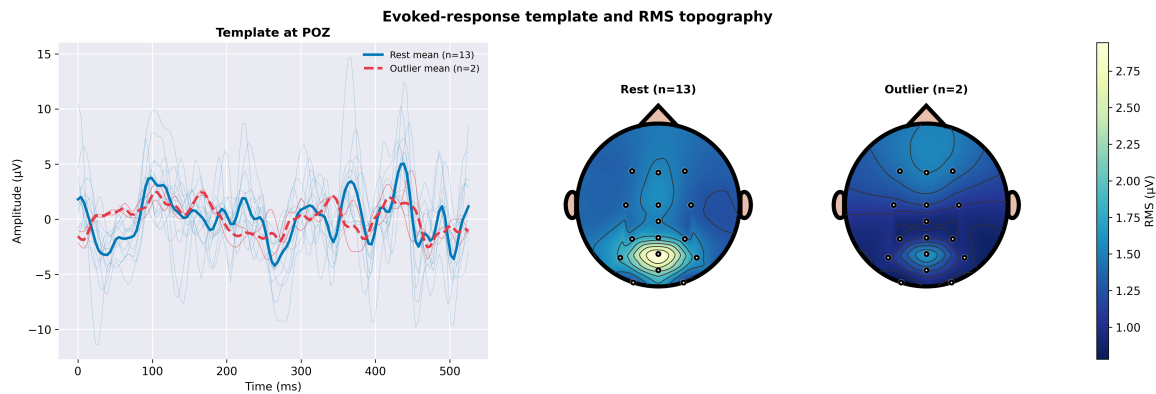


Figure 2: Evoked-response template and RMS topography averaged across regular and outlier subjects. The left panel shows the c-VEP template waveform at electrode POz for each subject (thin lines) together with the group mean (thick lines). The right panels display the mean root mean square (RMS) scalp topography for regular and outlier subjects, with a shared colour scale.

coding with an unsupervised online adaptation strategy.

Several limitations should be considered. The subject sample size was modest, which limits the interpretation of the outlier analyses, and the observed inter-subject variability may partly reflect the limited amount and diversity of data available for subject-independent training. Crucially, the current approach has yet to be validated within populations with severe motor disabilities, such as individuals with spinal cord injury or amyotrophic lateral sclerosis, whose neurophysiological profiles may differ

significantly from the current sample. In this regard, the proposed processing pipeline is expected to benefit from larger and more heterogeneous datasets, which should improve cross-subject robustness and reduce the impact of underrepresented response profiles. In addition, the adaptive strategy was deliberately conservative, which favors stability but may also limit personalization in difficult cases. Future work should therefore explore more flexible adaptation schemes.

CONCLUSION

This paper presents the first zero-calibration deep-learning framework for c-VEP spellers based on subject-independent training and unsupervised online adaptation. The proposed approach avoids supervised fine-tuning on the test subject and achieves performance levels that, for most users, approach those reported in conventional calibrated systems, thereby substantially reducing setup time while preserving high-speed communication capabilities. Nevertheless, a small subset of users remained challenging in the present study, highlighting the need for further work to improve robustness and personalization for underrepresented neurophysiological profiles.

ACKNOWLEDGEMENTS

This research was supported by Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), through Instituto de Salud Carlos III, co-funded by the European Regional Development Fund (ERDF); the Regional Government of Castilla y León (Junta de Castilla y León, Consejería de Educación) and the EU-FEDER under the project VA140P24; and the European Union through the Interreg VI-A Spain-Portugal Program (POCTEP) 2021-2027 project "0124_EUROAGE_MAS_4_E". A. Martín-Fernández was supported by a PIF grant from UVA.

REFERENCES

- [1] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): a literature review. *Journal of Neural Engineering*. 2021;18(6):061002.
- [2] Gembler FW, Rezeika A, Benda M, Volosyak I. Five Shades of Grey: Exploring Quintary m-Sequences for More User-Friendly c-VEP-Based BCIs. *Computational Intelligence and Neuroscience*. 2020;2020.
- [3] Sun Q, Zheng L, Pei W, Gao X, Wang Y. A 120-target brain-computer interface based on code-modulated visual evoked potentials. *Journal of Neuroscience Methods*. 2022;375(109597).
- [4] Moreno-Calderón S, Martínez-Cagigal V, Santamaría-Vázquez E, Pérez-Velasco S, Marcos-Martínez D, Hornero R. Combining brain-computer interfaces and multiplayer video games: an application based on c-VEPs. *Frontiers in Human Neuroscience*. 2023;17.
- [5] Verbaarschot C *et al.* A visual brain-computer interface as communication aid for patients with amyotrophic lateral sclerosis. *Clinical Neurophysiology*. 2021;132(10):2404–2415.
- [6] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. *Journal of Neural Engineering*. 2021;18(5):56007.
- [7] Stawicki P, Volosyak I. cVEP Training Data Validation—Towards Optimal Training Set Composition from Multi-Day Data. *Brain Sciences*. 2022;12(2).
- [8] Miao Y *et al.* High-performance c-VEP-BCI under minimal calibration. *Expert Systems with Applications*. 2024;249(PB):123679.
- [9] Zheng L, Dong Y, Tian S, Pei W, Gao X, Wang Y. A calibration-free c-VEP based BCI employing narrow-band random sequences. *Journal of Neural Engineering*. 2024;21(2).
- [10] Spüler M, Rosenstiel W, Bogdan M. Unsupervised online calibration of a c-vep brain-computer interface (bci). In: *Artificial Neural Networks and Machine Learning—ICANN 2013: 23rd International Conference on Artificial Neural Networks Sofia, Bulgaria, September 10-13, 2013. Proceedings 23. 2013*, 224–231.
- [11] Turi F, Gayraud NT, Clerc M. Auto-calibration of c-vep bci by word prediction. 2020.
- [12] Thielen J, Sosulski J, Tangermann M. Exploring new territory: Calibration-free decoding for c-vep bci. *arXiv preprint arXiv:2403.15521*. 2024.
- [13] Nagel S, Dreher W, Rosenstiel W, Spüler M. The effect of monitor raster latency on VEPs, ERPs and Brain-Computer Interface performance. *Journal of Neuroscience Methods*. 2018;295:45–50.
- [14] Santamaría-Vázquez E, Martínez-Cagigal V, Hornero R. Bit-wise reconstruction of non-binary visual stimulation patterns from eeg using deep learning: A promising alternative for user-friendly high-speed c-vep-based bcis. In: *International Work-Conference on Artificial Neural Networks*. 2023, 603–614.
- [15] Behboodi M, Kinney-Lang E, Etemad A, Kirton A, Abou-Zeid H. Leveraging foundation models for calibration-free c-vep bcis. *IEEE*, Oct. 2025, 4564–4571.
- [16] Martínez-Cagigal V. Dataset: Non-binary m-sequences for more comfortable brain-computer interfaces based on c-veps. 2025.
- [17] Santamaría-Vázquez E *et al.* MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. *Computer Methods and Programs in Biomedicine*. 2023;230(107357).
- [18] Martínez-Cagigal V, Santamaría-Vázquez E, Pérez-Velasco S, Marcos-Martínez D, Moreno-Calderón S, Hornero R. Non-binary m-sequences for more comfortable brain-computer interfaces based on c-VEPs. *Expert Systems with Applications*. 2023;232(120815).
- [19] Santamaría-Vázquez E, Martínez-Cagigal V, Vaquerizo-Villar F, Hornero R. Eeg-inception: A novel deep convolutional neural network for assistive erp-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2020;28:2773–2782.
- [20] Martínez-Cagigal V, Santamaría-Vázquez E, Pérez-Velasco S, Martín-Fernández A, Hornero R. Reevaluating performance in c-vep bcis: The impact of calibration time. *Biocybernetics and Biomedical Engineering*. 2025;45:685–696.

Adaptive Classification of EEG Signals for Bimanual Motor Imagery using Riemannian Geometry

A.E. Cortés-Téllez¹, J.R. López-Gutiérrez¹, A.A. Torres-García¹

¹Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México

E-mail: ahiram.cortest@inaoep.mx

ABSTRACT: Brain-computer interfaces (BCIs) experience severe performance degradation when transitioning from unilateral to bimanual decoding due to simultaneous bihemispheric interference. To address this, we developed the Adaptive Topological BCI (AT-BCI) architecture. By embedding temporal dynamics into Hankel matrices and applying Riemannian geometry to align neural distributions, the model enables zero-shot transfer without target-domain calibration. Treating the covariance shift as a purely geometric discrepancy isolates mixed signals and prevents the expansion of peripheral noise. We evaluated AT-BCI on the PhysioNet dataset, projecting unilateral motor imagery onto bimanual tasks. After applying a strict amplitude filter to exclude electromyographic artifacts and physiological desynchronization, evaluation on a validated cohort of 10 subjects yielded a mean F1-score of 0.7468 and a Cohen's Kappa of 0.4901. These results demonstrate that restricting spatial topology and enforcing geodesic alignment stabilizes bimanual control, advancing the clinical viability of neuroprosthetic interfaces without the burden of daily recalibration.

INTRODUCTION

Motor-imagery-based brain-computer interfaces (BCIs) traditionally rely on unimanual tasks to decode user intent [1, 2]. Standard classification models achieve high accuracy in controlled single-limb paradigms, but practical daily activities inherently demand bimanual coordination. Transitioning a BCI from a unimanual calibration phase to a bimanual execution environment introduces severe non-stationarities in the electroencephalographic (EEG) signals. This transition abruptly degrades classifier performance [3], presenting a major bottleneck for functional neurorehabilitation. The root of this degradation lies in the neurophysiology of simultaneous movement. The execution of a bimanual motor task induces a complex, bilateral activation across the sensorimotor network. This dual activation generates an interhemispheric masking effect, which disrupts the localized event-related desynchronization (ERD) patterns typically observed during isolated unimanual tasks [4, 5]. Con-

sequently, classical spatial filters, such as Common Spatial Patterns (CSP) [6], are compromised because the bilateral interference violates the underlying assumption of spatial separability between classes.

Mathematically, this discrepancy between the unimanual training domain and the bimanual testing domain constitutes a severe domain shift [7]. To restore decoding reliability without burdening the user with exhaustive bimanual retraining sessions, adaptive transfer learning techniques are required [8]. Building upon the augmented covariance matrix (ACM) framework introduced by Carrara et al. [9], which demonstrated that embedding delayed EEG signals into higher-dimensional spaces captures non-stationary spatial patterns, we propose to resolve this displacement. While generic recentring techniques attempt to align data distributions [10], they typically process broad frequency bands and fail to account for the temporal dynamics of interhemispheric communication, leading to sub-optimal adaptations. To overcome these limitations, we propose the Adaptive Topological BCI (AT-BCI) architecture. We hypothesize that by restricting the spatial topology strictly to the sensorimotor cortex, embedding temporal dynamics through an analytically derived delay, and resolving the domain shift via geometric alignment, we can neutralize bimanual interference [8, 11].

The main contributions of this work are:

- We introduce a pipeline that treats bimanual covariance shifts as an unsupervised geometric alignment problem on a Riemannian manifold, bypassing the need for target-domain calibration.
- We incorporate temporal embeddings derived analytically from the central frequency of the signal to capture the dynamic nature of bilateral motor intentions, avoiding exhaustive heuristic searches and preventing temporal overfitting.
- We provide an *in silico* evaluation across a validated cohort of 10 subjects from the PhysioNet dataset [12, 13], utilizing strict amplitude filters to discard severe electromyographic artifacts. The results indicate that AT-BCI miti-

gates performance drops during zero-shot transfer tasks.

The remainder of this paper is structured as follows: Section II details the dataset audit, the topological restriction, and the proposed unsupervised domain alignment. Section III presents the population classification results and visualizes the spatial projections. Section IV examines the physical implications of the findings, and Section V concludes the paper.

MATERIALS AND METHODS

The complete processing pipeline of the proposed Adaptive Topological BCI (AT-BCI) architecture is illustrated in Fig. 1. The framework is logically structured into four sequential phases: spatial and artifact restriction, temporal embedding, geometric domain alignment, and isolated parametric optimization.

Data Management and Artifact Filtering: Empirical evaluation relied on electroencephalography records from the public PhysioNet dataset [12, 13]. Zero-shot transfer depends mathematically on computing an accurate Riemannian mean from the unimanual calibration session. The geodesic distance metric exhibits high sensitivity to spatial outliers. Consequently, high-amplitude artifacts shift the centroid of the reference matrix, irreversibly corrupting the subsequent domain alignment.

To neutralize this operational vulnerability, the algorithm integrates a cleaning protocol based on peak-to-peak amplitude. Before estimating covariances, the system systematically purges epochs exceeding a $100\mu\text{V}$ threshold. Applying this filter excluded records with massive electromyographic contamination or biological desynchronization deficiency, consolidating a valid cohort of 10 subjects. Concurrently, the spatial topology was reduced to 15 channels concentrated across the sensorimotor strip (FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4). This dimensional reduction ensures that the Riemannian alignment operates on features of strictly cortical origin, preventing the model from expanding parasitic noise from prefrontal and occipital zones during temporal processing.

Spatiotemporal Embedding via Augmented Covariance: To structure the temporal dynamics of the signal, the architecture employs the augmented covariance matrix (ACM) framework proposed by Carrara and Papadopoulou [9]. The procedure concatenates the original data with time-delayed versions of itself, forcing the cross-covariance to encode temporal dependencies across different brain regions.

Prior implementations of this model located the optimal delay using grid search methods. Analysis of the logs indicated that this practice induces data leakage and causes sample overfitting. Evaluating the mathematical relationship between the time domain

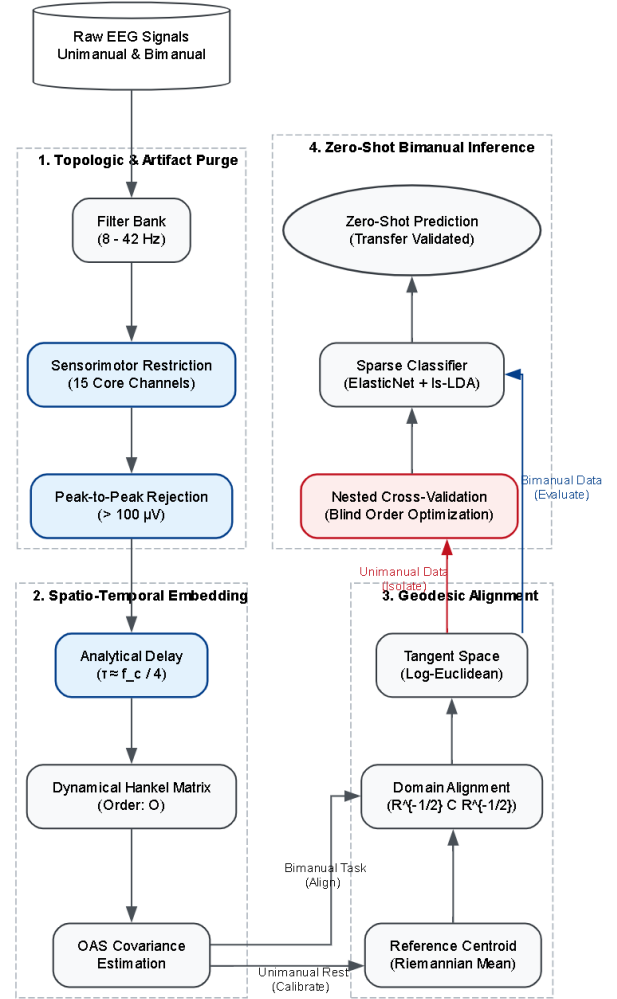


Figure 1: Flowchart of the AT-BCI architecture. The framework integrates a strict $100\mu\text{V}$ amplitude threshold and a 15-channel spatial restriction to prevent noise expansion. Temporal embedding relies on an analytically derived delay ($\tau \approx f_c/4$), while a nested cross-validation loop isolates the unimanual calibration from the bimanual zero-shot inference, eliminating data leakage.

and frequency coherence shows that capturing cross-dependencies requires an independent delay for each spectral band under evaluation. Therefore, the system abandons empirical exploration to compute the delay analytically. A quarter-wavelength phase shift derived from the central frequency (f_c) of each filter was applied using the equation:

$$\tau \approx \frac{f_c}{4} \quad (1)$$

This calculation guarantees that the matrix captures genuine hemispheric interactions and respects the physical dynamics of sensorimotor oscillations without forcing data convergence.

Unsupervised Domain Alignment and Nested Validation: The distribution discrepancy between the unimanual and bimanual domains was treated purely as a geometric problem. After computing the centroid of the unimanual rest data, geodesic alignment

was executed via the transformation:

$$C_{\text{aligned}} = R^{-1/2} C R^{-1/2} \quad (2)$$

where R acts as the reference matrix and C contains the task motor covariances [8]. The resulting matrices were projected onto the tangent space through a Log-Euclidean mapping to enable linear classification [2, 14].

To define the depth of the Hankel matrices without corrupting the model, a strict nested cross-validation was instrumented. The algorithm optimizes the autoregressive order by evaluating accuracy solely on the inner folds of the unimanual calibration data. Blocking access to the bimanual test set eliminates overfitting risks. This constraint ensures a genuine zero-shot transfer sustained by stable physiological characteristics, satisfying rigorous evaluation standards [6].

Algorithm 1 formalizes the computational sequence of the proposed methodology, explicitly detailing the integration of the analytical delay and the geodesic transformation prior to the classification phase.

Algorithm	1	Adaptive	Topo-
logical	BCI	(AT-BCI)	Execution

Require: Unimanual data $\mathcal{D}_U$, Bimanual data $\mathcal{D}_B$,
Filter central frequencies f_c

Ensure: Predicted bimanual intention labels $\hat{y}_B$

- 1: **Phase 1: Preprocessing & Restriction**
- 2: Restrict topology to 15 core sensorimotor channels.
- 3: Reject epochs with amplitude $> 100\mu\text{V}$.
- 4: **Phase 2: Embedding & Domain Alignment**
- 5: $O^* \leftarrow \text{OptimizeOrder}(\mathcal{D}_U)$ {Nested CV isolation}
- 6: $\tau \leftarrow f_c/4$ {Analytical phase shift calculation}
- 7: $\mathbf{C}_U, \mathbf{C}_B \leftarrow \text{ConstructHankel\&OAS}(\mathcal{D}_U, \mathcal{D}_B, O^*, \tau)$
- 8: $\mathbf{R} \leftarrow \text{RiemannianMean}(\mathbf{C}_{\text{rest}} \in \mathcal{D}_U)$ {Reference Centroid}
- 9: **Phase 3: Tangent Space Projection & Training**
- 10: $\mathbf{S}_U, \mathbf{S}_B \leftarrow \text{Log-Euclidean}(\mathbf{R}^{-1/2}\{\mathbf{C}_U, \mathbf{C}_B\}\mathbf{R}^{-1/2})$
- 11: $\text{Model} \leftarrow \text{TrainClassifier}(\mathbf{S}_U, \mathbf{y}_U)$ {ElasticNet + ls-LDA}
- 12: **Phase 4: Zero-Shot Bimanual Inference**
- 13: **return** $\hat{y}_B \leftarrow \text{Predict}(\text{Model}, \mathbf{S}_B)$ {No target recalibration}

RESULTS

Viability Audit and Global Performance: The original design targeted an initial cohort of 20 subjects.

However, to guarantee that decoding in the bimanual domain originated strictly from cortical activity and not electromyographic (EMG) contamination, a strict amplitude rejection filter ($> 100\mu\text{V}$) was applied. This audit revealed that several participants presented a massive loss of viable epochs due to constant muscular tension during the simultaneous motor task. After excluding individuals with structural signal deficiency, the analysis was consolidated on a valid cohort of 10 subjects.

On this physiologically rigorous group, the AT-BCI architecture achieved solid metrics. As the boxplots in Fig. 2 illustrate, the performance distribution shows the system projects motor intentions toward an uncalibrated domain (zero-shot) consistently exceeding chance levels. The architecture achieved a mean accuracy of 74.51% ($\pm 15.68\%$) and a mean F1-Score of 0.7468. These central tendencies, closely aligned with their medians (75.44% and 75.80%, respectively), demonstrate that the model maintains a low false-positive rate despite the variance introduced by the transfer task. Additionally, Cohen's Kappa averaged 0.4901 (± 0.3136), indicating moderate to substantial agreement under strong inter-hemispheric interference.

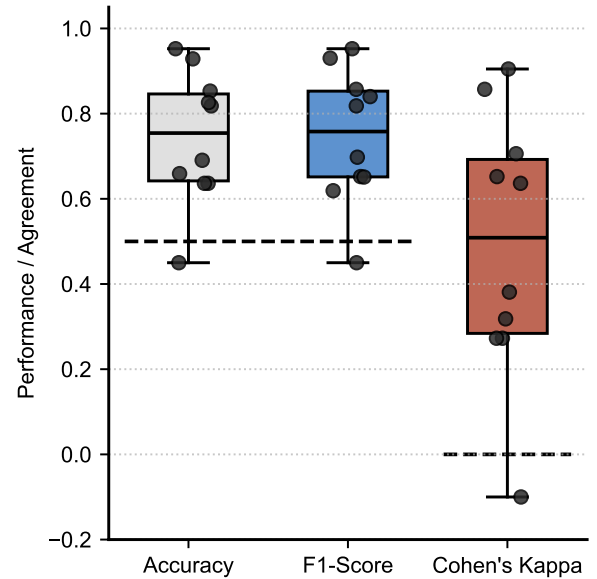


Figure 2: Distribution of decoding performance across the validated 10-subject cohort. The dashed horizontal line denotes the theoretical chance level, while the boxplots indicate the model consistently decodes bimanual intention with controlled variance.

Mitigation of BCI Illiteracy and Parametric Stability: For a neuroprosthesis to operate safely, the algorithm must surpass an operational competence threshold—typically set at a 60% F1-Score—to prevent erratic device control.

Fig. 3 details the individual performance of the valid cohort and exposes the empirical impact of the

nested cross-validation. The statistical audit confirms that 90% (9 out of 10) of the validated subjects surpassed the competence threshold. While one subject remained in the BCI illiteracy zone, achieving an operational viability of 90% represents a significant stabilization for zero-shot bimanual decoding without target-domain retraining. The dual-axis chart reveals a direct correlation between the parametric stability of the model and decoding success. Subjects converging toward a stable Hankel order during the isolated calibration phase maintain robust performance, whereas the underperforming outlier exhibited parametric instability across the inner folds. By resolving the domain shift as a purely geometric alignment in Riemannian space, the model rescues motor intentions that traditional spatial filters lose to bimanual noise.

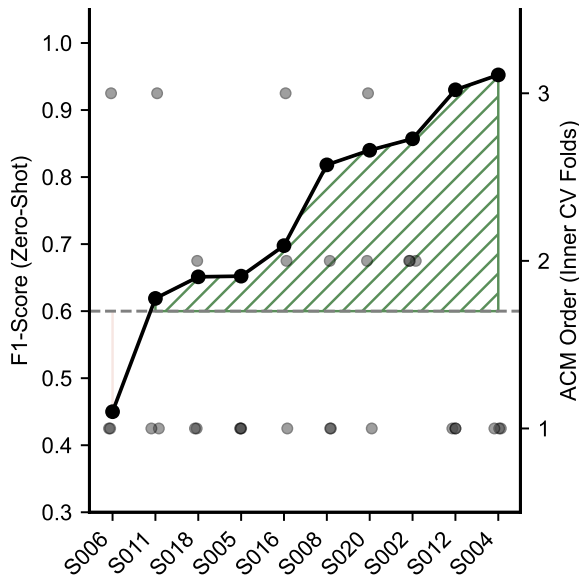


Figure 3: Subject-specific F1-Scores sorted in ascending order (left axis) correlated with the selected Hankel order during inner cross-validation (right axis). The shaded area confirms that 90% of the cohort surpassed the 60% operational competence threshold.

Tab. 1 outlines the individualized decoding metrics across the validated cohort. A fine-grained analysis reveals that subjects such as S004 and S012 achieved near-optimal transfer, with F1-scores exceeding 0.93 and strong Cohen’s Kappa agreements ($\kappa > 0.85$). Conversely, the structural domain shift severely impacted S006, yielding an accuracy below chance level (45.00%) and a negative Kappa value (-0.1000). This specific failure case highlights the biological constraints of zero-shot Riemannian calibration; when the underlying sensorimotor rhythms exhibit profound non-stationarity or a total lack of ERD lateralization, geometric adjustments alone cannot synthesize a viable classification space.

Physiological Evidence of Domain Shift: To em-

Table 1: Subject-Specific Zero-Shot Transfer Performance

Subject	Accuracy (%)	F1-Score	Cohen’s κ
S002	85.29	0.8571	0.7059
S004	95.24	0.9524	0.9048
S005	63.64	0.6522	0.2727
S006	45.00	0.4500	-0.1000
S008	81.82	0.8182	0.6364
S011	63.64	0.6190	0.2727
S012	92.86	0.9302	0.8571
S016	69.05	0.6977	0.3810
S018	65.91	0.6512	0.3182
S020	82.61	0.8400	0.6522
Mean	74.51	0.7468	0.4901
Std	15.68	0.1587	0.3136

pirically validate that geometric transfer occurs over real motor oscillations, the spatial distribution of Event-Related Desynchronization (ERD) was extracted.

The topographic maps (Fig. 4) visualize exclusively the 15 sensorimotor channels retained by the model. The evidence indicates that, in the μ (8–13 Hz) and β (16–24 Hz) bands, unimanual execution strongly lateralizes attenuation to the contralateral hemisphere. In contrast, bimanual imagery induces a complex and diluted ERD distribution across both hemispheres. Amputating peripheral channels prevents the augmented covariance matrix from inflating its dimensionality with parasitic variance, allowing the algorithm to align strictly relevant cortical dynamics.

DISCUSSION

The transition from unimanual BCI calibration to bimanual execution represents a significant barrier in functional neurorehabilitation. The results of this study indicate that the AT-BCI architecture overcomes this obstacle by directly addressing the underlying physiological interference rather than relying on generic linear corrections. The topographic analysis of event-related desynchronization (ERD) in Fig. 4 reveals the physical root of the problem: while unimanual execution yields a clear, localized pattern over the contralateral hemisphere, the bimanual task dilutes this attenuation, creating an interhemispheric spatial masking effect. Traditional spatial filters inherently fail in this context because their underlying mathematical assumptions break down when simultaneous movements alter brain synchronization. By restricting the spatial topology and applying geometric alignment to center the covariance matrices, the AT-BCI model successfully neutralizes this domain shift.

It is essential to address a critical issue regarding model validation within this field. Several prelimi-

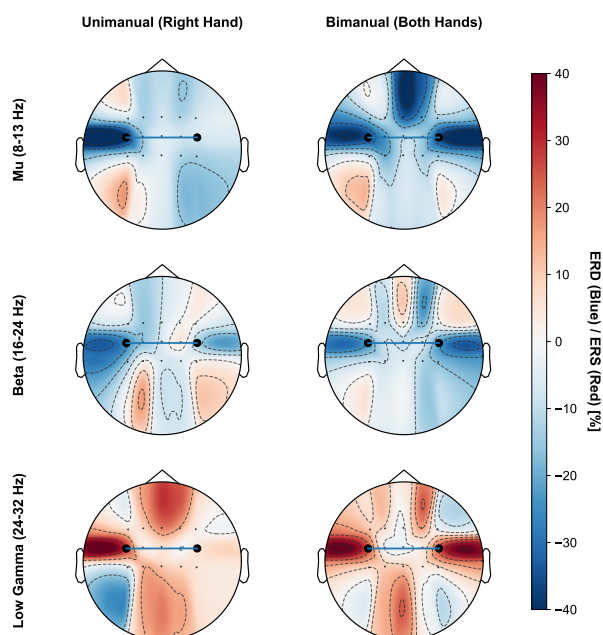


Figure 4: Spatial distribution of Event-Related Desynchronization (ERD) restricted to the 15 core sensorimotor channels. The topographic contrast between unimanual and bimanual tasks highlights the bilateral interference driving the domain shift.

nary approaches and baseline models in the literature report artificially high accuracy rates. As identified during the audit of our architecture, these performance spikes are often the product of a methodological overestimation or data leakage, where hyperparameters such as the embedding time delay are tuned empirically by exploring the test dataset. By eliminating this practice through the analytical derivation of the delay ($\tau \approx f_c/4$) and the strict parametric isolation provided by nested cross-validation, the mean accuracy achieved across our valid cohort (74.51%) represents a conservative but mathematically honest metric. This rigor ensures that the reported performance reflects a genuine zero-shot transfer capability, guaranteeing that the model will not fail due to overfitting when deployed with new patients in real clinical environments.

From a medical application perspective, mitigating BCI illiteracy stands out as a high-impact finding. Ensuring that 90% of the evaluated subjects surpass the 60% operational competence threshold has a direct clinical translation: patients could control bilateral exoskeletons without the physical and cognitive exhaustion imposed by strenuous daily recalibration sessions.

Nevertheless, the physical limits of Riemannian geometry must be acknowledged. The behavior of the single outlier who failed to reach the operational threshold (10% of the cohort) demonstrates that parametric stability is a strict prerequisite for reliable decoding. If a patient exhibits inherently weak sensorimotor rhythms or an excessively unstable sig-

nal, spatial alignment lacks robust underlying patterns to rescue, meaning such cases will continue to require subject-specific interventions.

Finally, while the framework demonstrated strong mathematical resilience in simulations, transitioning this model to a closed-loop robotic interface introduces new challenges. Computing the Riemannian Fréchet mean is an iterative and computationally heavy process. To maintain the extremely low latencies required for fluid object manipulation in the real world, future online implementations must adapt this geometry to process data within continuous sliding windows.

CONCLUSION

In this study, we presented the Adaptive Topological BCI (AT-BCI) architecture, which reformulates the decoding of bimanual motor imagery tasks. Instead of relying on traditional spatial filters, we approach the problem as an unsupervised geometric alignment task. By embedding the temporal dynamics of the neural signal via Hankel matrices and projecting these features onto a Riemannian manifold, the model successfully isolates and neutralizes the severe interhemispheric interference generated during simultaneous hand movements.

Evaluation of the framework demonstrated its empirical viability. Following the implementation of a strict cleaning protocol to eliminate electromyographic contamination, analysis on a validated cohort of 10 subjects yielded a mean accuracy of 74.51%, an F1-Score of 0.7468, and a Cohen's Kappa of 0.4901. Crucially, the nested cross-validation paradigm proved that combining an analytically derived time delay with geometric domain adaptation is the essential mechanism to bridge the performance gap between unimanual and bimanual tasks, avoiding data leakage and reducing inter-subject variability.

From a clinical translation perspective, the architecture actively mitigates the persistent challenge of BCI illiteracy. By recovering motor intentions that were previously obscured by noise, the system enabled 90% of the evaluated users to surpass the 60% functional competence safety threshold. By reducing the reliance on exhausting, daily recalibration sessions, AT-BCI establishes a mathematically stable foundation to operate bilateral neuroprosthetics more accessibly.

Finally, to advance this approach toward real-time applications, future work will focus on data retention. We project the evaluation and integration of automated artifact removal methods such as Artifact Subspace Reconstruction (ASR), MARA, or ADJUST. Implementing these tools will ensure signal quality in an automated manner, optimizing model robustness for closed-loop rehabilitation technologies

in real-world scenarios.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for the financial support and Master's scholarship granted to develop this research project.

REFERENCES

- [1] Millán JR, Rupp R, Müller-Putz GR, Murray-Smith R, Giugliemma C, Tangermann M, et al. Brain-computer interfaces, a review. *Machine learning*. 2010;14(1):31–50.
- [2] Yger F, Bercher M, Lotte F. Riemannian approaches in brain-computer interfaces: A review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2016;25(10):1753–1762.
- [3] Lotte F. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces. *Proceedings of the IEEE*. 2015;103(6):871–890.
- [4] Pfurtscheller G, Silva FH Lopes da. *Handbook of electroencephalography and clinical neurophysiology: Event-related desynchronization*. Elsevier, Amsterdam, Netherlands (1999).
- [5] Neuper C, Scherer R, Reiner M, Pfurtscheller G. Imagery of motor actions: Differential effects of kinesthetic and visual-motor mode of imagery in single-trial eeg. *Cognitive brain research*. 2006;25(3):668–677.
- [6] Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller KR. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*. 2008;25(1):41–56.
- [7] He H, Wu D. Transfer learning for brain-computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*. 2019;67(2):399–410.
- [8] Zanini P, Congedo M, Jutten C, Said S, Berthoumieu Y. Transfer learning: A riemannian geometry framework with applications to brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*. 2018;65(5):1107–1116.
- [9] Carrara I, Papadopoulou T. Classification of bci-eeg based on the augmented covariance matrix. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2024.
- [10] Kumar S et al. Performing bimanual tasks with a bci: Combining a brain-controlled hand exoskeleton with the functional limb. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2025.
- [11] Takens F. Detecting strange attractors in turbulence. *Dynamical systems and turbulence*, Warwick 1980. 1981:366–381.
- [12] Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR. Bci2000: A general-purpose brain-computer interface (bci) system. *IEEE Transactions on Biomedical Engineering*. 2004;51(6):1034–1043.
- [13] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215–e220.
- [14] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain-computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*. 2011;59(4):920–928.

Co-Adaptive MRCP-based BCIs: Modulation of Cortical Potentials by Enhanced Classifier Performance

Patrick Suwandjjeff¹, Gernot R. Müller-Putz^{1,2}

¹Institute of Neural Engineering, Graz University of Technology, Graz, Austria

²BioTechMed Graz, Austria

E-mail: gernot.mueller@tugraz.at

ABSTRACT: Movement-related cortical potentials (MRCPs) reflect voluntary motor intentions and are widely used as control signals in asynchronous brain-computer interfaces (BCIs). However, the extent to which task engagement and feedback influence MRCP dynamics remains unclear. We analyzed MRCPs across two identical online speller sessions to examine how task context and classifier reliability relate to preparatory brain activity. Healthy participants controlled a row/column speller by performing right-hand gestures, first using a classifier trained on cue-aligned offline data and subsequently using a classifier retrained on data collected during the online task. Trials segmented into four task stages of increasing complexity showed no significant differences in MRCP amplitudes, suggesting limited within-session effects. In contrast, MRCP amplitudes were significantly higher in the second session following retraining. This pattern suggests an association between improved classifier performance, task success, and enhanced MRCP expression, supporting a co-adaptive interpretation and highlighting the importance of session-wise factors in MRCP-based BCI operation.

INTRODUCTION

Brain-computer interfaces (BCIs) enable direct communication between the brain and external devices, allowing individuals with severe motor impairments, such as amyotrophic lateral sclerosis, to interact with technology using neural activity alone [1]. Among non-invasive approaches, EEG-based systems that decode movement intention are particularly promising [2-3]. A central neural correlate in this context is the movement-related cortical potential (MRCP), a slow cortical potential that begins up to two seconds before voluntary movement and peaks at movement initiation [4]. MRCPs are tightly time-locked to motor preparation, making them an attractive control signal for asynchronous BCIs. Despite their potential, MRCP-based BCIs face a major challenge: cortical signals are inherently non-stationary. Even during repeated performance of the same motor task, neural activity can vary substantially. Traditionally, this variability has been attributed to noise, fatigue, or nonspecific cognitive fluctuations. However, increasing evidence suggests that internal states, such as attention, motivation, expectancy, perceived agency, and reward processing, can modulate cortical potentials. This issue is

particularly relevant in co-adaptive BCI systems, where both the classifier and the user continuously adapt to each other [5-6]. Adaptive algorithms optimize decoding by updating decision boundaries or feature weights, while users adjust cognitive strategies, effort, and engagement based on perceived control and feedback reliability. In this bidirectional loop, subtle changes in classifier behaviour can influence the subjective experience of control: reliable detection enhances motivation and engagement, whereas inconsistent feedback reduces it. These cognitive-motivational adjustments can directly affect MRCP amplitude, slope, and spatial distribution, creating a dynamic interaction between neural signals and system performance. Another important consideration is that offline calibration data often do not fully capture the neural patterns that arise during interactive BCI use. In real-time speller tasks, participants experience continuous feedback, interface dynamics, and time pressure, which can modulate MRCPs in ways that are not present during calibration [7-8]. As a result, classifiers trained solely on offline data may underperform when deployed online, highlighting the need for retraining on task-relevant neural activity. In our previous work [2] with the same MRCP-based BCI spelling paradigm, retraining the classifier with online task data significantly improved system performance, measured in characters per minute (CPM). This raises the question: does enhanced control also alter the underlying neural signals? We hypothesize that differences in MRCP characteristics between sessions reflect attentional engagement and adaptation to the interface within the co-adaptive human-machine loop, which may subsequently enhance user motivation, rather than mere variability. Changes in classifier dynamics may influence perceived controllability, expectancy, and cognitive investment, reshaping preparatory cortical activity. Understanding how classifier properties affect cortical potentials is crucial for developing robust MRCP-based BCIs. If neural signals systematically depend on the interactive context, training and retraining strategies must consider these bidirectional effects. Examining session-to-session modulation under controlled task conditions is therefore a critical step toward separating technical adaptation from genuine neurocognitive changes in co-adaptive BCI systems.

MATERIALS AND METHODS

A. Participants and Setup

This study presents a subset of data drawn from a larger experiment involving 21 healthy volunteers (12 males; average age 26.3 ± 5.8 years). Brain activity was recorded using a 64-channel actiCAP EEG system (Brain Products GmbH, Gilching, Germany) with a sampling rate of 500 Hz. Sixty electrodes were positioned according to the 10-10 system on the head to capture EEG signals, while the remaining four electrodes recorded electrooculography (EOG), placed horizontally at the outer canthi of both eyes and vertically above and below the left eye. The reference electrode was set at FCz, with the ground on the right mastoid. To capture movement onsets, a self-developed motion capture system (Logitech HD Pro C920) at the institute was used. A marker attached to the participant's index finger enabled movement tracking. The device operated at a sampling rate of 30 Hz. Synchronization between EEG/EOG and motion capture data was achieved via the Lab Streaming Layer protocol.

B. Experimental Paradigm

The main study comprised both offline and online phases. This contribution focuses mainly on the online phase, which comprised two consecutive sessions. Sessions were separated by a short break of approximately 15 minutes, allowing participants to rest between sessions. In Session 1, participants controlled an online speller matrix using a classifier trained on cue-aligned offline data. During this offline session, EEG was recorded while participants executed 5 different hand gestures each collecting 150 trials. In those cue-based runs, a continuously rotating cross superimposed on a static fixation cross served as the visual cue, rotating 90° clockwise every 3-4 seconds [9]. We then performed a self-paced run for each gesture which was used as a test set to evaluate which gestures performs the best and this gesture was then selected for detecting clicks in Session 1. In Session 2, the same spelling task was repeated, however, the classifier was retrained using the online data collected during Session 1. In this session, we aligned the movements to the expected onsets during the online speller task. During both sessions, participants were instructed to spell the identical set of predefined goal words and sentences using a row/column scanning matrix, ensuring full comparability between sessions. Each session started with 2 easy goal words, followed by a break then we started with 2 medium words and at the end 2 full sentences. All were ordered randomly to avoid repetition. The matrix was arranged in a 5×6 grid, with rows and columns highlighted sequentially for 1.6 seconds each. Movement execution at target fields was detected by the classifier. An ESC field in each row allowed participants to exit incorrectly selected rows, and

a DEL field provided the option to correct spelling mistakes. The detailed procedure of the paradigm, including cue presentation, gesture selection, classifier training and retraining as well as the online speller task, is described in full in [3]. Figure 1 provides an overview of the experimental setup and an explanation of the two sessions classifier calibration and the online speller matrix.

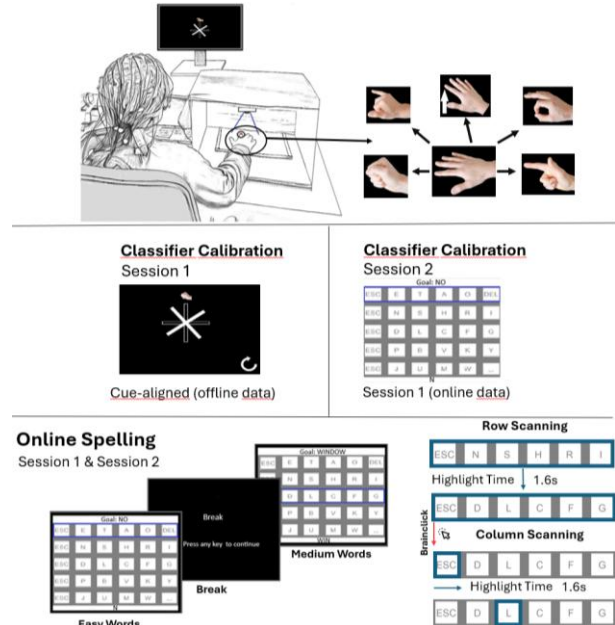


Figure 1: Experimental Setup. During offline phase 5 different right-handed gestures were performed. Session 1 used the best performing gesture (tested on test-set) for classifier training and control an online speller. Session 2 performed identical task, but the classifier was retrained on Session 1 data. Adapted from [3].

C. Data Processing & Online Classification

The signal processing pipeline included the following filtering steps: the EEG data was first band-pass filtered between 0.5 and 70 Hz to remove slow drifts and high-frequency noise. After this step, EOG artifacts were corrected using an eye-movement correction matrix [10]. The signals were then low-pass filtered at 3.5 Hz, downsampled to 10 Hz, and re-referenced to a common average reference. For filtering we used a 4th-order Butterworth IIR filter. To ensure methodological consistency, the offline data was processed using the same pipeline as applied during real-time operation, including the use of causal filters. All preprocessing and decoding procedures were implemented in BOLT (Brain Out-LeT), a Python-based framework for real-time neural decoding developed at the Wyss Center and built on the ezmsg architecture (<https://github.com/ezmsg-org/ezmsg>). For classification, a binary shrinkage linear discriminant analysis classifier was trained to detect movement onsets [11-12]. The calibration procedure followed the methods described in [2] and [3]. Specifically, the classifier for Session 1 was trained using

three cue-based runs consisting of 150 movement trials, while the Session 2 classifier followed the same procedure, using only the online data from Session 1 segmented into three runs.

D. Analysis

The analysis focused on MRCPs across Session 1 and Session 2 to assess neurophysiological and temporal characteristics. As the online speller task conditions were identical across sessions and only the decoder differed, additional analyses examined changes in MRCP characteristics and their relationship to behavioural performance. Low-frequency EEG amplitudes corresponding to MRCPs were used as features, and grand-average MRCPs were computed within a ± 2 s window around movement onset, pooling both sessions. Electrode C1 was selected for grand averages as it showed the strongest activity. Standard Error (SE) was included to reflect inter-subject variability. Significant differences between sessions (Session 1 - Session 2) were assessed using Wilcoxon rank-sum tests and corrected for multiple comparisons using the Benjamini-Hochberg procedure. Temporal variability between cue onset and movement onset was quantified by comparing cue-locked and movement-locked responses for each trial and participant. Session differences were evaluated to assess changes in MRCP amplitude, spatial distribution, and CPM at the single-subject level. Performance was quantified as CPM, where only characters that exactly matched the corresponding target position in the presented target words or sentences were counted as correct. Because scanning speed was fixed, CPM differences primarily reflect selection accuracy and efficiency. To examine within-session dynamics, trials were divided into four sequential stages (start, early, middle, late), each comprising 25% of trials. Grand-average MRCPs were computed for each stage across participants to assess potential learning or adaptation effects.

RESULTS

The grand-average MRCPs across all participants revealed significant differences between sessions at electrode C1 (Figure 2A). In particular, Session 2 showed a significant stronger negative peak amplitude ($-2.2 \mu\text{V}$) compared to Session 1 ($-1.2 \mu\text{V}$). The topographical analysis revealed significant differences at time points around movement onset (0 ms) and 125 ms (Figure 2B). These differences are primarily located over the contralateral side (right-handed movement execution) of the motor cortex as well as central motor areas. The subject-wise performance in CPM demonstrates an overall improvement when using the retrained classifier (Figure 3A). Nineteen out of the twenty participants showed an increase in CPM during Session 2, while only one participant did not exhibit a substantial improvement. Analysis of the temporal differences of the timing of movement execution relative to the cue onset

revealed reduced variability in Session 2 (Figure 3B). The interquartile range (IQR) for the movement onset relative to the cue was 0.18 s in Session 1 and 0.27 s in Session 2. For most subjects the timings in Session 2 are more centred around 0. The analysis of subject-wise MRCP amplitudes showed that most participants exhibited an increase in negativity during Session 2 (Figure 3C). The average values at C1 showed $-2.2 \mu\text{V}$ for Session 2 and $-1.2 \mu\text{V}$ for Session 1. The MRCP variability showed on a participant level a general pattern that the variability decreases for Session 2 (Figure 3D). Average values for Session 1 were $0.53 \mu\text{V}$ and $0.47 \mu\text{V}$ for Session 2. When comparing MRCP amplitudes between sessions, the calculated effect size was Cohen's $d = -0.5$, in contrast, the reduction in temporal variability showed a smaller effect size of $d = 0.27$. A comparison of the four stages of MRCP progression during Session 1 and Session 2 revealed stage-dependent differences in mean amplitude, although none reached statistical significance (Figure 4). In Session 1, the strongest negative MRCP amplitude was observed in the early stage ($-1.38 \mu\text{V}$), followed by the start stage ($-1.35 \mu\text{V}$) and the late stage ($-1.21 \mu\text{V}$), whereas the middle stage showed the weakest negativity ($-0.95 \mu\text{V}$). In Session 2, the strongest negative MRCP amplitude occurred in the middle stage ($-2.35 \mu\text{V}$), followed by the late stage ($-2.27 \mu\text{V}$) and the early stage ($-2.20 \mu\text{V}$), while the start stage exhibited the weakest negativity ($-1.98 \mu\text{V}$). However, statistical analysis revealed no significant differences between stages within either session.

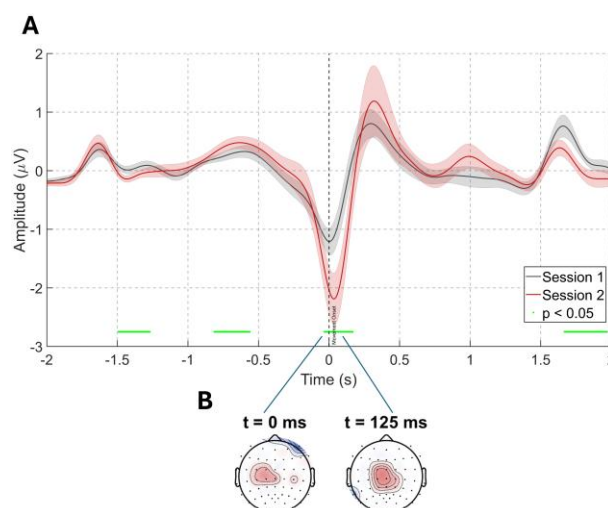


Figure 2: (A) Grand average MRCPs at electrode C1 for Session 1 (dark grey) and Session 2 (red), including SE, aligned to movement onset (-2 to +2 s). Significant differences between sessions are indicated in green. (B) Topographical distribution of significant differences between Session 1 and Session 2 from 0 ms to +125 ms relative to movement onset.

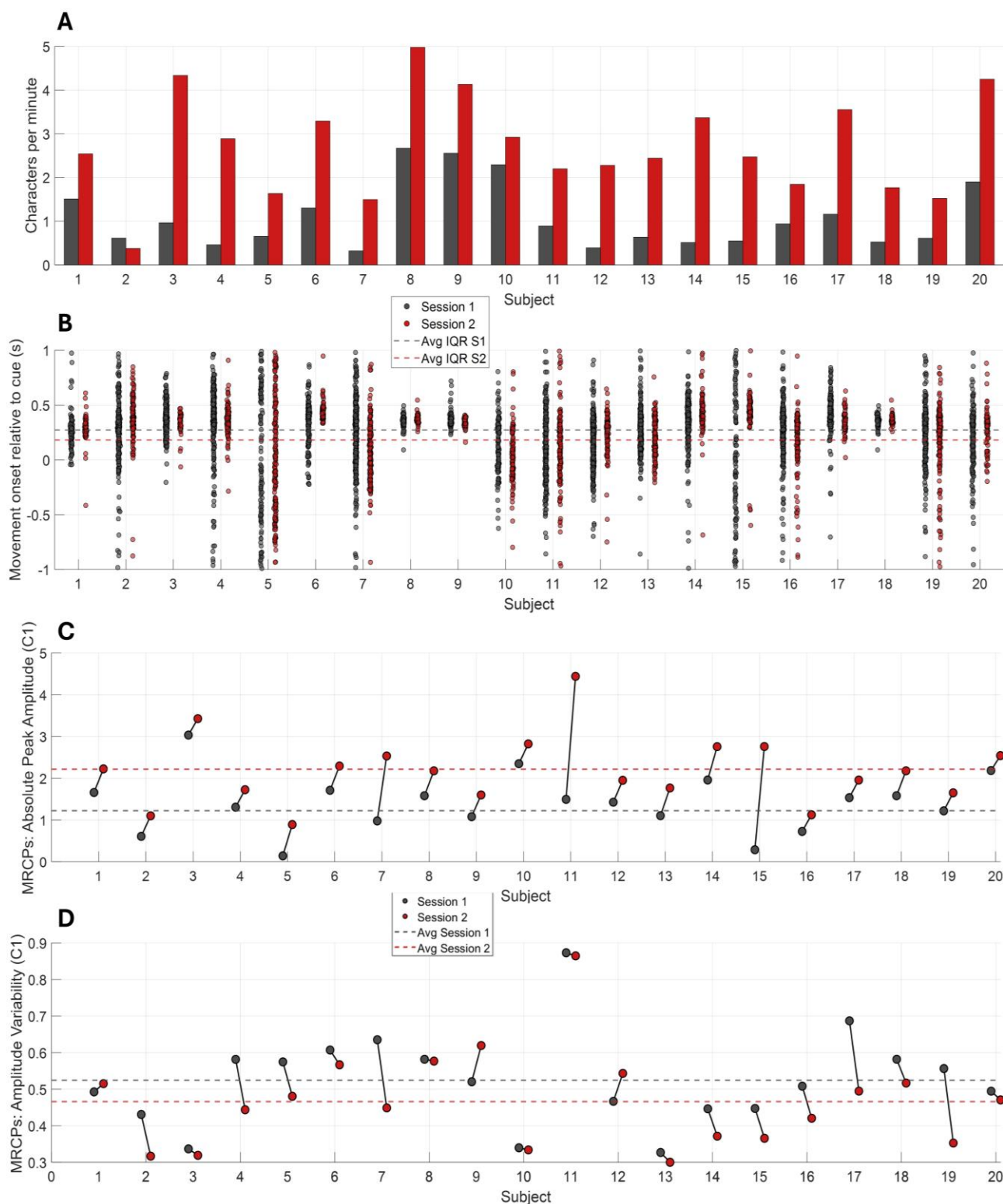


Figure 3: A) CPM per subject for Session 1 and Session 2. B) Time difference between movement onset and cue for both sessions per participant. The dashed dark-grey line shows the average IQR for Session 1, and the dashed red line shows the average IQR for Session 2. C) Subject-wise average MRCP peak amplitude for Session 1 and Session 2. D) MRCP amplitude variability per subject for both sessions. In all panels, Session 1 is shown in dark grey and Session 2 in red. The dashed lines in panel B indicate the average IQR for each session, while those in panels C and D indicate the average Amplitude/Variability for each session.

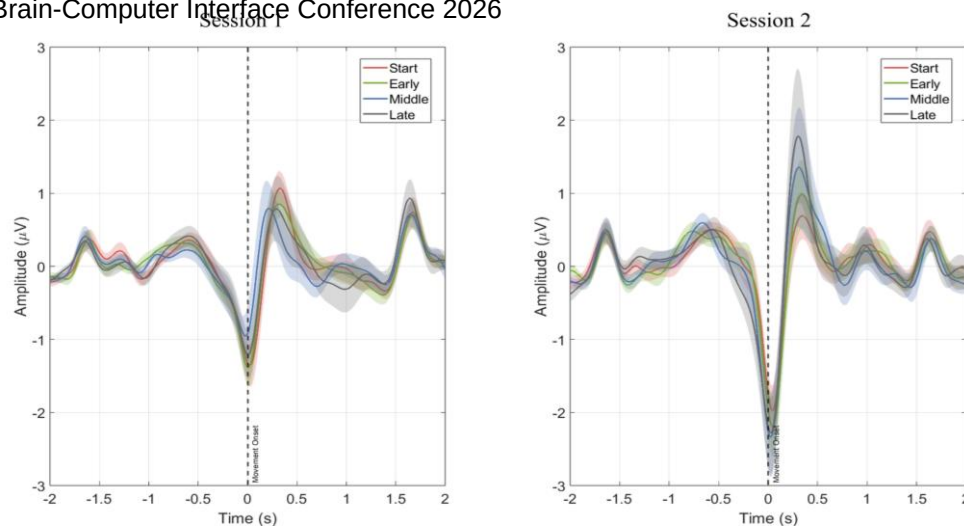


Figure 4: Grand average MRCPs of 4 different stages at electrode C1 for Session 1 and Session 2 including SE.

DISCUSSION

We observed a significant difference in motor-related brain activity between Session 1 and Session 2. Overall, the results indicate stronger and more consistent motor activation during the second session. This pattern is evident across multiple analyses. The grand-average MRCPs show a significantly stronger negative peak amplitude in Session 2 compared to Session 1 at electrode C1. Topographical maps further support this finding, revealing a larger negative MRCP amplitude at movement onset (0 ms) in Session 2. At 0 ms, the significant differences are primarily located over the contralateral motor cortex, with an additional effect at C4. At -125 ms relative to movement onset, the negative difference remains present but becomes more centrally distributed over the motor cortex. These spatial and temporal patterns are consistent with previous findings in the literature. Earlier work on slow cortical potentials and contingent negative variation (CNV) in BCI contexts has interpreted amplitude increases as correlates of learning and adaptive control, providing a framework for understanding preparatory brain activity changes in BCI systems [13]. An additional observation concerns the evolution of MRCP amplitudes across experimental stages. In Session 1, the strongest negative MRCP amplitude was observed during the early stage, followed by the start and late stages, while the middle stage showed the weakest negativity. In Session 2, the strongest MRCP amplitudes occurred during the middle and late stages, whereas the start stage showed the weakest negativity. However, these differences were not statistically significant in either session. Overall, the stage-wise analysis did not reveal a systematic evolution of MRCP amplitude within sessions, suggesting that within-session learning effects were minimal. Importantly, the increase in MRCP amplitude observed in Session 2 occurred after the classifier had been retrained and updated. Session 2 retraining used data recorded during the actual speller task, thereby improving alignment between training and testing distributions. This suggests an association between classifier adaptation, feedback reliability, and changes in motor-related brain activity. One possible explanation is

that more reliable feedback may support more stable and well-timed motor-related activity during task execution, potentially enhancing neural preparation and engagement compared to conventional feedback paradigms [14-16]. At the same time, it is also plausible that changes in neural activity and signal structure contributed to improved classifier performance, indicating a bidirectional relationship between neural dynamics and decoding accuracy. The results are therefore consistent with a co-adaptive process in which user and system mutually influence each other over time. In addition, psychological factors such as attention, engagement, and motivation may have contributed to the observed effects. Previous work has shown that these variables can modulate both BCI performance and neurophysiological signals suggesting a role in shaping MRCP dynamics alongside neural and system-driven factors. At the individual level, most participants showed stronger negative MRCP peaks in Session 2, although the magnitude of change varied across subjects. While most participants exhibited small increases in MRCP negativity, a few showed larger changes. Despite this variability, the overall effect size for the amplitude difference was moderate (Cohen's $d = 0.5$), while the reduction in variability showed a smaller effect ($d = 0.27$), supporting a general trend toward more consistent motor-related activity in Session 2. Several limitations should be considered when interpreting these findings. First, the direction of the relationship between MRCP modulation and classifier performance cannot be conclusively determined, as improved neural signal quality may also contribute to better decoding performance. Second, although sessions were conducted consecutively on the same day, factors such as task familiarity, attention, or fatigue may have influenced the results. Third, variability across participants suggests that the effects are not uniform and may depend on individual strategies or neural characteristics. Future work should address these factors by systematically manipulating feedback reliability and classifier performance. We acknowledge that a dedicated control condition in which Session 2 used the same classifier as Session 1 was not included. However, the within-session stage analysis did

not reveal systematic MRCP changes over time, suggesting that gradual learning effects within a session are unlikely to fully explain the between-session differences. Taken together, these findings suggest that MRCP modulation in BCI tasks is influenced by a combination of neural preparation, feedback characteristics, and learning-related processes. Rather than reflecting a single underlying mechanism, the observed effects likely emerge from the interaction between these factors within a co-adaptive BCI framework.

CONCLUSION

The present study suggests that motor-related brain activity became stronger and more consistent in the second online BCI session. This was reflected in increased MRCP amplitudes, reduced MRCP variability, decreased variability in movement timing relative to cue onset, and more stable motor activation patterns across participants. Crucially, these changes emerged after the classifier was retrained on task-related data, consistent with a co-adaptive interpretation of the system. In this framework, improved classifier performance may enhance the quality and interpretability of feedback, which in turn could support participants' engagement and more consistent motor preparation. At the same time, it is also plausible that increases in MRCP stability and amplitude contribute to improved classifier performance, suggesting a bidirectional relationship between neural signal quality and decoding accuracy. Together, these findings indicate that additional rounds of classifier retraining may further refine both neural modulation and classification performance, potentially increasing CPM by leveraging progressively more stable and discriminable neural signals. This co-adaptive interaction is particularly relevant for MRCP-based BCIs, which rely on self-initiated movements rather than externally driven visual stimuli. This design reduces visual demand and may facilitate a more natural interaction between user and system. Overall, the findings highlight the potential of adaptive BCI systems that continuously update classifiers based on task-related neural activity, enabling an interaction loop in which neural signals and decoding performance mutually influence each other over time.

ACKNOWLEDGEMENTS

This project is funded by the European Union's HORIZON-EIC-2021-PATHFINDER CHALLENGES program under grant agreement No 101070939 and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00198.

REFERENCES

[1] B. J. Edelman et al., "Non-invasive brain-computer interfaces: State of the art and trends," *IEEE Reviews in Biomedical Engineering*, 2024, doi:10.1109/RBME.2024.3449790.

[2] M. R. Crell, K. Kostoglou, P. Suwandjjeff, J. R. d. Cruz, and G. R. Müller-Putz, "A non-invasive, MRCP-based BCI for online communication," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 34, pp. 1272-1283, 2026, doi:10.1109/TNSRE.2026.3666564.

[3] P. Suwandjjeff, M. R. Crell, K. Kostoglou, and G. R. Müller-Putz, "Turning motor intentions into words: An MRCP-based BCI speller for motor-impaired users enhanced by task-specific calibration," *Journal of Neural Engineering*, 2026, under revision.

[4] H. Shibasaki and M. Hallett, "What is the Bereitschaftspotential?," *Clinical Neurophysiology*, vol. 117, no. 11, pp. 2341-2356, Nov. 2006.

[5] S. Wu, K. Bhadra, A.-L. Giraud, and S. Marchesotti, "Adaptive LDA classifier enhances real-time control of an EEG brain-computer interface for decoding imagined syllables," *Brain Sciences*, vol. 14, 2024, doi:10.3390/brainsci14030196.

[6] C. Vidaurre, A. Schlögl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "A fully on-line adaptive BCI," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 1214-1219, 2006.

[7] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, pp. R1-R13, 2007.

[8] J. Blumberg, J. Rickert, S. Waldert, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Adaptive classification for brain-computer interfaces," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2007, pp. 2536-2539.

[9] M. R. Crell, K. Kostoglou, K. Sterk, and G. R. Müller-Putz, "A novel paradigm for fast training data generation in asynchronous movement-based BCIs," *Frontiers in Human Neuroscience*, vol. 19, p. 1540155, Feb. 2025.

[10] R. J. Kobler, A. I. Sburlea, C. Lopes-Dias, A. Schwarz, M. Hirata, and G. R. Müller-Putz, "Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals," *NeuroImage*, vol. 218, p. 117000, 2020.

[11] R. Peck and J. Van Ness, "The use of shrinkage estimators in linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 5, pp. 530-537, May 1982.

[12] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components - A tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814-825, May 2011.

[13] Birbaumer N. Breaking the silence: brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology*. 2006 Nov; 43(6):517-32. doi: 10.1111/j.1469-8986.2006.00456.x. PMID: 17076808.

[14] T. Solf Frank et al., "The effect of multimodal and enriched feedback on SMR-BCI performance," *Clinical Neurophysiology*, vol. 127, no. 1, pp. 490-498, 2016.

[15] Á. Barbero and M. Grosse-Wentrup, "Biased feedback in brain-computer interfaces," *Journal of NeuroEngineering and Rehabilitation*, vol. 7, p. 34, 2010, doi:10.1186/1743-0003-7-34.

[16] A. Ramos-Murguialday et al., "Proprioceptive feedback and brain computer interface (BCI) based neuroprostheses," *PLoS ONE*, vol. 7, no. 10, p. e47048, 2012, doi:10.1371/journal.pone.0047048.

Brain-computer interface performance development over 30 hours of patient training with an auditory event-related potential protocol

Michael Tangermann¹, Alexander de Ranitz^{1*}, Soz Raheem^{1*}, Mariacristina Musso²

¹Donders Institute, Radboud University, Nijmegen, the Netherlands

²German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

*These authors contributed equally

E-mail: michael.tangermann@donders.ru.nl

ABSTRACT: Neural activity measured with electroencephalography can be decoded into control commands using a brain-computer interface (BCI). Longitudinal studies on BCI training are scarce. While some studies examined long-term training on motor imagery (MI) BCIs, multi-session user training using event-related potential (ERP) protocols remains largely unexplored, despite its relevance for novel cognitive rehabilitation protocols. *Methods:* In a post-hoc analysis, we investigated the ERP decoding performance in ten patients with stroke-induced aphasia, who completed a 30-hour multi-session feedback-based language training. Performance was assessed using an adaptive block-Toeplitz LDA classifier and quantified by the area under the ROC curve. *Results:* Decoding performance varied across patients and evolved over time. Several patients improved across sessions, indicating learning, while others showed stable performance. Increases in task difficulty caused sudden performance drops, often followed by recovery and gradual improvement. *Conclusion:* These findings provide insights into the longitudinal dynamics of ERP-based BCI training and highlight its potential for BCI-supported rehabilitation beyond MI paradigms.

INTRODUCTION

Brain signals such as the electroencephalogram (EEG) can be analyzed with machine learning methods and translated into control commands by a brain-computer interface (BCI) system. Such systems enable applications including device control, communication interfaces, mental state monitoring, or rehabilitation after stroke. Decoding information from brain signals is a central component of a BCI system. Decoding performance can be improved, not only with the availability of more data, but also with increased signal quality as a result of *user learning*. The latter means that the user learns to modulate their brain activity, such that more discriminative features can be extracted by the BCI, enabling improved decoding accuracy.

However, decoding performance can also deteriorate due to, e.g., disease progression or impedance problems in long-term implantations, both of which are observed in patients, for whom BCI systems are primarily designed.

Several studies report on the long-term *usage* of a BCI application for communication and control [1–4] or for artistic expression of patients [5]. As the goal of these is to provide a usable application, a recalibration of the BCI is typically avoided once the decoding performance reaches reliable control of the application. Furthermore, an explicit high-density, repeated, supervised *training* of the user, with the goal to further increase the decodability of the brain signals, is not performed, as the time investment for this *training* would leave the patient with less time for *using* the actual application. Nevertheless, user learning may be a side-effect of the continuous usage of the BCI application.

Contrary to these studies with a focus on long-term stability, longitudinal *changes* of brain signals are an explicit goal of (typically non-invasive) studies that employ BCIs to support neuro-rehabilitation [6]. They aim to support the patient to use and strengthen the remaining nodes and connections of, e.g., a stroke-lesioned brain network to overcome functional deficits. In motor rehabilitation protocols, this seems feasible by closing the loop between the (attempted) motor execution and sensory feedback, where the BCI has the role to detect the execution of the motor task. Therefore, it is beneficial if the patient learns to generate increasingly informative oscillatory band power changes over time, such that these can be decoded by the BCI as clearly as possible. As a result, the BCI can provide more informative feedback. For event-related potential (ERP) protocols, however, such longitudinal BCI learning scenarios are scarce. In [7], ten healthy participants used a P300 speller over 11 sessions. Half of the participants received only auditory stimuli and the other half a mixture of auditory and visual stimuli. The classifier was retrained before each session using data from previous sessions. For the auditory stimuli, accuracy improved notably over time, possibly indicating user learning. However, this improvement could in part be due to the classifier being trained on an increasing amount of data over time.

In [8], 16 healthy participants used an auditory P300 speller application over five sessions. The participants could see the letter selected by the BCI application, providing a form of feedback. The classifier was retrained

every session using calibration data collected at the start of the session. Their results showed a significant increase of the information transfer rate (ITR) over time. The authors stabilized the decoding accuracy over sessions by reducing the number of repetitions in a trial for participants with a high decoding accuracy, ensuring the task remained challenging. The increased ITR thus can be explained by the reduced time required per trial and indirectly reflects an improvement of the decoding performance.

While these studies indicate that longitudinal learning may also occur in ERP protocols, several questions remain open. Primarily, since these studies consist only of healthy participants, it is unclear whether learning also occurs in patients, the primary target group for most BCI systems. In the case of [7], it is difficult to discern whether learning has taken place as their experimental procedure and analysis were not focused on *user learning*. Additionally, no individual adaptation of the task difficulty was performed, possibly reducing learning pressure and resulting in some participants reaching a ceiling performance.

In this work, we address this gap by analyzing long-term training effects in an extended auditory ERP-BCI training procedure for patients with stroke-induced aphasia. For this purpose, we re-analyze a longitudinal ERP dataset collected in a study on BCI-assisted language rehabilitation [9], where the task difficulty was continuously adapted to keep the patients challenged.

MATERIALS AND METHODS

Dataset: The EEG data analyzed for this study has been recorded using a language rehabilitation protocol [9]. Each of ten patients (aged 38-76 years) with different levels of stroke-induced aphasia participated in approximately 30 hours of supervised auditory ERP-BCI feedback training. The intensive training used word stimuli to elicit ERP responses, spanned over multiple sessions, and took approx. four weeks per patient.

Sessions: Each patient started with two sessions to collect offline calibration data. In the original rehabilitation protocol [9], this data was used to calibrate the initial classifier. The final session was also an offline session and had served to study any pre-/post training differences. The intermediate sessions were online sessions, in which a trained classifier delivered feedback to the patient after each trial. In our analysis, only the data of the online sessions is considered, resulting in a total of 149 analysed sessions. The amount of sessions differed between patients, ranging from 14 to 28 sessions. However, the total amount of training sums up to 30 hours per patient, with the exception of P2, who underwent 22 hours of training only. Therefore, a patient with a large number of sessions on average had a shorter session duration than a patient with fewer sessions. As a consequence, the number of word presentations ranged between 2940 and 15144 per session. This variability can be clearly seen in Figure 1,

where the amount of recorded epochs, i.e. words presented, is displayed per session.

Experimental task: The online training consisted of a modified AMUSE protocol [10], using six word stimuli to repeatedly probe the patient's ability to detect repeated occurrences of a target word in a sequence of non-target words. No specific instructions on how to attend to targets or ignore non-targets were given. In most of the *trials*, the patient heard a pseudo-random sequence of 90 word stimuli in total, consisting of 75 non-target words and 15 target words. The EEG signal describing the response evoked by a single word stimulus is referred to as an *epoch*. In the original experiment, an adaptive classifier had been trained to estimate the class label (target or non-target) of each stimulus word based on the ERP information contained in the corresponding epoch.

Feedback: Patients received frequent feedback about the classification during the online training sessions. At the end of each trial, graded audiovisual feedback informed the patient whether the target word was decoded correctly. Especially positive feedback was provided in case dynamic stopping occurred, see [9]. It was hypothesized that providing feedback may help patients to find an individual strategy for word processing, which in turn may increase the positive, reinforcing feedback.

Experimental conditions: Multiple experimental conditions were used in this language training protocol. The conditions varied w.r.t. spatial presentation and stimulus onset asynchrony (SOA). The spatial condition determined whether words were presented to the patient through six loudspeakers surrounding the patient's head (6D), or through a mono stream delivered via headphones (HP). The latter made the task more difficult, as spatial information was removed. The SOA defined the time between the onsets of two consecutive word presentations. For most conditions, it was either 250 ms or 350 ms, but rarely reached up to 1000 ms. Depending on the classification performance of the first two offline sessions, the first online session started with the *6D 250* or the *6D 350* condition.

Changing the task difficulty:

Between sessions and, on a few occasions, within sessions, the experimenters changed the task difficulty with the goal to challenge, but not to overburden the patient. The most important change involved making the task more difficult by switching from the 6D condition to the HP condition, or by shortening the SOA. In addition, single words that had shown very high decoding accuracy on two consecutive sessions when they were in the target role were replaced between sessions. The training intensity was also maintained by using dynamic stopping of ongoing trials [11] that showed a clear decoding of the target word. As a consequence, trials were shorter and more trials—each resulting in feedback to the patient—could be executed in the same amount of time. As, according to the patients, the condition changes had the strongest impact on the task difficulty, we have included this factor into our analysis.

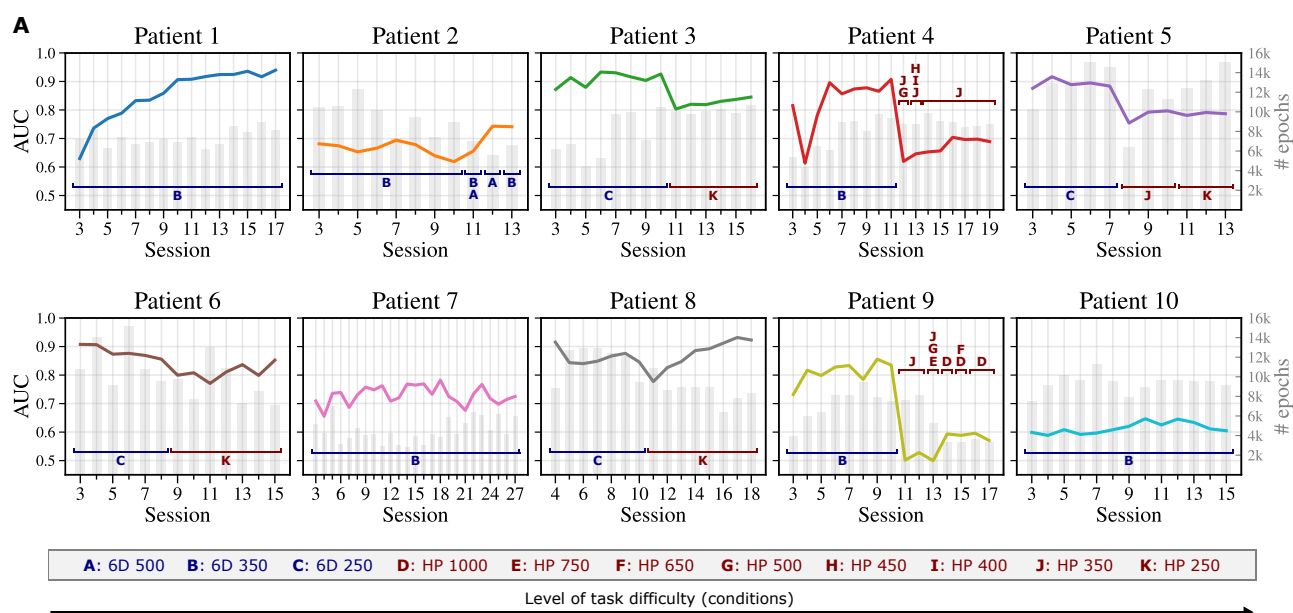


Figure 1: **Individual session-wise learning curves.** Colored lines provide the ROC-AUC 5-fold cross-validation performance per patient and session. The number of sessions differs between patients. Gray bars indicate the number of epochs recorded per session. For condition details please see the main text. Conditions marked in red are substantially more difficult than those in blue.

Data pre-processing: The raw 31-channel EEG data was band-pass filtered between 0.5 Hz and 16 Hz and cut into epochs around each stimulus event. Per epoch and channel, the average amplitude of 18 consecutive time intervals was extracted. The intervals had a width of 50 ms, starting at 100 ms post stimulus onset, and ending 1000 ms after stimulus onset. This resulted in 31×18 average microvolt values, which were flattened into a feature vector of shape 1×558 .

Epochs were rated outliers and removed when exceeding 25 inter-quartile ranges (IQR) based on the absolute magnitude of each feature vector. The procedure was applied independently for each subject and resulted in the removal of 738 epochs in total.

Classification: To track the performance development over time, we deployed a binary classifier. Based on the feature vector of an epoch, it classified the corresponding word as either a target (attended by the patient) or non-target (ignored). To generate a trial-level prediction, i.e., which out of the six words was attended, these epoch-level predictions can be combined at the end of the stimulus sequence. For the purposes of this paper, however, we only consider the epoch-level binary predictions in our post-hoc analysis. These were delivered by the recently introduced block-Toeplitz linear discriminant analysis (btLDA) [12]. It efficiently regularizes the covariance estimation in ERP protocols, and is advantageous specifically with a larger number of temporal features or limited training data.

Evaluation of decoding performance: Per patient, we estimated the classification performance across the training duration in two ways: In the *session-wise* analysis, each session was treated as a separate dataset. This analysis may suffer from performance variances caused by

the varying amount of words, and thus data points, available per session. In the *windows-based* analysis, the on-line data of a patient was cut into overlapping windows of 5000 epochs each. Windows moved with a stride of 1000 epochs. This analysis did not respect the session structure, but instead ensured that each classifier was trained on an equal amount of data. This allows for a meaningful comparison between windows across time. Each window of 5000 epochs was subsequently treated as a single dataset to train and evaluate the classifier on.

For both analyses, we report the area under the receiver-operating characteristic curve (ROC-AUC) of the binary epoch-level classification based on a 5-fold cross-validation, as the ROC-AUC is insensitive to class imbalances and classifier bias. It should be noted that both analyses are adaptive over time, but are not strictly causal due to the within-session or within-window cross-validation procedure.

RESULTS

We present the results in three views, which differ in how the segments of the dataset have been determined, for which the AUC score is computed:

Learning curves over sessions: Figure 1 shows the learning curves for each patient for the *session-based analysis*. The amount of data points contained in each session differs between and within patients, as indicated by the lengths of the vertical gray bars in Figure 1. To check whether sessions with more epochs lead to better classification, a correlation test was conducted. We observed a significant ($p = 0.002$) but weak correlation of $r = 0.26$ between the session size (in epochs) and the AUC obtained. However, experimental condition may act

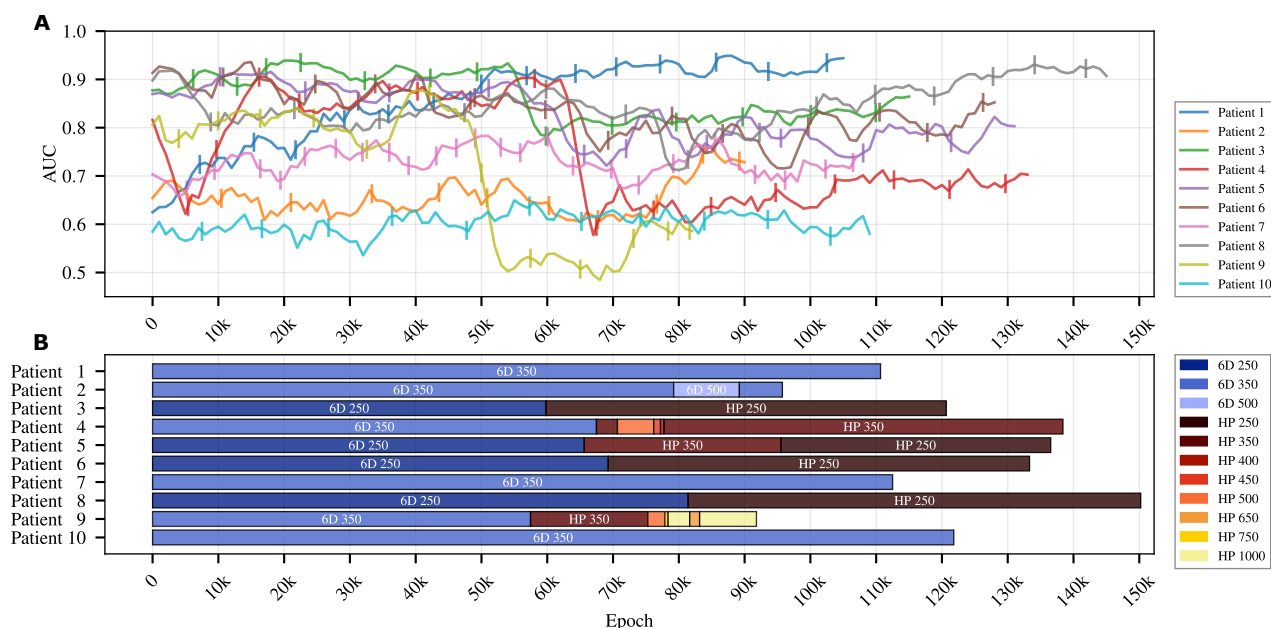


Figure 2: **Windows-based performance curves.** (A) Five-fold cross-validation ROC-AUC for the moving window starting at that epoch. Per curve, vertical bars indicate the session boundaries. (B) Overview of experimental conditions used for each patient across all epochs. ‘6D’ refers to the stereo setup with 6 speakers surrounding the patient, and ‘HP’ refers to the headphones condition. The subsequent number indicates the SOA in milliseconds. Please note that due to the data windowing in the training and validation process, the curves in the upper plot end between 5000 and 5999 epochs earlier than the corresponding bar in the lower plot.

as a confounding variable. For example, shorter SOAs are often used with high-performing patients, and sessions with shorter SOAs tend to have more epochs.

Learning curves over constant window sizes: Figure 2 shows the learning curves for each patient over time for the *window-based* analysis (top), and indicates which experimental conditions have been used (bottom). Unlike Figure 1, per patient a classifier slides with a window size of 5000 epochs over the data and treats each *window* as a separate dataset, eliminating the effect that varying dataset sizes across sessions may have on the classification performance. However, this approach does not respect session boundaries, which can result in windows that contain data from multiple sessions.

Learning curves over constant experimental conditions: Figure 3 shows the performance development relative to the first window, and separately for data *segments* from consecutive sessions (of a single patient) that share the same experimental condition. Only segments containing at least two sessions and 20,000 epochs are considered. We observed positive slopes for the HP conditions (right plot), while for the 6D conditions (left plot) no dominating slope direction could be identified, as depicted in the corresponding histograms.

DISCUSSION

We have derived the BCI performance curves for data obtained from stroke patients who received a longitudinal ERP-based language training protocol. To our knowledge, this data is unique in that task difficulty was continuously adjusted to remain challenging throughout the 30

hours of training. This design characteristic of the original study may influence the shape of learning curves compared to a mere long-term *usage* of an ERP application—a well-working application would not be made more difficult on purpose.

In the following, we use the term *learning* to denote changes in classification performances over time. Please note that the clinically targeted language improvement has been discussed in [9] and will not be regarded here.

Overall learning effects: On the group level, Musso and colleagues reported training induced stronger ERP components and reduced latencies [9]. Not unexpected due to their different lesions and stroke-induced deficits (see [9]), our offline post-hoc analysis revealed substantial differences between patients, as presented in the learning curves. The data of patient P1 exemplifies a continuously improved binary classification performance, starting with an AUC of just over 0.6 and reaching an AUC of more than 0.9 after 30 hours of training. As a clear learning curve is observed not only in the session-wise analysis, but also using the moving window classifier in Figure 2, P1’s improvement cannot solely be caused by larger training data sets per session, but more likely represents a proof of concept that learning in this patient group is feasible. On the other end of the spectrum is the learning curve of patient P10. It shows a relatively stable performance over 30 hours without a major improvement in classification.

Influence of conditions: We found that learning effects can be minimal in one experimental condition, but large in another, as exemplified by patient 8: While P8’s performance in the 6D 250 condition remained rather sta-

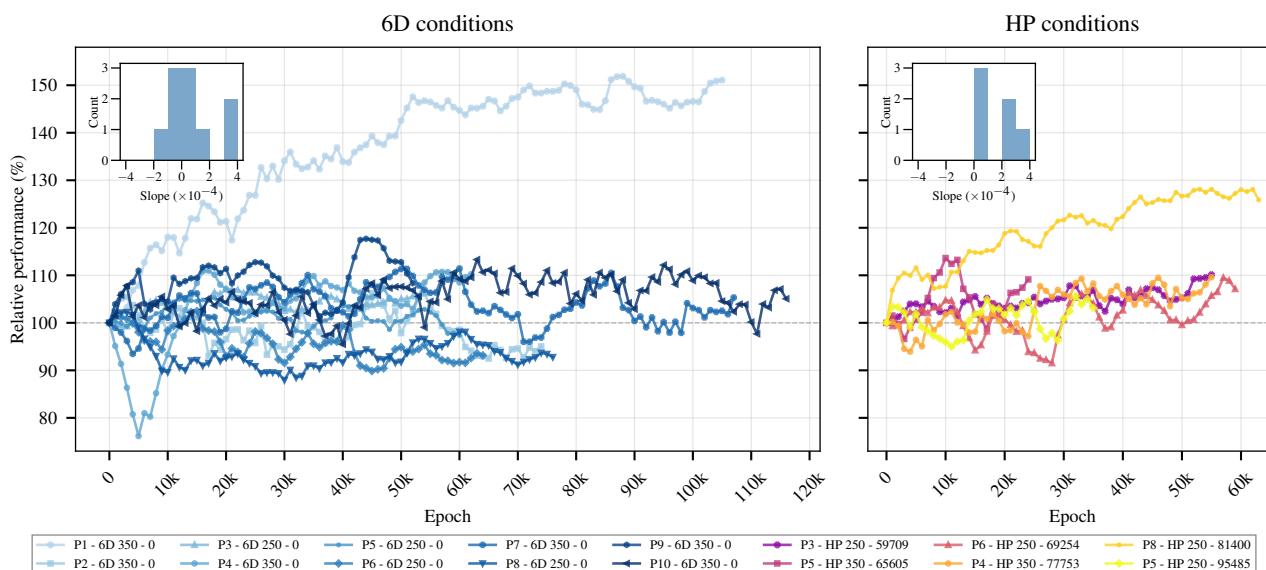


Figure 3: **Relative performance curves for data segments of stable conditions.** Each line represents a segment, that shares the same experimental condition and contains data of consecutive sessions. Multiple lines (segments) are shown per patient, if multiple conditions have been used during his/her training. Conditions are denoted as in Figure 1. The segment legend denotes the patient number combined with the condition information and the epoch number the segment started at. **Left:** Performance curves of data segments that used the 6D condition. **Right:** Performance curves for HP segments. Note that no patient started the online training with the HP condition. The curves depict the *relative* performances with respect to the 5-fold AUC score obtained at the first data window of each segment. For both conditions, histograms show the distribution of linear slopes for the segments contained.

ble at an AUC between 0.8 and 0.9, the change to the more challenging *HP 250* condition started an interesting process: After an initial drop to AUC values below 0.7, the learning curve recovered over the following 70.000 epochs (8 sessions) and eventually surpassed the initial 6D performance. This suggests that increasing task difficulty may stimulate the learning for this patient, while the original 6D condition may not have provided a sufficiently large learning pressure or motivation to further improve the ERP classification.

While we have not analyzed a potential influence of SOA values on the learning curves, the results of P8 exemplify our finding that increasing learning curves were observed for all HP segments, but only for six out of ten 6D segments, see the histograms in Figure 3. Certainly, it would be premature to claim that the harder HP conditions are more beneficial for learning, as the unbalanced histogram finding could be caused by ceiling- or floor effects as well, or by the need to first adjust to the harder HP condition after the easier 6D condition.

Comparison with ERP learning curves from other studies: The studied patient data may be special in the sense that the patients' initial performances were probably lower than those of healthy controls. While healthy controls may quickly run into a ceiling performance, patients in principle could have increasing learning curves.

This is similar to what is seen in [7] and [8]. In the study by Klobassa and colleagues, participants using a P300 speller with relatively easy audio-visual stimuli show good performance from the start, followed by little to no learning over sessions. However, participants using

the more challenging auditory-only P300 speller initially had a much lower performance, but do show meaningful learning over time. In the study by Baykara and colleagues, significant learning is only observed in the first three sessions, possibly indicating that a ceiling performance is reached rather quickly.

Session-based and windows-based analysis: Our *session-based* analysis highlights the influence of different sessions on the decoding performance. Since sessions were conducted on different days, this analysis best reflects day-to-day differences caused by factors such as sleep quality, mood, cap fitting, and any learning that occurs between sessions. This session-based analysis, however, has the disadvantage that results are confounded by varying dataset size. Our *window-based* analysis used a constant amount of 5000 epochs for training and evaluating classifier performances, which is not a given for the session-based analysis. While this makes btLDA performances comparable, the window size is a critical trade-off between training set size (and thus classification performance), the ability to follow non-stationary EEG distributions, inter-session effects and inter-condition effects. The latter may be specifically critical, as a window that bridges between an easier and a more difficult condition smears out and dampens any condition-induced performance differences over time. Around these transitions, the moving window classifier mixes data from two experimental conditions, for which different features may be informative. This can result in a seemingly reduced classifier performance for that bridging window, even if the true performances are equal and overall higher in both

conditions, if analyzed separately. Due to these differences in the two analyses, we decided to include both the session-based and windows-based flavour in this work.

Adaptive, non-causal analysis: Our post-hoc analysis used sessions or sliding time windows to both train and evaluate a classifier. This allows the classifier to adapt to long-term changes of the signals, which are an intended consequence of the rehabilitation procedure. As each session or window was effectively treated as a separate dataset, the reported performance mainly reflects the internal data quality within that specific session or window. This design choice, however, did not allow to evaluate sessions or windows relative to a common reference dataset, which means that a similar performance of two windows does not necessarily imply similarity in the underlying features.

Comparison of online and offline sessions: Our analysis was restricted to the online sessions of the ERP training and has excluded the offline sessions which did not provide feedback and were conducted before and after the 30 hours of online feedback training. Musso and colleagues [9] reported a significant grand average improvement of the binary classification performance when comparing pre- (AUC approx. 0.64) and post-training (AUC approx. 0.76) offline sessions. Please note, however, that these values were obtained using a shrinkage-regularized LDA classifier, which may perform less than the btLDA classifier in our study [12].

Limitations: The interpretation of our results is challenged by the diversity in patients and the—despite a constant task—changing difficulty. As sessions were of differing length, and the time between sessions also varied, it is difficult to compare performances between patients. We are aware that improvements in BCI decoding do not necessarily result in improved language performance. This question will be addressed in future work with larger patient numbers.

CONCLUSION

We analyzed a rare experimental paradigm consisting of multi-session auditory ERP training with trial-wise feedback about the ERP classification and a continuously challenging task difficulty. The study was conducted with stroke patients, who at the beginning of the training showed, compared to healthy participants in a similar age range, a lower ERP classification performance.

Overall, we observed that the stimulus conditions determined most of the observed performance changes. Breaking down the data into multi-session segments of constant stimulus conditions revealed high variability between patients in their ability to improve the ERP decoding performance over sessions. Increasing learning curves were observed systematically for the more demanding headphone mono condition, compared to the easier 6D stereo condition, suggesting that higher task difficulty may promote learning in ERP-based training.

ACKNOWLEDGMENTS

Authors received support from the Dutch Research Council (NWO) via the Dutch Brain Interface Initiative (DBI2), grant 024.005.022 and from BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG), grant EXC 1086.

AUTHOR CONTRIBUTIONS

MT, SR, AdR: Conceptualization, Methodology, Writing - original draft - review & editing; SR, AdR: Formal Analysis MT: Supervision; MM: Writing - review & editing MT, MM: Funding acquisition

REFERENCES

- [1] Wolpaw JR et al. Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. *Neurology*. 2018;91(3):e258–e267.
- [2] Vansteensel MJ et al. Longevity of a brain–computer interface for amyotrophic lateral sclerosis. *New England J of Medicine*. 2024;391(7):619–626.
- [3] Milekovic T et al. Stable long-term BCI-enabled communication in ALS and locked-in syndrome using LFP signals. *J of neurophysiology*. 2018;120(7):343–360.
- [4] Nijboer F et al. A P300-based brain–computer interface for people with amyotrophic lateral sclerosis. *Clin. neurophysiology*. 2008;119(8):1909–1916.
- [5] Botrel L, Holz EM, Kübler A. Using brain painting at home for 5 years: Stability of the P300 during prolonged BCI usage by two end-users with ALS. In: *Int. Conference on Augmented Cognition*. 2017, 282–292.
- [6] Pichiorri F et al. Brain–computer interface boosts motor imagery practice during stroke recovery. *Annals of neurology*. 2015;77(5):851–865.
- [7] Klobassa DS et al. Toward a high-throughput auditory P300-based brain–computer interface. *Clin. neurophysiology*. 2009;120(7):1252–1261.
- [8] Baykara E et al. Effects of training and motivation on auditory P300 brain–computer interface performance. *Clin. Neurophysiology*. 2016;127(1):379–387.
- [9] Musso M et al. Aphasia recovery by language training using a brain–computer interface: A proof-of-concept study. *Brain communications*. 2022;4(1):fcac008.
- [10] Schreuder M, Rost T, Tangermann M. Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. *Frontiers in neuroscience*. 2011;5:112.
- [11] Schreuder M, Höhne J, Blankertz B, Haufe S, Dickhaus T, Tangermann M. Optimizing event-related potential based brain–computer interfaces: A systematic evaluation of dynamic stopping methods. *J of neural eng.* 2013;10(3):036025.
- [12] Sosulski J, Tangermann M. Introducing block-toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain–computer interfaces. *J of neural eng.* 2022;19(6):066001.

UniSwarm-SSVEP: Dynamic Content Mapping for Efficient SSVEP-Based Swarm Robot Control

Rui Du¹, Mingyi Wei¹, Hongxin Zhang¹

¹ Beijing University of Post and Telecommunications, Beijing, China

E-mail: hongxinzhang@bupt.edu.cn

ABSTRACT: Steady-state visual evoked potential (SSVEP)-based brain-computer interfaces (BCIs) are widely used for device control, but static one-to-one stimulus-command mapping limits scalability in swarm tasks, and simply increasing the number of stimulus targets may reduce recognition performance. This paper proposes UniSwarm-SSVEP, a swarm robot control system with a Dynamic Content Mapping Module (DCMM). By updating stimulus meanings according to the current control state, DCMM allows tens of physical targets to be mapped to hundreds of logical commands and supports flexible switching among single-vehicle, subgroup, and whole-swarm control modes. To validate this mechanism, we constructed a 20-target UniSwarm-SSVEP test system for a 153-command swarm control task. Under FBCCA, UniSwarm-SSVEP achieved an average accuracy of 91.56% and a standard information transfer rate (ITR) of 143.46 bits/min. Task-discriminant component analysis (TDCA)-based validation further improved the accuracy to 95.78% and the standard ITR to 268.80 bits/min, demonstrating the scalability and decoding compatibility of the proposed system.

INTRODUCTION

Brain-computer interfaces (BCIs) enable direct interaction between users and external devices by decoding neural activity [4]. Among noninvasive BCI paradigms, steady-state visual evoked potential (SSVEP)-based BCIs have been widely used owing to their high signal-to-noise ratio (SNR), high information transfer rate (ITR), and low training burden [1,2]. These advantages have made SSVEP-BCIs an effective approach for controlling robotic arms, unmanned vehicles, and other external devices [3,10,11,15]. However, most existing SSVEP-based control systems still suffer from inefficient utilization of brain-control resources. In this study, brain-control resources mainly refer to the user's stable attention, gaze capacity, and reliable EEG responses. Conventional SSVEP interfaces usually adopt a static stimulus-command mapping mode, in which each physical stimulus target is assigned to a fixed command meaning throughout the task. As the number of controlled devices, command types, or control scopes increases, this static mapping mode requires more physical targets or more complex interface switching, thereby consuming additional visual and neural

resources for command selection rather than efficient device control.

Previous studies have attempted to improve BCI-based swarm control through SSVEP-controlled robot swarms, VR-based heterogeneous swarms, VR-assisted UAV formations, shared control, and predefined group commands [5-12]. Although these approaches improve interaction to some extent, many of them still follow the static one-to-one mapping strategy exemplified in [13]. Under this framework, systems can usually broadcast predefined commands to the whole swarm or to limited groups, but flexible switching among individual robot control, group coordination, and global commands remains difficult without expanding the physical stimulus set. However, previous multi-target SSVEP studies have shown that classification accuracy and communication speed tend to decrease as the number of simultaneously displayed targets increases [14].

To address this limitation, this study proposes UniSwarm-SSVEP, an SSVEP-based swarm robot control system with a Dynamic Content Mapping Module (DCMM). Different from static mapping, DCMM dynamically updates the control object and command scope associated with each stimulus according to the current system state. Therefore, the same physical stimulus target can represent different individual robots, robot groups, or global commands in different contexts. This mechanism decouples the number of physical stimuli from the size of the logical command space, enabling scalable swarm scheduling while maintaining a compact and stable visual interface.

To evaluate the effectiveness of the proposed dynamic mapping mechanism, we designed a baseline comparison between UniSwarm-SSVEP and a 153-target static-mapping baseline constructed according to the strict one-to-one stimulus-command relationship in [13]. In this implementation, we used FBCCA to compare the static-mapping baseline system with UniSwarm-SSVEP and highlight the advantage of dynamic content mapping. In addition, task-discriminant component analysis (TDCA) was further introduced to examine whether dynamic content mapping can be combined with a high-performance decoder for shorter decision windows [16]. This experimental design verifies the advantage of dynamic mapping from both baseline comparison and algorithm-enhanced validation. This offers a practical engineering solution for high-efficiency scheduling swarm devices via brain control.

MATERIALS AND METHODS

1. overall architecture

The overall architecture of the system is as Figure 1 shows:

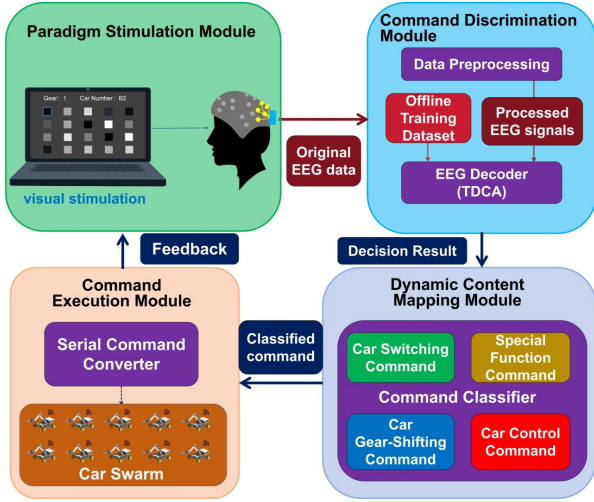


Figure 1: The overall architecture of UniSwarm-SSVEP. The system consists of four main modules: the Paradigm Stimulation Module, the Command Discrimination Module, the Dynamic Content Mapping Module and the Command Execution Module.

1.1 Paradigm Stimulation Module

This system adopts an SSVEP paradigm as the stimulation interface. The interface is divided into four functional regions: the Car Control Region, the Car Gear-Shifting Region, the Special Function Region and the Car Switching Region. Stimulus coding adopts the classic SSVEP frequency coding scheme, and the stimulus is presented in the form of a picture sequence to enhance the visual locking effect and improve signal quality.

1.2 Command Discrimination Module

The Command Discrimination Module supports interchangeable SSVEP decoders. In this study, filter bank canonical correlation analysis (FBCCA) served as the training-free baseline decoder, while task-discriminant component analysis (TDCA) served as the calibrated high-performance decoder [16]. The following section briefly introduces the TDCA formulation, because FBCCA is a standard correlation-based method and is used here only as the controlled baseline.

TDCA is one of the best-performing algorithms for current low-frequency SSVEP signal recognition. After augmenting the training data twice in both spatial and temporal dimensions, a shared spatial filter for different classes of signals is solved using two-dimensional linear discriminant analysis.

Calculate the between-class scatter matrix S_b and the

within-class scatter matrix S_w of the training data across all categories.

$$S_b = \frac{1}{N_f} \sum_{n=1}^{N_f} (\bar{A}_n - \bar{A}) (\bar{A}_n - \bar{A})^T$$

$$S_w = \frac{1}{N_f N_t} \sum_{n=1}^{N_f} \sum_{i=1}^{N_t} (A_n^{(i)} - \bar{A}_n) (A_n^{(i)} - \bar{A}_n)^T \quad (1)$$

$$\bar{A}_n = \frac{1}{N_t} \sum_{i=1}^{N_t} A_n^{(i)}, \quad \bar{A} = \frac{1}{N_f} \sum_{n=1}^{N_f} \bar{A}_n \quad (2)$$

After simplification, the optimization objective of TDCA is shown in Equation (3), where W is the spatial filter to be solved, and $\hat{W}$ is the estimate of W . Derived from the optimization problem, the first N_k eigenvectors of $\hat{W}$ can be used to construct the TDCA spatial filter W .

$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{\operatorname{tr}(W^T S_b W)}{\operatorname{tr}(W^T S_w W)} \quad (3)$$

After zero-padding a single test signal, the correlation coefficient ρ_n between the second-order augmented test signal and the spatially filtered template signal is calculated. The category corresponding to the maximum value of ρ_n is taken as the recognition result, as shown in Equation (4).

$$\rho_n = \operatorname{corr}(\hat{W}^T A', \hat{W}^T \bar{A}_n) \quad (4)$$

1.3 Dynamic Content Mapping Module

The Dynamic Content Mapping Module (DCMM) is the core component that distinguishes UniSwarm-SSVEP from conventional static-mapping SSVEP interfaces. It receives decoded labels from the Command Discrimination Module and interprets them according to the current control state, including the control scope, group index, car index, and gear state. Therefore, the semantic meaning of each physical stimulus target is dynamically updated rather than fixed. Thus, the same physical stimulus target can represent different commands in individual, group, or global control contexts.

After a decoded label enters DCMM, car-control and special-function labels are converted into executable commands for the currently selected object, whereas gear-shifting and car-switching labels update the system state and refresh the stimulus interface. In this way, DCMM forms a closed-loop process among EEG decoding, state transition, command generation, and interface updating, as illustrated in Figure 2.

The command expansion capability of DCMM can be described by the equivalent number of commands that would be required in a static-mapping interface:

$$\begin{aligned} Target_{sw} &= Target_{grp} \times (M+1) + 1 \\ Target_{car} &= Target_{ctrl} + Target_{gear} \\ N_{eq.} &= Target_{car} \times Target_{sw} \end{aligned} \quad (5)$$

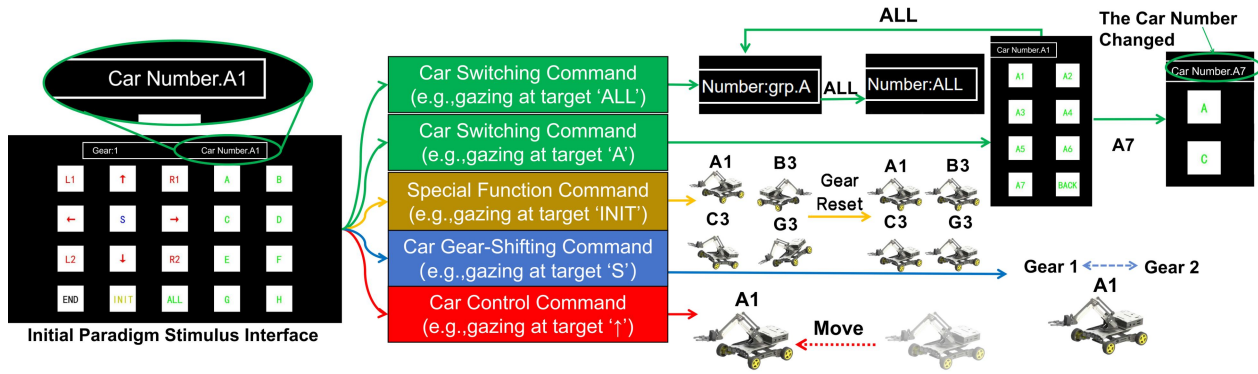


Figure 2: Functional architecture and dynamic content mapping logic of DCMM. Starting from the initial stimulus interface, decoded labels are interpreted according to the current control state. Car-switching commands, such as selecting “ALL” or a specific car label, update the control scope or selected vehicle; special-function commands perform synchronization operations; gear-shifting commands update the gear state; and car-control commands generate executable motion commands for the current object. After each state transition, the stimulus labels are refreshed, allowing the same physical target to represent different commands in individual, group, or global control contexts.

where $Target_{sw}$ denotes the number of selectable swarm control objects or scopes, $Target_{grp}$ denotes the number of car groups, and M denotes the total number of cars in each group. For each group, the selectable objects include M individual cars and one group-level scope, while the additional “+1” represents the ALL scope for global swarm control. $Target_{car}$ denotes the number of command targets available for each selected object, where $Target_{ctrl}$ is the number of control-command targets and $Target_{gear}$ is the number of gear-shifting targets. N_{eq} is the equivalent number of commands under a static-mapping interface. This formulation reflects the general expansion mechanism of DCMM: a limited number of physical stimulus targets can be reused across different control states to generate a much larger logical command space.

For the online experiment described in Section 2.3, the 12-car swarm was divided into four groups. Substituting the corresponding numbers of selectable control scopes and command targets into Equation (5) gives $N_{eq} = 153$. In the 20-target UniSwarm-SSVEP interface, 13 targets were directly related to task execution, whereas the remaining seven targets were reserved for modular design and future extension. Therefore, the 153-command value is specific to the experimental configuration used in this study, rather than a fixed upper limit of DCMM.

1.4 Command Execution Module

This module is responsible for communicating with the lower computer of the car swarm, receiving the specific control commands forwarded by the Command Dispatch Module, and sending them to the target cars via Wi-Fi to achieve precise motion control.

2. Experimental Design

2.1 Experimental Design

The experiment consisted of two parts: offline algorithm verification and online task testing. We collected EEG

signals from the same 15 subjects for both the offline and online experiments. The subjects were mainly students aged 18-25 years, including 9 males and 6 females. Each subject was asked to read and sign an informed consent form before the experiment. The Institution Review Board of Beijing University of Post and Technology approved this study.

2.2 Offline Experiment

2.2.1 Offline Data Acquisition

In the offline experiment, EEG signals were recorded using twenty stimulus blocks with distinct frequencies. The frequency range was 8-16 Hz with an interval of 0.4 Hz, and the phase difference between adjacent targets was 0.35π . In each trial, the subject was required to gaze at the specified stimulus block according to the cue. Each trial lasted 2 s, with an inter-trial interval of 1 s. Each block contained one trial for each of the 20 frequencies, and 4 blocks were conducted per experiment.

2.2.2 Offline Experimental Design

The offline data were used to evaluate the signal characteristics and algorithm performance of the system. EEG topographic maps were analyzed to examine the visual response patterns across subjects. In addition, FBCCA and TDCA were applied to the offline dataset to compare their classification performance and determine the decision window used for the subsequent online validation.

2.3 Online Experiment

To evaluate the dynamic content mapping mechanism, the same 15 subjects controlled 12 cars arranged into four groups and completed predefined individual-, group-, and global-level control tasks. According to the calculation in Section 1.3, this online configuration corresponded to 153 equivalent commands.

For the baseline comparison, we constructed a 153-target static-mapping baseline following the one-to-one stimulus-command relationship in [13]. UniSwarm-SSVEP used a 20-target dynamic interface to complete the same command set, whereas the baseline

represented the 153 commands with 153 static stimulus targets. Figure 3 compares the stimulus interfaces of the 153-target static-mapping baseline and the 20-target UniSwarm-SSVEP system.

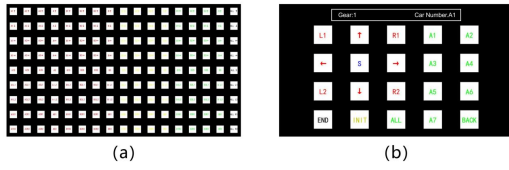


Figure 3: Stimulus interfaces used in the online comparison. (a) is the 153-target static-mapping stimulus interface, while (b) is the dynamic-mapping stimulus interface of UniSwarm-SSVEP.

In this comparison, we used FBCCA to compare the static-mapping baseline system with UniSwarm-SSVEP and highlight the advantage of dynamic content mapping. TDCA was evaluated only within UniSwarm-SSVEP for subsequent algorithm-enhanced validation, because its calibration time increases with the number of targets or classes [16,17], making it impractical for the 153-target static baseline.

In this experimental design, the number of additional gaze shifts to vehicle-switching stimulus blocks was one quarter of the number of direct vehicle-control commands.

RESULTS

1. Signal Characteristic Analysis

To examine the visual response characteristics of the acquired EEG signals, the EEG topographic maps of the 15 subjects under 10 Hz stimulation are shown in Figure 4.

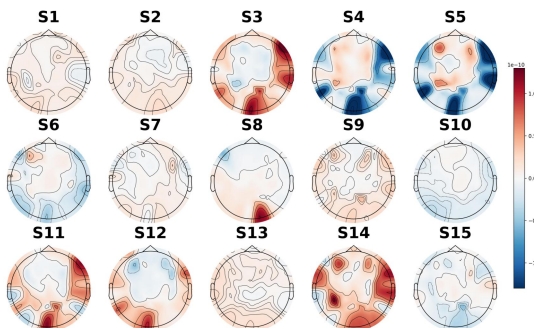


Figure 4: EEG topographic maps of the 15 subjects under 10 Hz stimulation. Most subjects showed clear occipital responses, indicating that the stimulation paradigm elicited stable visual responses across participants.

2. Offline Algorithm Performance Validation

To validate the superiority of the TDCA algorithm and determine its optimal stopping point, we conducted an offline analysis of classification accuracy across varying time windows. Specifically, both the FBCCA and TDCA algorithms were applied to the

offline dataset, with decisions made at 0.1 s intervals. The reported accuracy represents the average performance across 15 trials for each time window. Figure 5 illustrates the accuracy trajectories of the FBCCA and TDCA algorithms as a function of time window length. The results indicate that TDCA demonstrates a significant advantage over FBCCA, particularly in short time windows.

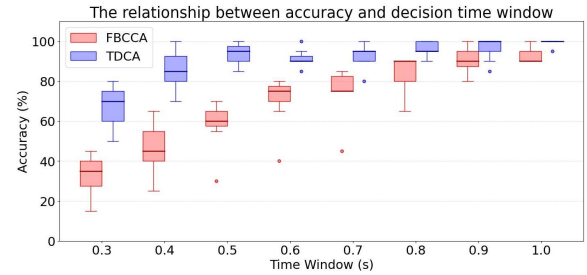


Figure 5: Classification accuracy of FBCCA and TDCA across decision windows. TDCA reached its highest performance around 0.8-1.0 s.

3. Online System Performance

Based on the online experiment described in Section 2.3, Table 1 summarizes the accuracy, standard ITR, and effective ITR of the static-mapping baseline and UniSwarm-SSVEP under FBCCA, together with UniSwarm-SSVEP under TDCA. All summary values are reported as mean $\pm$ standard deviation across subjects.

3.1 Baseline Comparison Between Static and Dynamic Mapping

This subsection focuses on the FBCCA-based comparison between static and dynamic mapping.

3.1.1 Classification Accuracy

Previous multi-target SSVEP studies have shown that classification performance tends to decrease as the number of simultaneously displayed stimulus targets increases [14]. This tendency was also reflected in the 153-target static-mapping baseline. As shown in Table 1, the static-mapping baseline achieved an average accuracy of $30.67 \pm 7.15\%$ across 15 subjects. In comparison, UniSwarm-SSVEP maintained an average accuracy of $91.56 \pm 6.02\%$ while realizing the same 153 equivalent commands with a 20-target dynamic interface.

This result suggests that dynamic content mapping can reduce the dependence of online recognition performance on the size of the logical command space. By reusing a compact set of physical stimulus targets across different control states, UniSwarm-SSVEP avoids the need to display all command targets simultaneously and preserves a simpler visual layout for online selection.

3.1.2 Standard ITR and Effective ITR

We used information transfer rate (ITR) to quantify the communication efficiency of the two systems.

$$ITR = \frac{60}{T} \left(\log_2 N + P \log_2 P + (1-P) \log_2 \frac{1-P}{N-1} \right) \quad (6)$$

Table 1: Comparison of Accuracy and ITR in Online Experiments

Sub ID	static-mapping baseline(FBCCA)		UniSwarm-SSVEP(FBCCA)			UniSwarm-SSVEP(TDCA)		
	Acc (%)	ITR (bits/min)	Acc (%)	ITR (bits/min)	ITR _{eff} (bits/min)	Acc (%)	ITR (bits/min)	ITR _{eff} (bits/min)
1	20.00	29.48	93.33	147.40	205.47	100.00	294.68	395.86
2	26.67	44.22	90.00	137.13	194.04	93.33	251.26	350.22
3	30.00	52.10	100.00	172.88	232.24	100.00	294.68	395.86
4	23.33	36.67	96.67	158.79	270.65	100.00	294.68	395.86
5	30.00	52.10	93.33	147.40	205.47	100.00	294.68	395.86
6	20.00	29.48	80.00	110.02	187.53	83.33	202.07	294.51
7	26.70	44.22	90.00	137.13	194.04	100.00	294.68	395.86
8	40.00	77.51	90.00	137.13	194.04	90.00	233.74	330.74
9	36.67	68.76	100.00	172.88	232.24	100.00	294.68	395.86
10	40.00	77.51	90.00	137.13	194.04	93.33	251.26	350.22
11	33.33	60.28	86.67	127.57	183.18	93.33	251.26	350.22
12	26.67	44.22	83.33	118.55	172.78	90.00	233.74	330.74
13	43.33	86.52	93.33	147.40	205.47	100.00	294.68	395.86
14	33.33	60.28	100.00	172.88	232.24	100.00	294.68	395.86
15	30.00	52.10	86.67	127.57	183.18	93.33	251.26	350.22
Avg	30.67	54.36	91.56	143.46	205.77	95.78	268.80	368.25
(Mean±SD)	±7.15	±17.44	±6.02	±19.42	±25.78	±5.27	±31.05	±33.43

Where N denotes the number of physical stimulus targets, P denotes the classification accuracy, and T denotes the average target selection time. The static-mapping baseline and UniSwarm-SSVEP both used a 1.5 s decision window; therefore, T was set to 1.5 s for standard ITR calculation.

Under the standard ITR metric, the static-mapping baseline achieved 54.36 ± 17.44 bits/min, whereas UniSwarm-SSVEP achieved 143.46 ± 19.42 bits/min. This result indicates that UniSwarm-SSVEP had higher information transfer capability when completing the same online control task.

To further describe the equivalent command capacity enabled by dynamic content mapping, we additionally calculated effective information transfer rate (ITR_{eff}) by replacing the physical target number N with the equivalent command number $N_{eq.}$ from Equation (5).

$$ITR_{eff} = \frac{60}{T_{eq.}} \left(\log_2 N_{eq.} + P \log_2 P + (1-P) \log_2 \frac{1-P}{N_{eq.}-1} \right) \quad (7)$$

For this supplementary metric, according to the experimental design described in Section 2.3, the adjusted selection time of UniSwarm-SSVEP was set to 1.875 s to account for the additional gaze shifts to vehicle-switching stimulus blocks. Under this supplementary metric, UniSwarm-SSVEP (FBCCA) achieved an effective ITR of 205.77 ± 25.78 bits/min, further reflecting the command-capacity gain provided by dynamic content mapping.

3.2 Algorithm-Enhanced Validation With TDCA

To further examine whether dynamic content mapping can be combined with a high-performance decoder, we applied TDCA to UniSwarm-SSVEP. TDCA improves class separability in short EEG windows through subject-specific templates and task-discriminant spatial

filtering, as described in Section 1.2. According to the offline analysis, TDCA reached its highest performance around 0.8-1.0 s, and the optimal decision window for subsequent online validation was further determined as 0.88 s.

As shown in Table 1, UniSwarm-SSVEP (TDCA) achieved an average accuracy of $95.78 \pm 5.27\%$, a standard ITR of 268.80 ± 31.05 bits/min, and an effective ITR of 368.25 ± 33.43 bits/min. Compared with UniSwarm-SSVEP (FBCCA), TDCA further improved both recognition accuracy and communication efficiency under the same dynamic mapping framework.

DISCUSSION

1. Core Advantages of Dynamic Content Mapping in Scalable BCI Systems

The core advantage of dynamic content mapping is that it expands the logical command space while maintaining a compact physical stimulus interface. In this study, the 20-target UniSwarm-SSVEP interface realized 153 equivalent commands in a 12-car task, but this value is not the upper limit. The same interface can theoretically support up to eight groups of seven cars, corresponding to a 56-car swarm and about 600 equivalent commands; adding more vehicle-switching commands could further expand this space. Compared with the multi-target SSVEP system reported in [14], UniSwarm-SSVEP also achieved higher accuracy, indicating that DCMM improves scalability without sacrificing recognition performance.

Under the same FBCCA decoder, UniSwarm-SSVEP outperformed the static-mapping baseline in both accuracy and standard ITR. These results indicate that DCMM reduces the visual and neural burden caused by displaying a large number of targets simultaneously,

while preserving efficient command selection.

2. Limitations and Future Work

Current limitations include the restriction to young adults and homogeneous car clusters, necessitating further validation across diverse ages, neurophysiological states, and heterogeneous robot types. Future work will optimize stimulus paradigms and decoding algorithms to enhance robustness, extending the system to schedule mixed swarms including cars, UAVs, and robotic manipulators.

CONCLUSION

This paper proposed UniSwarm-SSVEP, an SSVEP-BCI swarm robot control system based on dynamic content mapping. To validate the proposed dynamic content mapping mechanism, we constructed a 20-target UniSwarm-SSVEP test system for a 153-command swarm control task, involving single-vehicle control, subgroup coordination, and whole-swarm control. In the online comparison under the same FBCCA decoder, UniSwarm-SSVEP achieved an average accuracy of 91.56%, a standard ITR of 143.46 bits/min, and an effective ITR of 205.77 bits/min, outperforming the 153-target static-mapping baseline. The TDCA-based validation further improved the average accuracy to 95.78%, with a standard ITR of 268.80 bits/min and an effective ITR of 368.25 bits/min. These results indicate that UniSwarm-SSVEP provides a practical and scalable approach for efficient SSVEP-based swarm robot control.

REFERENCES

- [1] Mridha MF, Das SC, Ali MF, et al. Brain-Computer Interface: Advancement and Challenges. *Sensors* 2021;21(17): 5746. doi: 10.3390/s21175746
- [2] Wolpaw JR, Wolpaw EW. *Brain-Computer Interfaces: Principles and Practice*. Oxford University Press, 2012.
- [3] Zhang Y, Qian K, Xie SQ, et al. SSVEP-Based Brain-Computer Interface Controlled Robotic Platform With Velocity Modulation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2023;31: 3448-3458. doi: 10.1109/TNSRE.2023.3308778
- [4] Guger C, Schlögl A, Neuper C, Müller-Putz G, editors. *Proceedings of the 8th Graz Brain-Computer Interface Conference 2019*; 2019 Sep 17-20; Graz, Austria. Verlag der Technischen Universität Graz; 2019. doi: 10.3217/978-3-85125-682-6
- [5] Rosca SD, Leba M, Sibisanu RC. SSVEP Based BCI Control of a Robot Swarm. In: *World Conference on Information Systems and Technologies (WorldCIST)*; 2022. pp. 296-305. doi: 10.1007/978-3-031-04826-5_29
- [6] Huo Z, Deng T, Zhang L, et al. A BCI-Based Motion Control System for Heterogeneous Robot Swarm. In: *Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO)*; 2022. doi: 10.1109/ROBIO56267.2022.10016398
- [7] Deng T, Huo Z, Zhang L, et al. A VR-based BCI interactive system for UAV swarm control. *Biomedical Signal Processing and Control* 2023;86: 105065. doi: 10.1016/j.bspc.2023.105065
- [8] Jeong JH, Lee DH, Ahn HJ, Lee SW. Towards Brain-Computer Interfaces for Drone Swarm Control. *arXiv preprint arXiv: 2002.00519*, 2020.
- [9] Karavas GK, et al. A hybrid BMI for control of robotic swarms: Preliminary results. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; 2017. doi: 10.1109/IROS.2017.8206390
- [10] Zhu Y, Li Y, Jin J, et al. A Hybrid BCI Based on SSVEP and EOG for Robotic Arm Control. *Frontiers in Neurobotics* 2020;14: 583641. doi: 10.3389/fnbot.2020.583641
- [11] Xu Y, Ding C, Shu X, et al. Shared control of a robotic arm using non-invasive brain-computer interface and computer vision guidance. *Robotics and Autonomous Systems* 2019;115: 121-129. doi: 10.1016/j.robot.2019.02.014
- [12] Angrisani L, Arpaia P, Esposito A, Moccaldi N, Parvis M. A wearable SSVEP-based brain-computer interface with off-the-shelf components. In: *Proceedings of the 8th Graz Brain-Computer Interface Conference*; 2019. doi: 10.3217/978-3-85125-682-6-09
- [13] Gemblar F, Stawicki P, Volosyak I. How Many Electrodes are Needed for Multi-target SSVEP-BCI control: Exploring the Minimum number of signal Electrodes for CCA and MEC. In: *Proceedings of the 7th Graz Brain-Computer Interface Conference*; 2017. doi: 10.3217/978-3-85125-533-1-29
- [14] Gemblar F, Stawicki P, Volosyak I. Exploring the possibilities and limitations of multitarget SSVEP-based BCI applications. In: *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; 2016. pp. 1488-1491. doi: 10.1109/EMBC.2016.7590991
- [15] Gala R, Xu J, Wu Z, Nuernberger A. Decoding SSVEP on time and frequency domain using Convolutional Neural Network. In: *Proceedings of the 8th Graz Brain-Computer Interface Conference*; 2019. doi: 10.3217/978-3-85125-682-6-03
- [16] Liu B, Chen X, Shi N, Wang Y, Gao S, Gao X. Improving the Performance of Individually Calibrated SSVEP-BCI by Task-Discriminant Component Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2021;29: 1998-2007. doi: 10.1109/TNSRE.2021.3114340
- [17] Wong C M, Wang Z, Wang B, et al. Inter- and Intra-Subject Transfer Reduces Calibration Effort for High-Speed SSVEP-Based BCIs. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2020, 28(10): 2123-2135. doi: 10.1109/TNSRE.2020.3019276.

Influence of Median Nerve Stimulation Intensity on ERD/ERS, BCI Performance and User Experience

Loïc Bechon¹, Fabien Lotte¹, Stéphanie Fleck², Sébastien Rimbart¹

¹ Inria Center at Univ. Bordeaux / LaBRI, Talence, France

² University of Lorraine, Perseus, Metz, France
E-mail: loic.bechon@inria.fr

ABSTRACT: Median Nerve Stimulation (MNS) has recently emerged as a promising approach to enhance Motor Imagery (MI)-based Brain-Computer Interfaces (BCIs) by providing additional somatosensory input that modulates sensori-motor rhythms. While previous studies have shown that MNS interacts with MI-related activity and can improve BCI performance, the influence of stimulation intensity on the resulting cortical patterns remains largely unexplored. In this study, we investigate how different MNS intensities affect ERD/ERS patterns, offline classification performance, and user experience (UX) with these stimulations. Thirteen participants performed right-hand motor imagery while receiving three stimulation levels relative to their motor threshold: low (below threshold), medium (twitch-inducing), and high (above threshold). Results show similar ERD/ERS patterns across stimulation intensities and no significant differences in classification accuracy. Participants reported very low levels of discomfort associated with the MNS. These findings suggest that relatively low-intensity MNS is sufficient to elicit detectable sensorimotor responses while improving user comfort, supporting the design of more usable MNS-based BCI systems. This opens the possibility of using MNS at the lowest effective level, which is particularly beneficial from a user experience perspective for prolonged BCI interaction.

INTRODUCTION

Brain-Computer Interfaces (BCIs) enable users to interact with external systems by translating brain activity into control signals [1]. Among the various BCI paradigms, Motor Imagery (MI)-based BCIs are one of the most widely studied approaches. In MI-BCIs, users imagine performing movements without executing them, typically involving limb movements. Such mental tasks modulate sensorimotor rhythms over the motor cortex, producing Event-Related Desynchronization and Synchronization (ERD/ERS) patterns in the mu and beta frequency bands of the electroencephalographic (EEG) [2]. In particular, contralateral ERD during MI, followed by a post-task beta rebound (ERS), can be detected in EEG signals and used as discriminative features to infer the user's intention. Despite significant progress, MI-based BCIs still face important limitations. Classification accuracy often remains below practical usability thresholds (around

70–75% for two-class MI tasks), and performance varies considerably across users and sessions [3].

One promising approach to improve MI detection consists in combining MI tasks with external somatosensory stimulation. Various forms of stimulation have been explored, including tactile stimulation, functional electrical stimulation (FES), neuromuscular electrical stimulation (NMES), and tendon vibration [4, 5]. Among them, Median Nerve Stimulation (MNS) has recently emerged as a promising candidate. Electrical stimulation of the median nerve produces reproducible cortical responses characterized by an ERD during stimulation followed by a beta-band ERS. When MNS occurs during MI, this post-stimulation ERS is strongly reduced or suppressed, suggesting an interaction between sensory input and motor cortex activation that can improve MI detection. Previous studies reported classification improvements exceeding 12% when integrating MNS within MI-based paradigms, as compared to MI alone [6].

However, several key parameters of MNS-based BCI paradigms remain poorly understood. Previous work has investigated temporal aspects of stimulation, showing, for instance, that closely spaced MNS pulses can alter sensorimotor dynamics and lead to dominant ERD patterns due to overlapping cortical responses [7, 8]. More generally, studies on somatosensory evoked potentials (SEPs) have demonstrated that stimulation intensity strongly modulates cortical responses, affecting the amplitude, latency, and spatial distribution of early somatosensory components [9]. These findings suggest that stimulation intensity may also influence sensorimotor oscillations such as ERD/ERS. In practice, MNS intensity is commonly adjusted around the motor threshold, defined as the level producing a small involuntary movement between the thumb and index finger [10–12]. Yet it remains unclear whether such micromovements are necessary to evoke useful cortical patterns for BCI purposes, or whether lower intensities could produce comparable neural signatures. Conversely, higher intensities might enhance ERD/ERS responses but could also increase discomfort and reduce task concentration. From a human-computer interaction perspective, stimulation should ideally remain imperceptible while still eliciting sufficiently strong and detectable cortical responses.

In this study, we therefore investigate how median nerve stimulation intensity influences (i) ERD/ERS patterns

over the sensorimotor cortex, (ii) offline classification performance for discriminating MNS alone from MNS combined with motor imagery, and (iii) participants' perception of the stimulation. Our goal is to better understand how to calibrate stimulation intensity to design effective and usable MNS-based BCI systems.

MATERIALS AND METHODS

Participants:

Fifteen right-handed healthy volunteers (6 females, 9 males; mean age 23.56 ± 3.94 years) provided written informed consent and participated in this study, approved by Inria's ethical committee (the COERLE: approval number: 2025-67). Inclusion criteria required normal or corrected-to-normal vision and no neurological or psychiatric disorders. Participants sat in a chair with armrests for arm stabilization, facing a screen with experimental instructions. Prior to the session, they watched a short video explaining the motor imagery tasks. Two participants were excluded due to poor EEG signal quality, leaving thirteen in the final analysis.

Experimental Paradigm:

Each participant completed a single experimental session. After informed consent, a short calibration procedure was performed to determine the individual stimulation intensity. EEG and stimulation electrodes were then installed before the experiment. The protocol consisted of three blocks, each containing four runs of 40 trials. Three stimulation intensities were tested relative to the motor threshold: low ($T - 2$ mA), medium (*Threshold*), and high ($T + 2$ mA). As the low condition was set below the motor threshold, no visible twitch occurred. Stimulation intensity was fixed within each block following a Latin square design (Fig. 1) to counterbalance the order of conditions across participants (Low–Medium–High, Medium–High–Low, High–Low–Medium). During each trial, participants performed either right-hand motor imagery for 5 s with MNS delivered at 0.75 s or waited to receive only MNS, depending on the visual instruction displayed on the screen. A visual cue was presented only during MI trials and remained visible throughout the trial. Trials were randomized within runs, yielding 80 per task and condition. Short breaks (1–3 min) were provided between runs, and participants completed short questionnaires between blocks.

EEG Acquisition and MNS:

EEG signals were recorded using a 64-channel EEG system (ANT Neuro) with active wet electrodes positioned according to the international 10–20 system. Signals were acquired at 500 Hz and processed online using the OpenViBE platform (version 3.6.0). For subsequent analyses, the signals were downsampled to 128 Hz. MNS was delivered using a Micromed ENERGY LIGHT stimulator. Two stimulation electrodes (Ambu Neuroline 700) were placed at the wrist over the median nerve following standard positioning described in [6]. Electrical stimulation consisted of single pulses of 100 μ s duration. Before

the experiment, several test stimulations were delivered to determine the individual stimulation levels. Participants were asked to report their perceived sensation and confirm that the stimulation was neither painful nor uncomfortable. None of the participants reported excessive discomfort during this calibration procedure.

Time–Frequency and Topographic Analysis:

Offline EEG preprocessing, time–frequency analyses, and topographic analyses were performed using the EEGLAB toolbox (version 2025.0.0)[13] in Matlab 2020 (The MathWorks Inc., Natick, MA, USA). From the 64 recorded electrodes, a subset of 36 electrodes covering the sensorimotor areas was selected for further analysis (see Fig. 1). A common average reference (CAR) was computed on this subset of electrodes.

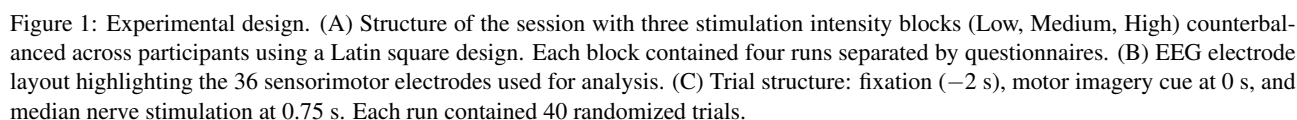
For time-frequency analyses, EEG signals were segmented into 7 s epochs ranging from -2 s to 5 s relative to task onset. Event-related spectral perturbations (ERSPs) were computed to visualize the time-frequency dynamics associated with the different stimulation conditions and to characterize ERD/ERS patterns during MNS alone and during MNS combined with motor imagery. Topographic analyses were performed in the 8–30 Hz frequency band and within the 750–2750 ms time window after task onset. In our paradigm, this window corresponds to the period encompassing the stimulation and the associated post-stimulation ERD/ERS activity. It was therefore selected to visualize the spatial distribution of task-related sensorimotor modulations across conditions.

Classification and Performance Evaluation:

Offline classification was performed using MNE-Python (v1.11.0) and Scikit-Learn (v1.8) on preprocessed EEG data exported from EEGLAB. The task consisted of discriminating between two conditions: MNS alone and MNS combined with right-hand motor imagery. Features were extracted from the 8–30 Hz frequency band within the 750–2750 ms time window after task onset, corresponding to the stimulation period and the associated sensorimotor response. Four classifiers were evaluated on 80 trials per class per condition, per participant with 4-fold cross-validation: Common Spatial Patterns (decomposed in 4 component) combined with Linear Discriminant Analysis(CSP+LDA), Minimum Distance to Riemannian Mean (MDM), and Tangent Space classifiers with logistic regression using either a Ledoit–Wolf (LWF_TS+LR) or Oracle Approximating (ORC_TS+LR) covariance estimator. Performance was evaluated using classification accuracy (see Tab. 1). Differences between stimulation intensity conditions and group were assessed using one-way Anova for global comparison, then paired Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons (see Fig. 4).

User Experience Assessment:

User experience and perception of the stimulation were assessed throughout the experiment. Right-handedness was verified using the Edinburgh Handedness Inventory (EHI) [14]. Participants' subjective state was evaluated before and after the session using the NExT question-



of ERD/ERS patterns remain qualitatively similar across Low, Medium, and High stimulation intensities. Statistical comparisons (right column in Fig. 2 and bottom row in Fig. 3) confirm that with permutation-based one-way ANOVA test and False Discovery Rate correction, no significant differences are observed between intensity conditions.

Classifier	Mean accuracy
CSP+LDA	0.739
MDM	0.697
LWF_TS+LR	0.750
ORC_TS+LR	0.754

ERD/ERS Patterns Across Stimulation Intensities: Figure 2 shows time-frequency dynamics observed at electrode C3 across stimulation intensities and task conditions. In the MNS-only condition, stimulation induces a typical sensorimotor pattern across all intensities, characterized by an initial ERS at stimulation onset followed by a short ERD in the mu band (7-13 Hz) and a post-stimulation beta rebound (15-30 Hz), consistent with previous reports of MNS-induced responses [6]. With MNS+MI, the initial ERS is reduced and replaced by a sustained ERD lasting for most of the task duration, and the beta rebound is largely suppressed, reflecting the dominance of MI-related desynchronization. These temporal dynamics are consistent with the spatial patterns shown in Fig. 3. In the MNS-only condition, a central ERS is mainly visible around the Cz region, whereas the MNS+MI condition produces a bilateral ERD over sensorimotor areas centered around C3 and C4. Importantly, both the temporal dynamics and the spatial distribution

Classification Results: Classification performance obtained with the different classifiers is summarized in Tab. 1. Among the evaluated methods, the Tangent Space classifier with logistic regression using the Oracle Approximating Shrinkage covariance estimator (ORC_TS+LR) achieved the highest average performance (mean accuracy = 0.754, SD = 0.090). Although this value is approximately 7% lower than that reported in previous MNS-based BCI studies [6], tangent space classifiers remained the most effective overall. Pairwise comparisons between classifiers using Wilcoxon signed-rank tests with Bonferroni correction showed that MDM performed significantly worse than CSP+LDA ($p < 0.01$), LWF_TS+LR ($p < 0.01$), and ORC_TS+LR ($p < 0.01$), while no significant differences were observed among the other classifiers. Based on these results, subsequent analyses focused on the ORC_TS+LR classifier. We further investigated whether classification performance was influenced by the order of stimulation conditions introduced by the Latin square design. A one-way ANOVA revealed no significant differences between groups corresponding to the three possible condition orders ($p =$

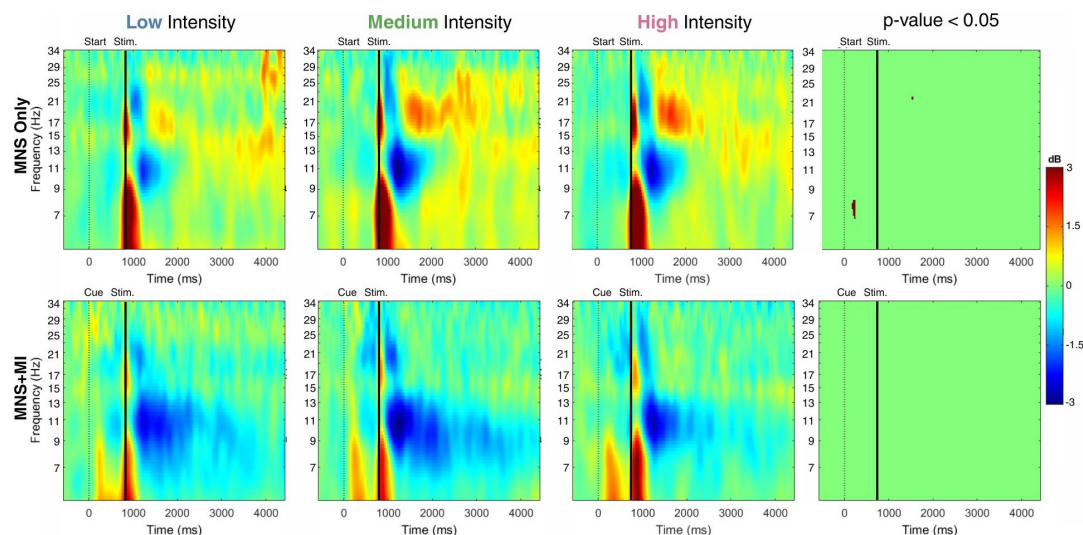


Figure 2: Time-frequency representations at C3 for three stimulation intensities (top: MNS-only; bottom: MNS+MI). Dashed line: cue onset; solid line: stimulation. MNS induces ERS, mu-band ERD (7–13 Hz), and beta rebound (15–30 Hz); during MNS+MI, the rebound is suppressed and replaced by sustained ERD. No significant differences across intensities (right column, $p < 0.05$).

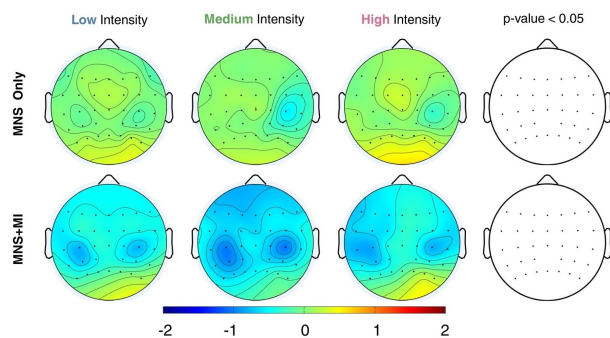


Figure 3: Topographic maps of ERD/ERS (%) in the mu+beta band (1–3 s after task onset) for each stimulation intensity and task. Red indicates ERS and blue ERD. Electrodes highlighted in red in the right panel denote significant differences between conditions ($p < 0.05$).

0.392), nor between stimulation intensity conditions ($p = 0.814$). These results indicate that classification accuracy was not affected by stimulation intensity or by the order in which the intensity conditions were presented.

Additional analyses were conducted using a broader 5–35 Hz frequency band, motivated by previous studies suggesting that including lower-frequency activity, particularly in the theta band, may improve MNS-based MI detection [19]. Figure 4 presents classification accuracy across stimulation intensities for the ORC_TS+LR classifier in both frequency bands. In the conventional 8–30 Hz band, no significant differences were observed between stimulation intensities. In contrast, the 5–35 Hz analysis revealed a small but significant difference between the Low and Medium conditions, while all other comparisons remained non-significant. Although limited, this effect suggests that including lower-frequency components may reveal subtle intensity-related differences not captured by the standard sensorimotor band.

Subjective User Experience Assessment:

Participants reported low to moderate fatigue before the session (FIFS: mean = 4.15, SD = 1.99), which increased significantly by the end of the experiment (mean = 6.61, SD = 2.29; $p = 0.002$; 10-item scale), likely due to the sustained attention and immobility required during the protocol. Emotional state remained globally stable: NEXT results showed a decrease in depression scores ($p = 0.005$), while meCUE emotional scores indicated neither strong negative (mean = 3.38, SD = 0.89; 5-item scale) nor strong positive emotions (mean = 3.92, SD = 0.75). UX-BCI responses indicated moderate difficulty in performing the MI task (Activity: mean = 2.56, SD = 0.78; Kinesthetic MI: mean = 3.38, SD = 0.82), consistent with typical MI-based BCI studies [20]. Overall, MNS was well tolerated across intensity conditions. Participants reported a generally neutral perception of the stimulation (mean = 2.94, SD = 0.46; 5-item scale), commonly described as tingling or vibration-like sensations in the hand. Reported discomfort remained very low (mean = 0.35, SD = 0.46; 10-item scale), and no noticeable differences were observed between stimulation intensities. The urge to move occasionally reported during stimulation remained moderate (mean = 3.30, SD = 1.12) and did not lead to overt movements. These results indicate that the different stimulation intensities were similarly well tolerated and did not notably affect subjective perception.

DISCUSSION

This study investigated the influence of median nerve stimulation (MNS) intensity on sensorimotor EEG responses and motor imagery (MI) detection performance. Similar ERD/ERS patterns were observed across the three stimulation intensities, with no significant differences in either ERSP analyses or classification accuracy. One possible explanation is that the tested intensity range

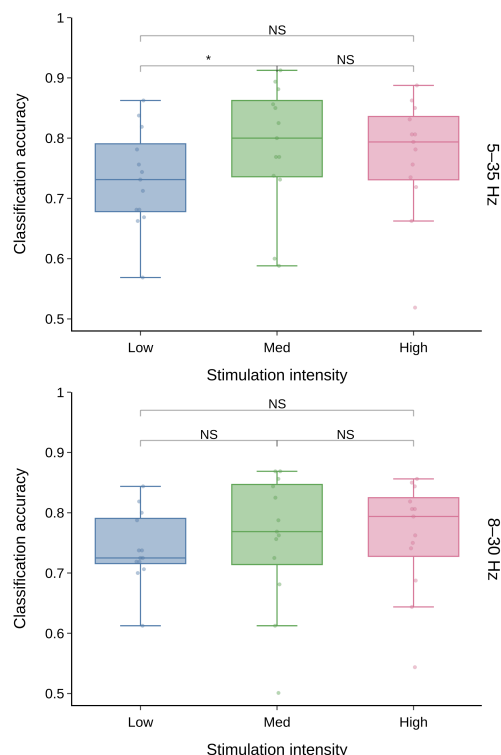


Figure 4: Offline classification accuracy for ORC_TS+LR classifier across stimulation intensities (Low, Med, High) for the 5-35 Hz and 8-30 Hz bands. Points represent individual participants. A significant difference was observed between Low and Med in the 5-35 Hz band ($*p < 0.05$), while other comparisons were not significant.

(± 2 mA around the motor threshold) was insufficient to induce measurable cortical differences. Once afferent pathways are sufficiently activated, increasing stimulation intensity may not substantially modify the resulting cortical oscillatory responses. In other words, stimulation levels near the motor threshold may already recruit the neural mechanisms underlying the observed sensorimotor activity, leading to comparable ERD/ERS patterns across the tested intensity range. Nevertheless, the stability of the observed patterns suggests that MI-related responses remain robust within this range, indicating that relatively low stimulation intensities may already be sufficient to elicit detectable sensorimotor activity. User questionnaires further indicated that the protocol was well tolerated and elicited subjective responses comparable to those of conventional MI-BCI paradigms, suggesting that repeated MNS remains acceptable for participants and compatible with practical BCI applications.

Comparison with Previous MNS-Based BCI Studies:

A previous work reported accuracy improvements of about +12% when combining MNS with MI compared to MI alone [6], and in some datasets, gains exceeding 18% have been observed with mean accuracies around 85% in MI vs Rest paradigms [21]. In contrast, the best performance in the present study reached approximately 80% accuracy (5–35 Hz band, medium intensity; Fig. 4). This

difference is somewhat surprising, as time–frequency and topographical analyses still show a clear gating effect when MNS is combined with MI: the post-stimulation beta rebound is strongly reduced during MI, likely due to dominant ERD activity [6, 8]. Methodological factors may explain this discrepancy. In the present study, participants performed the task with eyes open, whereas previous studies were conducted with eyes closed, a condition known to modulate sensorimotor rhythms [22]. Participants were also seated, while previous experiments often involved a lying position in clinical monitoring contexts [6]. In addition, the present protocol did not include a baseline MI vs Rest condition without stimulation, which would have allowed a direct estimation of the performance gain induced by MNS.

Interestingly, classification results suggest that the broader 5–35 Hz frequency band slightly improves performance compared to the classical sensorimotor band (8–30 Hz). This indicates that lower-frequency components, including the theta band, may contribute to MI detection in MNS-based paradigms, though the contribution of the 30–35 Hz range remains to be investigated. Topographical analyses suggest that a central and fronto-central rebound observed in the MNS-only condition becomes more prominent when theta frequencies are included, thereby increasing the contrast with the MNS+MI condition. This is consistent with previous findings reporting that MNS can modulate low-frequency oscillatory activity beyond classical mu and beta rhythms [19].

Exploratory Analysis of Stimulation Intensity:

A complementary analysis examined whether the absolute stimulation intensity influenced classification performance across participants. Spearman correlation revealed a weak but significant positive relationship between stimulation intensity and ORC_TS+LR classifier accuracy in the 8–30 Hz band ($p = 0.37$, $p = 0.019$), while no significant correlation was observed in the broader 5–35 Hz band ($p = 0.21$, $p = 0.207$). Although this suggests that higher stimulation amplitudes may slightly enhance sensorimotor discriminability in the narrower band, this effect was not reflected in comparisons between the three discrete stimulation conditions. One possible explanation is that once afferent pathways are sufficiently activated, increasing stimulation intensity may not substantially modify the resulting cortical oscillatory responses.

Influence of Analysis Window:

Another possible explanation for the absence of clear differences between stimulation intensities lies in the use of a fixed analysis window for ERSP and classification analyses (8–30 Hz, 750–2750 ms). While this approach ensures consistent comparisons across participants, it may mask individual variability in the temporal or spectral characteristics of sensorimotor responses. Future analyses could therefore explore more individualized approaches, for example by extracting participant-specific ERD minima or optimizing subject-specific time–frequency features for classification, which may better capture inter-individual variability in sensori-

motor responses.

CONCLUSION

This study investigated the influence of median nerve stimulation (MNS) intensity on sensorimotor EEG responses and motor imagery (MI) detection performance. Similar ERD/ERS patterns were observed across stimulation intensities below, at, and above the motor threshold, and classification accuracy was not significantly affected by stimulation intensity. These results indicate that stimulation levels around or below the motor threshold are sufficient to elicit robust sensorimotor responses for MNS-based MI detection. Importantly, participants reported very low discomfort, suggesting that such low-intensity stimulation remains well tolerated during prolonged sessions. Overall, these findings suggest that MNS-based BCI paradigms can operate effectively with minimal stimulation levels, opening the way to more lightweight, comfortable, and user-friendly systems. Future work should investigate subject-specific optimization of stimulation parameters and analysis windows to better capture inter-individual variability and further improve MI detection performance.

ACKNOWLEDGMENTS

This work was supported by ANR projects BCI4IA (ANR-22-CE19-0016-03) and STIM-BCI (ANR-24-CE33-3736).

REFERENCES

- [1] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*. 2002.
- [2] Neuper C, Wörtz M, Pfurtscheller G. ERD/ERS patterns reflecting sensorimotor activation and deactivation. In: *Progress in Brain Research*. Elsevier, 2006, 211–222.
- [3] Lotte F et al. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering*. 2018;15(3):031005.
- [4] Yao L, Jiang N, Mrachacz-Kersting N, Zhu X, Farina D, Wang Y. Reducing the Calibration Time in Somatosensory BCI by Using Tactile ERD. *IEEE Trans Neural Syst Rehab*. 2022;30:1870–1876.
- [5] Yi W et al. Enhancing performance of a motor imagery based brain-computer interface by incorporating electrical stimulation-induced SSSEP. *Journal of Neural Engineering*. 2017;14(2):026002.
- [6] Rimbart S, Riff P, Gayraud N, Schmartz D, Bougrain L. Median Nerve Stimulation Based BCI: A New Approach to Detect Intraoperative Awareness During General Anesthesia. *Frontiers in Neuroscience*. 2019;13.
- [7] Houlgreave MS, Morera Maiquez B, Brookes MJ, Jackson SR. The oscillatory effects of rhythmic median nerve stimulation. *NeuroImage*. 2022;251:118990.
- [8] Rimbart S. Investigating the overlap of ERD/ERS modulations for a Median Nerve Stimulation-based BCI. In: *5th International Neuroergonomics Conference 2024*. Bordeaux, France, Jul. 2024.
- [9] Gerber J, Meinck HM. The effect of changing stimulus intensities on median nerve somatosensory-evoked potentials. *Electromyography and clinical neurophysiology*. 2000;40 8:477–82.
- [10] Salenius S, Portin K, Kajola M, Salmelin R, Hari R. Cortical control of human motoneuron firing during isometric contraction. *Journal of Neurophysiology*. 1997;77(6):3401–3405.
- [11] Backes W, Mess W, van Kranen-Mastenbroek V, Reulen J. Somatosensory cortex responses to median nerve stimulation: Fmri effects of current amplitude and selective attention. *Clinical Neurophysiology*. 2000;111(10):1738–1744.
- [12] Cebolla AM et al. Movement gating of beta/gamma oscillations involved in the n30 somatosensory evoked potential. *Hum Brain Map*. 2009;30(5):1568–1579.
- [13] Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*. 2004;134(1):9–21.
- [14] Oldfield R. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*. 1971;9(1):97–113.
- [15] Bismuth J, Vialatte F, Lefaucheur JP. Relieving peripheral neuropathic pain by increasing the power-ratio of low- over high- activities in the central cortical region with EEG-based neurofeedback: Study protocol for a controlled pilot trial (SMRPain study). *Neurophysiologie Clinique*. 2020;50(1):5–20.
- [16] Davis MP et al. Four-Item Fatigue Screen: Replacing the Brief Fatigue Index. *American Journal of Hospice and Palliative Medicine*. 2013;30(7):652–656.
- [17] Linn S. The injury severity score—Importance and uses. *Annals of Epidemiology*. 1995;5(6):440–446.
- [18] Minge M, Thüning M, Wagner I, Kuhr CV. The meCUE Questionnaire: A Modular Tool for Measuring User Experience. In: *Advances in Ergonomics Modeling, Usability & Special Populations*. Cham, 2017, 115–128.
- [19] Avilov O, Rimbart S, Popov AO, Bougrain L. Optimizing motor intention detection with deep learning: Towards management of intraoperative awareness. *IEEE Trans Biomed Eng*. 2021;68:3087–3097.
- [20] Yousefi M, Frey J, Rimbart S, Altamira GH, Fleck S. Sense-It: Towards new types of BCI-based sensorimotor neurofeedback for motor rehabilitation. In: *Proc. IEEE SMC*. 2025, 6891–6898.
- [21] Rimbart S, Guerci P, Gayraud N, Meistelman C, Bougrain L. Innovative brain-computer interface based on motor cortex activity to detect accidental awareness during general anesthesia. In: *Proc. IEEE SMC*. 2019.
- [22] Rimbart S, Al-Chwa R, Zaepffel M, Bougrain L. Electroencephalographic modulations during an open- or closed-eyes motor task. *PeerJ*. 2018;6:e4492.

Mission Impossible? Inhibiting Frontal Midline Theta with Neurofeedback

M. Pfeiffer¹, L. Botrel¹, A. Kübler¹

¹ Institute of Psychology, University of Würzburg, Marcusstr. 9-11, Würzburg D-97070, Germany

E-mail: maria.pfeiffer@uni-wuerzburg.de

ABSTRACT: Frontal midline theta (FMT) is an oscillation closely linked to cognitive control and performance monitoring. A recent meta-analysis suggests that FMT can be upregulated via neurofeedback (NF), yet successful downregulation remains largely unsubstantiated and has primarily been tested in single-session designs.

We tested whether multi-session training enables FMT inhibition via NF. Five participants completed six NF sessions targeting FMT reduction (4-8 Hz at Fz). Contrary to the training goal, FMT increased during NF relative to the pre-session resting-state baseline. A similar pattern was observed in a previous pilot study using individualized FMT peak-frequency NF. Together, these findings suggest that traditional NF approaches may face fundamental difficulties in achieving reliable FMT downregulation.

INTRODUCTION

Neurofeedback (NF) is realised through a closed-loop brain-computer interface (BCI) which provides participants with real-time information about their brain activity and thereby allows them to learn its intentional modulation [1]. Early demonstrations of electroencephalography (EEG) NF already showed that oscillatory activity can be shaped through operant conditioning principles [2]. NF is a broadly used method in cognitive and clinical neuroscience, particularly with EEG due to its high temporal resolution and practical feasibility in repeated-session training paradigms [1].

Frontal Midline Theta (FMT; 4–8 Hz) is a well-established electrophysiological marker of the need for cognitive control. Increases in FMT are consistently observed during tasks requiring sustained attention, conflict monitoring, error detection, and feedback-based learning [3,4]. FMT is thought to reflect the recruitment of top-down control processes in situations demanding behavioural adjustment. It is also involved in memory formation and retrieval [5] and increases during meditation [6] or the experience of a flow state [7]. Converging evidence implicates the midcingulate cortex (MCC; often referred to as dorsal anterior cingulate cortex, dACC) as a primary generator of FMT, further linking the rhythm to reinforcement learning and adaptive control mechanisms [8,9].

Neurofeedback of Frontal Midline Theta has received increasing attention as a target to enhance performance of memory or other executive functions [10]. Consistent

with the positive link between FMT and cognitive performance, most NF studies have aimed at FMT upregulation for enhanced performance in task-switching, episodic and working memory tasks or behavioural inhibition [11].

Although theta downregulation has been investigated in other NF contexts, most prominently theta/beta protocols for ADHD, these protocols typically target theta/beta ratio changes at central or frontocentral sites rather than FMT as a specific cognitive-control rhythm [12]. Attempts to inhibit FMT via NF have received considerably less attention despite potential relevance for domains where excessive control or conflict-related processing may be maladaptive (e.g., anxiety-related hypervigilance or persistent conflict processing). In our recent meta-analysis, we showed that only a few studies aimed at decreasing FMT, either as a control group or to improve golf putting performance. We found that evidence for FMT upregulation by NF is more consistent than evidence for FMT downregulation [11].

Reducing Frontal Midline Theta with Neurofeedback: While we identified multiple studies aiming at FMT reduction, only one reported a clear reduction of FMT [13] and this emerged only as a first-to-last run decrease. When NF was compared to the preceding baseline, FMT increased, suggesting that apparent reductions may reflect a gradual attenuation of an initial NF-induced FMT rise. However, all included studies were single-session protocols and none used more narrow, individualized frequency bands. We therefore hypothesized that extended training combined with more precise frequency targeting might enable FMT downregulation. Hence, we implemented a multi-session NF protocol targeting FMT reduction motivated by the broader aim of reducing excessive conflict-related processing in approach–avoidance contexts.

As multiple successful FMT modulation protocols focused on individual peak training, we first implemented a similar approach (pilot 1). Participants performed a task designed to elicit approach-avoidance conflicts and hence FMT [14]. From this task an individual peak frequency was derived. NF was then delivered within a narrow 2 Hz band surrounding this peak frequency. However, we found that the peak frequency was not stable across sessions [15], rendering the provided feedback band unreliable despite its

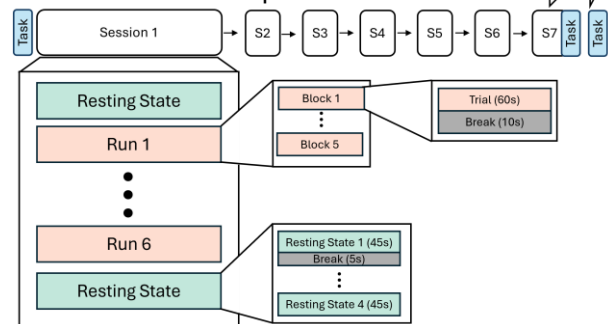


Figure 1: Timeline of pilot 1: Seven sessions were scheduled over the course of two weeks. Each session started and ended with four 45 second recordings of resting-state, separated by 5 seconds breaks. The NF training consisted of six runs, interleaved by self-paced breaks. Each run contained five blocks of 60 seconds NF trial and 10 second break. An approach–avoidance task that reliably elicits FMT bursts was conducted before the first session, immediately after the final NF session, and again seven days later.

narrower bandwidth. NF modulation of FMT was not achieved. We therefore conducted a second pilot study using a conventional broadband FMT target (4–8 Hz) rather than individualized peak frequencies. Yet this approach likewise failed to produce reliable FMT downregulation.

We here present data from both pilot studies. In each study, five participants trained to reduce their FMT in multiple NF sessions distributed over the course of two weeks. Together, these two pilot studies provide converging evidence regarding the difficulty of achieving FMT downregulation using conventional NF approaches. Only data from the task used to determine peak frequency in the first pilot study were reported previously [15]. The NF data were not reported at the time because the lack of successful modulation was initially attributed to unstable peak frequency estimates, which may have resulted in an unreliable individualized feedback signal.

MATERIALS AND METHODS

Participants: Five participants were recruited for each pilot study via the university’s online portal (pilot 1: three female, age: $M = 24.4$ years, $SD = 1.5$; pilot 2: four female; age $M = 25$ years, $SD = 2.8$). All participants provided written informed consent, were ≥ 18 years old, right-handed, non-color blind, without a history of psychiatric disorder, and naïve to NF. Participants received either course credit or monetary compensation (12.50€ per hour). All procedures were carried out in accordance with the ethical standards of the Association of German Professional Psychologists (Deutsche Gesellschaft für Psychologie, DGPs) and the Declaration of Helsinki, and were approved by the Ethics Committee of the Institute of Psychology, Julius-Maximilians-University Würzburg (GZEK 2023-45).

Study Procedure: An overview of the session structure for pilot 1 with the individualized frequency is shown in Fig. 1 and for pilot 2 with the 4-8 Hz band in

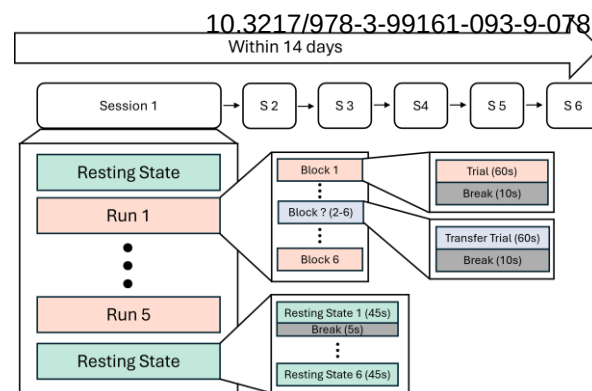


Figure 2: Timeline of pilot 2: Six sessions were scheduled over the course of two weeks. Each session started and ended with six 45 second recordings of resting-state, which were interleaved by 5 seconds breaks. The NF training consisted of five runs, interleaved by self-paced breaks. Each run contained six blocks of 60 seconds NF trial and 10 second break, during each run, one of the trials was a transfer trial, which did not contain any feedback.

Fig. 2. Participants completed seven NF sessions within 14 days, scheduled at approximately the same time of day. Each session began and ended with a three-minute resting-state recording. During resting-state, participants watched an animation of a slowly rotating 3D sphere. Pilot 2 was structured identically but shortened to six sessions containing 5 runs and had extended 5 minutes resting state measurements. To determine the individual FMT peak frequency and assess potential changes during conflict processing, participants performed an approach–avoidance task [14] at three time points: before NF training, immediately after the final session, and at a seven-day follow-up (only for pilot 1). Specifics on how the peak frequency was obtained are described in detail in [15].

Each session comprised six NF runs (pilot 2: five runs). Each run consisted of five trials lasting one minute each, separated by 10-second breaks. In pilot 2, each run comprised one additional, transfer trial, in which no feedback was presented. During transfer trials participants were instructed to continue applying their current mental strategy to modulate brain activity. Transfer trial positions were pseudo-randomized such that each trial position (except the first) served once as transfer; the first trial of a run was never assigned as transfer. Between runs, participants could take breaks of self-determined length and were asked to write down their thoughts and strategies during the preceding run.

Neurofeedback: Visual feedback was presented on a black screen using three bar charts. A wide central dark-cyan bar provided real-time feedback of FMT activity; a thinner pink bar indicated artifacts in the EEG in real time; and a thin light-cyan bar on the left provided continuous feedback of correct modulation across each trial. Participants were instructed to sit still, keep their eyes open, face the monitor, and attend to the bars. They were asked to increase the height of the dark-cyan bar via their thoughts and to minimize the pink artifact bar by relaxing and reducing muscle activity. No specific mental strategy instructions were provided; participants

EEG acquisition: EEG was recorded using ActiCap electrodes and BrainAmp amplifiers (Brain Products GmbH, Gilching, Germany). Two electrodes (O1/O2) were repurposed for the electrooculogram (EOG) recording (right eye), and the remaining 30 electrodes were placed according to the international 10–20 system. The reference electrode was placed at FCz and the ground at FPz. Data were streamed and recorded via LabStreamingLayer (LSL-Connector/LabRecorder) at 250 Hz.

Online signal processing and feedback computation: Online processing relied on a Python-based NF framework. The pre-NF resting-state was used to calibrate the system by deriving (i) artifact thresholds for EEG and EOG channels, (ii) an Artifact Subspace Reconstruction (ASR) model, and (iii) the baseline distribution of FMT activity used for z-normalization. For calibration and online processing, signals were re-referenced to a common average reference (CAR) and detrended (DC offset removal). Artifact thresholds were determined as the smallest absolute value at which the z-score exceeded 2.58, constrained to 80–180 μ V. Data were then bandpass-filtered (1–40 Hz) and used to fit ASR for online artifact correction. PSD was computed using Welch’s method. Mean and standard deviation of band power in a 2-Hz band surrounding the peak frequency at Fz were stored as the baseline distribution (pilot 2: band power between 4–8 Hz).

During NF trials, FMT was computed from a 2-second sliding window updated every 100 ms. The stream was CAR-referenced, detrended, bandpass-filtered (1–40 Hz), corrected using the pre-trained ASR, and transformed to PSD with Welch’s method. Power in the respective band (pilot 1: individualized 2 Hz, pilot 2: 4–8 Hz) at Fz was extracted and normalized to baseline z-scores. These z-scores were mapped to the central dark-cyan feedback bar with an inverted y-axis such that lower z-scores (FMT desynchronization) produced a higher bar; bar limits were set to -2 and $+2$.

Offline processing and FMT quantification: For offline analyses, EEG was re-referenced to CAR and bandpass-filtered (1–100 Hz; 8th-order Butterworth IIR). Continuous data were segmented into 5-second epochs. Artifacts were addressed using ICA (infomax; max 500 iterations). Components explaining 99% of the variance were retained and labelled via ICLabel; non-brain components were removed if classified with $\geq 70\%$ certainty. Remaining epochs exceeding 150 μ V were discarded. Data were then transformed using surface Laplacian (current source density) and bandpass-filtered again (1–40 Hz; 8th-order Butterworth IIR). PSD at Fz was computed using Welch’s method (Hamming window, 50% overlap), and theta power (4–8 Hz for both studies) was extracted as the FMT measure.

Statistical analysis: FMT power was log-transformed and averaged within participants across runs of each session for each condition (resting-state baseline, NF, transfer). To avoid carry-over effects of the NF, only the

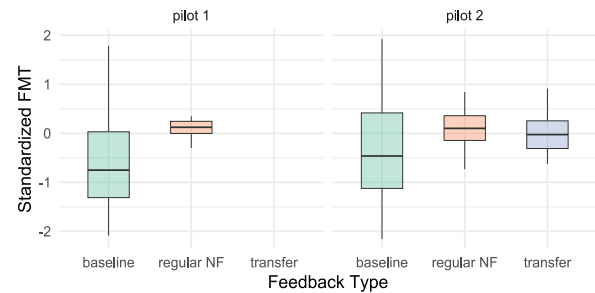


Figure 3: Frontal midline theta (FMT) activity for both datasets across the three task conditions. Values were standardized within each participant and session relative to the pre-NF resting-state baseline. Model estimates indicated higher FMT in the NF condition compared to baseline for both datasets. To avoid overinterpretation, statistical contrasts are not marked in the figure.

resting-state recording before NF served as baseline. For both pilots, we tested the effect of condition with three linear mixed-effects models. Given the small sample size of both datasets, we refrained from a stepwise parameter selection approach for the mixed-model and employed a theoretically motivated model structure with condition as fixed effect and participant as a random intercept: $FMT \sim \text{condition} + (1 | \text{participant})$.

The first two models examined condition differences relative to resting-state baseline (reference level) in the two different datasets. The third model analysed baseline-adjusted FMT values, using the regular NF condition as reference, to assess differences between feedback and transfer trials while reducing between-session variability. All statistical analyses were conducted using R [16] and the glmmTMB package [17].

RESULTS

All recordings were processed with the offline pipeline described above. One transfer trial from a single participant of pilot 2 was excluded due to poor data quality; all remaining runs were retained for analysis.

Participant-standardized condition means are shown in Fig. 3. Given the pilot nature of both studies, statistical results should be interpreted as exploratory and illustrative rather than confirmatory. No participant showed consistent decreases below baseline across sessions. Instead, FMT values during NF were generally elevated relative to baseline throughout the training period.

Reported Strategies: In both pilots, participants reported using diverse strategies (e.g., positive thoughts, recalling the day, imagining music, breath control), but frequent within-run switching and the small sample size prevented evaluating their effects on FMT regulation.

FMT increases relative to baseline:

The linear mixed-effect model was first fitted to log-transformed FMT values averaged within participants across runs. The resting-state baseline served as the reference level. In both datasets FMT activity was higher

Proceedings of the 10th Graz Brain-Computer Interface Conference 2026
 condition than during the baseline ($\beta = 0.14$, $SE = 0.05$, $z = 3.07$, $p < 0.01$). Similar increases were observed in the transfer trials of pilot 2 ($\beta = 0.10$, $SE = 0.05$, $z = 2.32$, $p < 0.05$). Both model fits included a random intercept for participants (pilot 1: $\sigma^2 = 0.26$; pilot 2: $\sigma^2 = 0.57$), indicating substantial between-participant variability. While both models explained a large proportion of total variance (conditional R^2 : pilot 1 = 71.1%; pilot 2 = 91.8%), the variance attributable to condition alone was small (both marginal $R^2 < 1\%$), suggesting that most explained variance reflected between-subject variability independent of condition.

FMT increases differ between feedback conditions:
 For pilot 2, the model was fitted to FMT values adjusted relative to the pre-NF resting-state baseline of the respective day to reduce inter-individual and inter-session variability. The NF condition served as the reference level. Relative to the regular NF, FMT activity was only slightly lower during the transfer condition ($\beta = -0.04$, $SE = 0.03$, $z = -1.3$, $p = 0.19$) but more so during resting-state ($\beta = -0.14$, $SE = 0.05$, $z = -3.05$, $p < 0.01$). This model explained less total variance (conditional $R^2 = 33.0\%$) but a slightly larger proportion was attributable to condition (marginal $R^2 = 1.9\%$).

DISCUSSION

The present study examined whether FMT can be reliably downregulated using multi-session NF. Across two pilot studies employing different targeting approaches, we found no evidence for successful FMT downregulation. Instead, FMT activity increased relative to resting-state baseline during NF in both protocols. In pilot 1, feedback was based on an individualized peak-frequency band derived from a task chosen to elicit FMT bursts. Individualized frequency targeting has been proposed as a potentially more precise approach for oscillatory NF. However, we observed that peak frequencies were not stable across sessions, consistent with previous findings [15]. This instability likely reduces the reliability of individualized feedback bands and may partly explain the lack of modulation observed in this pilot. In pilot 2, we therefore reverted to a conventional broadband theta range (4–8 Hz). Despite this simplification, the pattern remained unchanged: FMT activity during NF was again elevated relative to baseline and no consistent decreases were observed across participants.

Our results are not only consistent with the reported negative effect size computed from studies attempting downregulation [11] but also with a most recent study which reported FMT decreases only after, but not during the NF training [18]. This pattern suggests that increases in FMT during NF may be a recurring phenomenon rather than a study-specific artifact.

While the modulation of certain oscillatory rhythms via NF has been demonstrated in various contexts, rhythms that are tightly coupled to cognitive control processes may be more difficult to voluntarily suppress within the typical NF paradigm.

10.3217/978-3-89161-093-9_078
 However, the studies presented here included only small samples ($N=5$ participants per pilot) and thus do not allow for generalization. Furthermore, NF studies are highly heterogeneous in their design, including differences in feedback modality, session number, and signal processing pipelines. While the present protocols attempted to address some of these factors (e.g., individualized frequency targeting in Pilot 1 and multi-session training in both pilots), other design variations may yield different outcomes. Future studies with larger samples and systematically varied NF parameters (e.g. longer training periods) will be necessary to determine under which conditions FMT modulation may be achievable.

Despite these limitations, the present work highlights the importance of reporting negative or nil findings in NF research [19]. Since multi-session NF studies are resource-intensive, meaningful sample sizes are often only achieved through the aggregation of studies by meta-analysis. However, the absence of reported failures may bias the literature toward overly optimistic conclusions regarding modulation success. Transparent reporting of unsuccessful protocols is therefore essential for the cumulative evaluation of NF methodologies and growth of the field [20].

CONCLUSION

Across two pilots we found no evidence for successful downregulation of FMT. Contrary to the training goal, however consistent with previous reports, FMT activity was consistently elevated during NF relative to resting-state baseline. We highlight a potential challenge when attempting to modulate neural rhythms that are closely linked to cognitive control processes. Future research should be directed at the systematic evaluation of the task-specific effects of NF and their interference with specific targets.

REFERENCES

- [1] Sitaram R, Ros T, Stoeckel L, Haller S, Scharnowski F, Lewis-Peacock J, et al. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience* 2017;18(2):86–100.
- [2] Kamiya J. Conditioned discrimination of the EEG alpha rhythm in humans. *Proceedings of the Western Psychological Association* 1962.
- [3] Cavanagh JF, Frank MJ, Klein TJ, Allen JJB. Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage* 2010;49(4):3198–209.
- [4] Cavanagh JF, Frank MJ. Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences* 2014;18(8):414–21.
- [5] Hsieh L-T, Ranganath C. Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. *NeuroImage* 2014;85:721–9.
- [6] Cahn BR, Polich J. Meditation states and traits: EEG, ERP, and neuroimaging studies. *Psychological Bulletin* 2006;132(2):180–211.
- [7] Katahira K, Yamazaki Y, Yamaoka C, Ozaki H, Nakagawa S, Nagata N. EEG Correlates of the

- Alpha Rhythm in the Mental Arithmetic Task. *Front Psychol* 2018;9:300.
- [8] Hayden BY, Platt ML. Neurons in Anterior Cingulate Cortex Multiplex Information about Reward and Action. *J Neurosci* 2010;30(9):3339–46.
- [9] Vogt BA. Midcingulate cortex: Structure, connections, homologies, functions and diseases. *Journal of Chemical Neuroanatomy* 2016;74:28–46.
- [10] Yeh W-H, Ju Y-J, Liu Y-T, Wang T-Y. Systematic Review and Meta-Analysis on the Effects of Neurofeedback Training of Theta Activity on Working Memory and Episodic Memory in Healthy Population. *IJERPH* 2022;19(17):11037.
- [11] Pfeiffer M, Kübler A, Hilger K. Modulation of Human Frontal Midline Theta by Neurofeedback: A Systematic Review and Quantitative Meta-Analysis. *Neuroscience & Biobehavioral Reviews* 2024:105696.
- [12] Van Doren J, Arns M, Heinrich H, Vollebregt MA, Strehl U, K. Loo S. Sustained effects of neurofeedback in ADHD: a systematic review and meta-analysis. *European Child & Adolescent Psychiatry* 2019;28(3):293–305.
- [13] Chen TT, Wang KP, Chang WH, Kao CW, Hung TM. Effects of the function-specific instruction approach to neurofeedback training on frontal midline theta waves and golf putting performance. *Psychology of Sport and Exercise* 2022;61(December 2021):102211.
- [14] Rodrigues J, Müller M, Mühlberger A, Hewig J. Mind the movement: Frontal asymmetry stands for behavioral motivation, bilateral frontal activation for behavior. *Psychophysiology* 2018;55(1):1–19.
- [15] Pfeiffer M, Masson E, Kübler A, Rodrigues J. Reliability of individual task-related frontal-midline-theta frequency for neurofeedback training –exploratory investigations. *Proceedings of the 9th Graz Brain-Computer Interface Conference* 2024.
- [16] R Core Team. R: A language and environment for statistical computing. 2020.
- [17] Brooks M E, Kristensen K, Benthem K J ,van, Magnusson A, Berg C W, Nielsen A, et al. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 2017;9(2):378.
- [18] Kanatschnig T, Berger LM, Schrapf N, Tilp M, Kober SE. The Theta Paradox – Delayed Reduction of Frontal Midline Theta Following Downregulation Neurofeedback Training. *bioRxiv*. 2026 Feb 5. Preprint. doi:10.64898/2026.02.03.703090v1.
- [19] Lotte F, Jeunet C, Chavarriaga R, Bougrain L, Thompson DE, Scherer R, et al. Turning negative into positives! Exploiting ‘negative’ results in Brain–Machine Interface (BMI) research. *Brain-Computer Interfaces* 2019;6(4):178–89.
- [20] Ros T, Enriquez-Goppert S, Zorn R, Young KD, Wood G, Whitfield-Gabrieli S, et al. Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). *Brain* 2020;143(6):1674–85.

Pre-Experiment Resting Baseline Improves Motor Imagery BCI Decoding in a Robotic Hand Rehabilitation Context

Aniana Cruz¹, Gabriel Pires^{1, 2}

¹ Institute for Systems and Robotics, University of Coimbra, Coimbra, Portugal

² Engineering Department of the Polytechnic Institute of Tomar, Tomar, Portugal.

E-mail: anianabrito@isr.uc.pt

ABSTRACT: Motor imagery (MI)-based brain-computer interfaces (BCIs) rely on event-related desynchronization and synchronization (ERD/ERS), which are defined relative to a resting-state baseline. Therefore, baseline selection directly affects both the interpretation of the ERD/ERS patterns and MI classification performance. In this work, we systematically evaluate the impact of baseline definition and temporal window selection on ERD/ERS patterns and classification performance in a post-stroke rehabilitation protocol involving a robotic hand exoskeleton, considering both calibration and on-line sessions. Three resting-state baselines were compared: a long rest recorded before the experiment, a long rest recorded after task execution, and a short inter-trial rest commonly used in MI-BCI paradigms. The results show that the pre-experiment resting baseline yielded the strongest and most stable ERD patterns, as well as the highest classification performance in both calibration and online data. Rest recorded after task execution resulted in attenuated ERD and slight performance degradation. The short inter-trial rest led to substantial variability and performance loss.

INTRODUCTION

Motor imagery (MI), the mental simulation of movement, without its physical execution, has been increasingly used as a control signal in brain-computer interfaces (BCIs) across applications ranging from clinical uses, such as post-stroke rehabilitation [1], to non-clinical applications such as gaming [2]. Because MI activates the same neural networks involved in movement execution [3], MI-based BCIs exploit modulations of sensorimotor rhythms (SMRs) over motor cortical areas. During MI, a decrease in the power relative to the reference baseline is typically observed, referred to as event-related desynchronization (ERD), followed by an increase in power known as event-related synchronization (ERS) [4]. These modulations predominantly occur in the mu (8-13 Hz) and beta (14-30 Hz) frequency bands. In MI-BCI paradigms, the baseline generally corresponds to a relaxed resting-state condition, during which SMR power is maximized, thereby enhancing the discrimination between rest and MI states. Since ERD and ERS are defined as relative changes in band power with respect to the baseline, their estimation

is inherently dependent on the choice of baseline.

Several studies have shown that baseline selection plays a critical role in ERD/ERS computation and interpretation [5–7]. In [5], the authors investigated the impact of baseline selection on maximizing ERD by comparing static and dynamic baselines. Their results showed that subject-specific baseline selection could enhance mu-band power. Study [6] explores how two baseline conditions (observation of a static image and a period of stillness) affect mu-ERD/ERS modulation in toddlers during action-observation and action-execution tasks. The results showed that baseline effectiveness varied across subjects, although static-image observation generally produced stronger mu-ERD. In [7], the effects of four different baseline time-window selections were investigated with respect to both ERD/ERS modulation and BCI performance. The results indicated that ERD may be underestimated when the baseline is close to the upcoming trial, as it may include anticipatory neural activity. Contrarily, ERD can be overestimated when the baseline is selected near the end of the previous trial. Previous studies have analyzed baselines obtained in different conditions (e.g., different tasks, some artifacts), which may induce distinct mental states. In most MI-BCI paradigms, the baseline is defined as a short rest period between consecutive MI trials [7]. Because trials typically follow each other closely, such baseline windows may be contaminated by cue-related activity and may include subject concentration and anticipation, thereby distorting the estimated ERD. Incorrect ERD computation can negatively affect MI decoding performance, particularly in stroke patients, where typical contralateral lateralization of sensorimotor activation is frequently disrupted [8].

In this study, we consider a post-stroke rehabilitation scenario in which a BCI is used to control a robotic hand exoskeleton. To optimize this protocol, we systematically compare the impact of three different baselines on the performance of MI classification, namely: 1) a long clean resting state recorded before the beginning of the experiment; 2) a long resting period recorded after task execution; and 3) a short rest period recorded between MI trials.

MATERIALS AND METHODS

Participants: Ten healthy participants (6 female, mean age 26.6 years (std 4.9)) participated in the study. None of them had previous experience with MI-based BCI. Each participant signed an informed consent explaining the purpose of the study, the procedures, the possible risks, and the authorization for the publication of the results.

EEG data acquisition: The EEG signal was recorded using a g.USBamp bioamplifier (g.tec medical engineering GmbH, Austria) from electrodes FC3, FCz, FC4, C3, Cz, C4, CP3, CP1, CPz, CP2, CP4, and Pz according to the international extended 10-20 standard system. The reference electrode was placed at the right earlobe, and the ground was located at FPz. The EEG signals were pre-processed using a notch filter at 50 Hz and a band-pass filter between 8–30 Hz and sampled at 256 Hz.

Experimental Protocol: The experiment consisted of three sessions (Fig. 1):

- **Calibration Session 1:** conducted to collect EEG data for model calibration. This session included a long resting-state recording before task execution, referred to as long baseline 1 (LB1), representing a clean and true resting-state condition. In addition, short baseline periods (SB) were recorded between MI tasks during the trials, representing task-dependent resting states that may include workload and fatigue effects.
- **Calibration Session 2:** used to increase the amount of training data and improve classifier robustness. At the beginning of this session, a second long resting-state was acquired, referred to as long baseline 2 (LB2), representing a resting-state potentially affected by cognitive fatigue after task execution. Similarly to Calibration Session 1, baseline SB was also recorded.
- **Online Session:** performed to evaluate the trained classifier in real time during MI-based robotic rehabilitation, including both visual feedback and somatosensory feedback delivered through actuation of the robotic hand exoskeleton.

The rehabilitation protocol included a robotic hand performing hand opening and hand closing movements. Accordingly, during the calibration sessions, participants were instructed to perform MI of the same movements with their right hand so that the imagined actions were aligned with those executed by the robotic hand. The calibration sessions started with a long resting-state recording lasting 90 s, during which participants were instructed to relax and remain at rest. This was followed by repeated trials (20 times), including relax, rest, and motor imagery periods. Each trial began with a 2-s relax period, during which participants were allowed to rest and be free. This was followed by a 6 s rest period, during which participants were asked to rest and minimize artifacts such as movement and blinking. Subsequently, a 6 s motor

imagery period was presented, during which participants performed motor imagery of the predefined movement (hand opening or closing, according to the visual cue). Participants were instructed to start performing the corresponding MI task as soon as the cue appeared and to continue until the cue disappeared. The total duration of each trial was approximately 16 s. Following the calibration sessions, an online session was conducted, consisting of relax and motor imagery periods. During the MI periods, classification was performed in real time using a model trained on calibration data from rest SB and MI periods. The classifier was applied using a sliding-window approach during the motor imagery period. Windows of 1.5 s length with a 1 s overlap were extracted, denoted as W^i , where $i=1,2,\dots,9$, with onset times shifted relative to the motor imagery onset (0.5 s, 1.0 s, 1.5 s, etc., as denoted in Fig. 1). At the end of each window, the system provided visual feedback: a green bar increased if MI was correctly detected. Simultaneously, somatosensory feedback was delivered through robotic hand actuation (trigger-based control). If at least one classification window was correctly identified, the robotic hand, called Hand Of Hope [9], executed the intended movement (opening or closing hand) (Fig. 2).

Feature Extraction and Classification: The 12 EEG channels were segmented into 1.5-second epochs beginning 0.5 seconds after the onset of each cue. All epochs were first band-pass filtered between 8 and 30 Hz using a Butterworth filter. Then, common spatial patterns (CSP) [10] with two spatial filters ($m = 2$) were applied for feature extraction. The variance (var_t) of each spatial filter projection was computed, followed by normalization and logarithmic transformation ($\log_{10} \left(\frac{\text{var}_t}{\sum_{t=1}^m \text{var}_t} \right)$). Finally, the resulting feature vectors were classified using a Linear Discriminant Analysis (LDA) classifier. Although two different MI tasks were included to increase participant engagement and reduce task monotony, the classifier performed binary classification between MI and rest, since previous studies have shown that this approach is more suitable for dealing with the weak contralateral activation commonly observed in stroke patients [11]. Consequently, the hand opening and hand closing conditions were combined into a single MI class. The training dataset consisted of 40 MI epochs and 40 rest epochs.

Performance Metrics - online data: In online operation, the classifier was evaluated only during MI tasks. Therefore, performance was assessed using metrics based on the true positive rate (TPR). Three metrics were considered: i) the sliding-window TPR (Win_{TPR}^i), which evaluates classification performance for each window W^i ; ii) the mean sliding-window TPR (Win_{TPR}), which is the average performance across all windows; and iii) the trial-level detection rate (Trial_{DR}), which evaluates whether at least one window was correctly classified within a trial. These metrics are defined as follows:

- (i) Win_{TPR}^i , the percentage of correctly classified W^i across all trials:

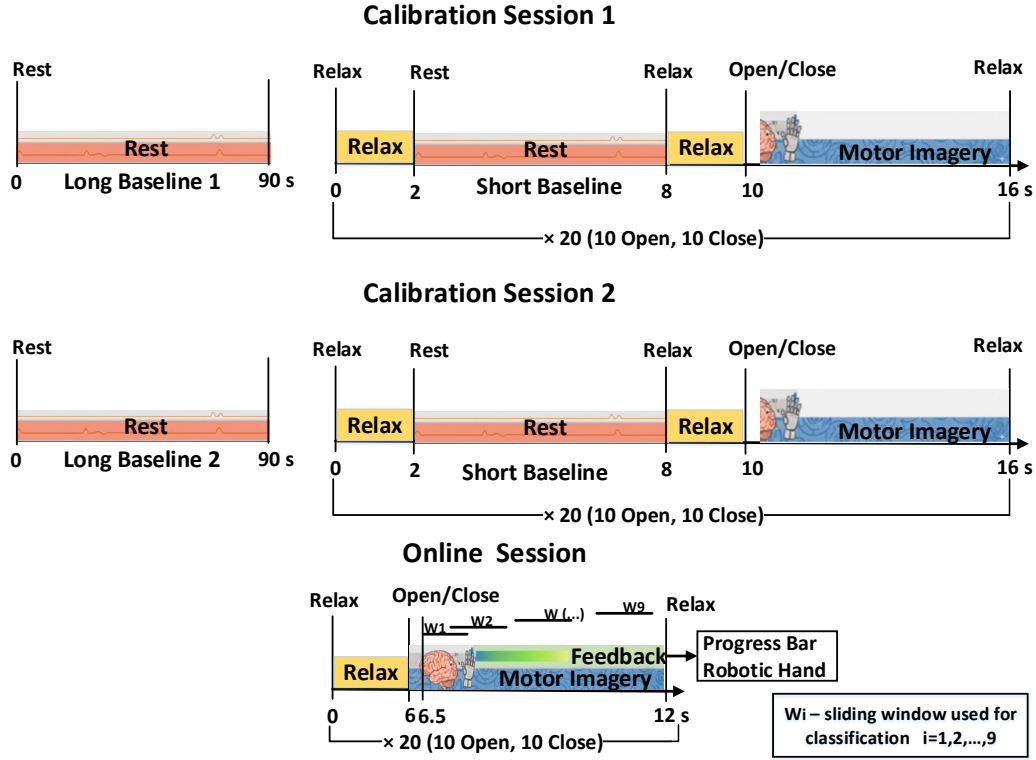


Figure 1: Schematic representation of the rehabilitation protocol, including the calibration sessions, and the online session. Two types of feedback were provided: a progressing bar (visual feedback) and robotic hand actuation (somatosensory feedback). Three baseline conditions were defined: long baseline 1 (LB1), recorded at the beginning of the experiment; long baseline 2 (LB2), recorded after Calibration Session 1; and short baseline (SB), recorded between MI trials.



Figure 2: Experimental setup including the EEG acquisition system, the visual interface, and the robotic hand (Hand Of Hope).

- (iii) $Trial_{DR}$, percentage of trials in which at least one window was correctly classified:

$$Trial_{DR} = \frac{1}{T} \sum_{t=1}^T \gamma_t \times 100 \quad (3)$$

where

$$\gamma_t = \begin{cases} 1, & \text{if at least one } W^i \text{ in trial } t \text{ is correct} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Performance Metrics - calibration data: The performance was computed using the sliding-window accuracy (Win_{Acc}^i), defined as:

$$Win_{Acc}^i = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where positives correspond to MI windows and negatives correspond to rest windows. Thus, whereas online performance is assessed only during MI periods, here classifier performance is assessed using both MI and rest windows.

Neurophysiological analysis: The effect of baseline selection on ERD and ERS was evaluated offline, taking data recorded in the 3 sessions. Event-related spectral perturbations (ERSP) were computed using EEGLAB in the 8–30 Hz frequency band from the CSP projections. To reduce inter-session variability, CSP projections were

$$Win_{TPR}^i = \frac{1}{T} \sum_{j=1}^T \delta_j^i \times 100, \quad \delta_j^i = \begin{cases} 1, & \text{if } \hat{y}_j^i = y_j^i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where T corresponds to the total number of trials per window i (W^i), $\hat{y}_j^i$ and y_j^i denote the predicted and true labels of the W^i , respectively.

- (ii) Win_{TPR} computed as:

$$Win_{TPR} = \frac{1}{N} \sum_{i=1}^N Win_{Acc}^i \quad (2)$$

where $N = 9$ is the total number of windows.

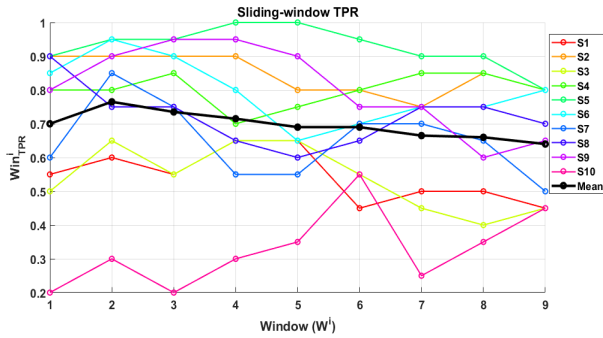


Figure 3: The sliding-window TPR (eq. 1) achieved during the online operation.

normalized by the standard deviation computed across all projected samples within each session. ERSP analysis focused on the minimum ERSP value across the four CSP projections. These values were computed separately for each baseline/MI window combination.

The spectral characteristics of each baseline condition were examined using the power spectral density (PSD), estimated with the Welch method. The relative band power was computed for the theta (4-8 Hz), alpha (8-13 Hz), and beta (13-30 Hz) bands to quantify the contribution of each frequency band to the overall EEG activity, as follows:

$$RP_{band} = \frac{\sum_{f=f_1}^{f_2} PSD(f)}{\sum_{f=f_{min}}^{f_{max}} PSD(f)} \quad (6)$$

where $PSD(f)$ represents the power spectral density at frequency f , f_1 and f_2 denote the lower and upper frequency limits of the band of interest, and $f_{min} = 4$ and $f_{max} = 30$ define the total frequency range. The two motor cortex channels, C3 and C4, were analyzed since they are commonly used in MI-based BCI systems.

RESULTS

Online performance: Table 1 presents the mean TPR obtained during online operation. The results show that in most of the trials, participants achieved at least one correctly classified window, which successfully triggered the robotic hand. The $Trial_{DR}$ ranged from 85% to 100% across participants. For Win_{TPR} , performance was on average approximately 70%. The TPR obtained in each window is presented in Fig. 3 for each subject. As observed, participant S10 had difficulty performing MI, while participants S1 and S3 were able to sustain MI only for short periods. Some participants showed poorer performance in the later windows, possibly because the robotic hand executed the movement after the first correctly classified window, which may have reduced their motivation to continue the task.

Offline classification performance and ERD/ERS analysis: Offline analyses were carried out using the three baseline conditions: LB1, LB2, and SB. The classification problem was defined as MI versus rest, with the rest class corresponding to each baseline condition. Classi-

fication was performed separately for each baseline type and for each MI window onset combination. For the calibration data, classification performance was evaluated using a 5-fold cross-validation procedure. For data recorded during the online session, the classifier was trained on the calibration dataset and evaluated on the online data, thereby simulating real-time operation.

Figure 4 presents the classification performance and ERSP results obtained from the calibration data, averaged across all participants. Both classification accuracy and minimum ERSP values varied substantially with baseline selection and MI window onset. When LB1 was used as the baseline, classification accuracy was consistently high and stable across all MI windows, with Win_{Acc}^i ranging from 81 to 83.3%. This indicates robust separability between MI and resting-state conditions for all windows. Consistently, ERSP analysis revealed strong desynchronization during MI with low variability across window shifts. Using LB2 resulted in lower classification accuracy and increased variability to MI window onset, with accuracies ranging from 77.0% to 83.4%, suggesting reduced discriminability between MI and rest. ERSP values were also attenuated and exhibited greater temporal variability, particularly for later MI window onsets. A marked degradation in performance was observed when using SB as the baseline. In this condition, classification accuracy became highly dependent on the relative timing of baseline and MI windows. The highest accuracies (approximately 70.0–74.6%) were obtained for early MI window onsets (0.5–2.0 s). Similarly, ERSP analysis revealed stronger desynchronization for early MI windows and weaker ERSP values for later window onsets.

Results obtained from the online data showed similar trends but with increased variability (Fig. 5). LB1 again yielded the most stable performance across MI window onsets, with the highest and most consistent performance and ERSP values. For LB2, both classification performance and ERSP values were reduced in the online condition. When using SB, Win_{TPR}^i exhibited substantial variability, ranging from near-chance levels (52%) to high TPR of approximately 82%. The particularly low TPR observed for later MI window onsets may indicate that participants did not sustain motor imagery until the end of the cue period.

Spectral Analysis: The spectral analysis of the three baseline conditions shows consistent relative power distributions across the theta, alpha, and beta bands for the channels C3 and C4 (Fig. 6). The theta band presents relatively low power across all baseline conditions, with only minor variations. LB1 and LB2 have a slightly higher alpha relative power compared to SB, which may indicate a more relaxed resting state. In contrast, the SB condition shows relatively increased beta activity, possibly reflecting residual task engagement or motor anticipation between MI trials. Although consistent trends were observed across baseline conditions, Friedman tests revealed no significant effects for alpha power at C3 ($p = 0.097$) and C4 ($p = 0.895$), nor for beta power at

Table 1: The trial-level detection rate (eq. 3) and mean sliding-window TPR (eq. 2) obtained during the online operation.

Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Mean	STD
$Trial_{DR}(\%)$	95	100	100	90	100	100	100	100	100	85	97.0	5.4
$Win_{TPR}(\%)$	54.4	83.9	53.9	80.0	92.8	79.4	65.0	72.2	80.6	32.8	69.5	18.0

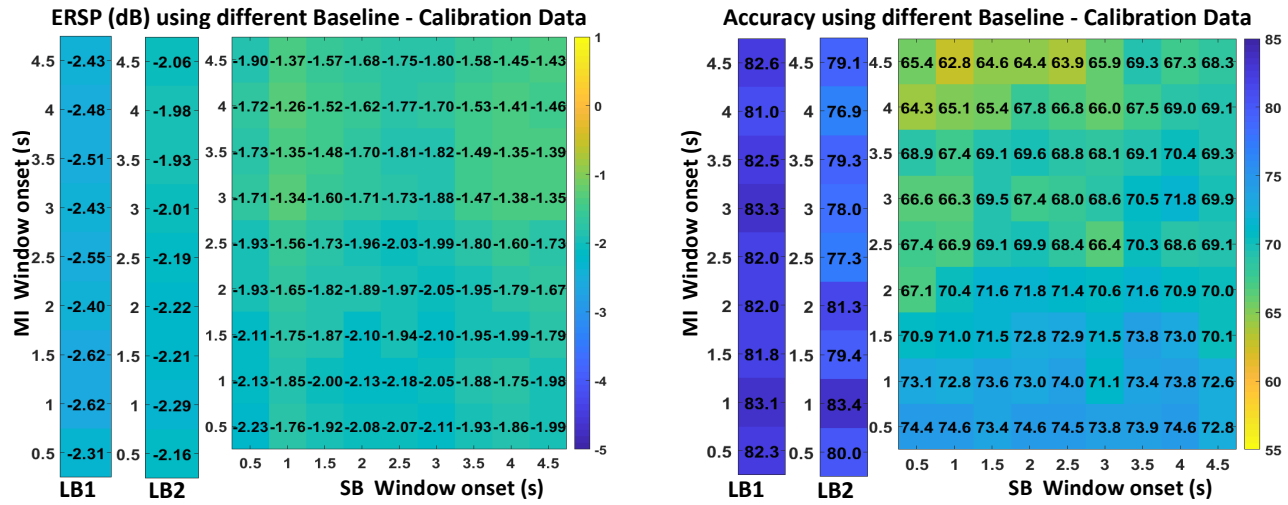


Figure 4: Confusion matrices for ERSP and classification performance from calibration data: Left - ERSP for different baseline conditions (LB1, LB2, SB (per window)) across MI window onsets. Right - Classification performance for the same baseline conditions.

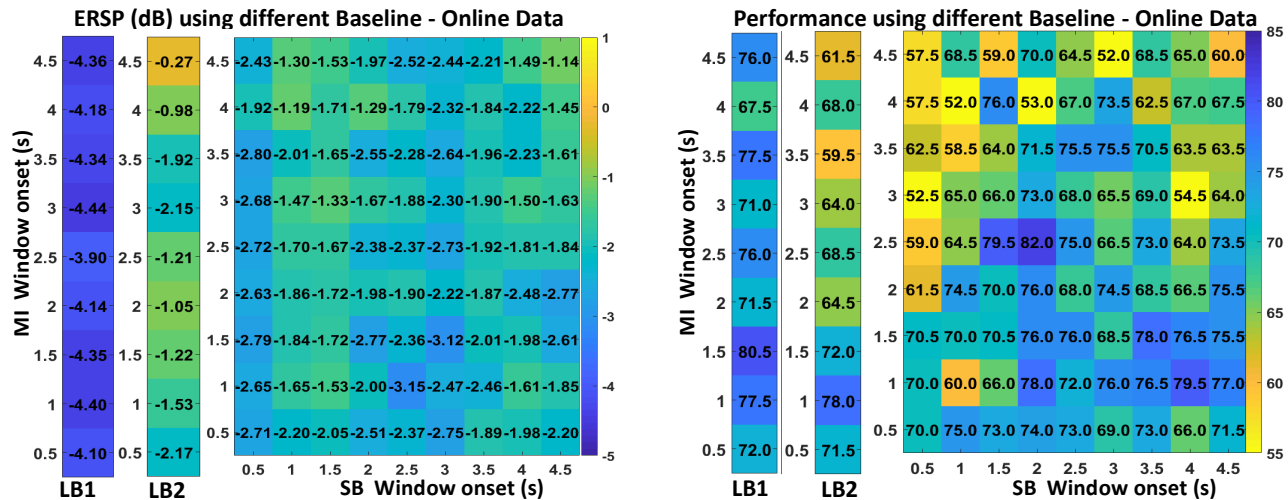


Figure 5: ERSP and classification performance for different baseline conditions (LB1, LB2, SB (per window)) across MI window onsets during the online phase.

C3 ($p = 0.121$) and C4 ($p = 0.895$).

DISCUSSION

Baseline definition and temporal window selection substantially influence both ERSP and classification performance, with important implications for online BCI operation. Using LB1 consistently yielded the strongest and most stable desynchronization, which translated into the highest and most stable performance in both calibration and online data. In contrast, LB2 exhibited attenuated ERSP and moderate performance degradation, suggesting that rest recorded after task execution

is partially influenced by residual cognitive engagement, thereby reducing discriminability from MI. The most pronounced effects were observed when using SB, where classification accuracy strongly depended on the selected time window onset. Although high accuracies could be achieved by carefully selecting MI and baseline timing window, performance was unstable. A repeated-measures ANOVA revealed significant differences among baseline conditions in the calibration data ($F = 10.6$, $p = 0.0009$). Post-hoc pairwise comparisons indicated that LB1 achieved significantly higher accuracy than SB ($p = 0.0005$), and no significant differences were ob-

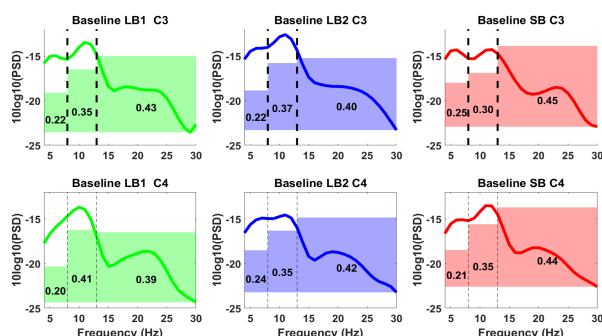


Figure 6: Grand average of power spectral density for channels C3 and C4 during the three resting conditions. The shaded areas indicate the theta, alpha, and beta frequency bands. The values inside each band correspond to the relative power for that frequency range.

served between LB1 and LB2 ($p = 0.6$). The comparison between LB2 and SB approached significance ($p = 0.05$). In contrast, no significant differences between baseline conditions were found for the online data ($F = 0.7$, $p = 0.51$). Across all baseline conditions, performance was higher for calibration data than for online data. However, these metrics are not directly comparable, since calibration performance was evaluated using accuracy, and online performance was assessed using TPR. Nevertheless, the lower online performance is consistent with the performance degradation reported in [12]. This study showed that when feedback does not match the participant's expectations, it can lead to frustration and disrupt focus. Furthermore, in our setup, the robotic hand actuation triggered after the first correctly classified window may have also reduced participants' engagement in motor imagery, considering they had already achieved their goal.

CONCLUSION

This work systematically investigated the impact of baseline definition and temporal window selection on both classification performance and ERD/ERS, under calibration and online conditions for post-stroke rehabilitation. The results consistently demonstrated that a long resting period recorded before task execution provides the best overall performance among the three baseline conditions. It constituted the most stable baseline, producing strong ERD together with the highest and most robust performance across both calibration and online data. In contrast, long rest recorded after task execution led to reduced ERD magnitude and degraded classification performance, likely due to residual cognitive or task-related engagement. The greatest performance degradation was observed when using a short rest period between MI trials as the baseline, possibly due to contamination by residual sensorimotor activity. Although this baseline is commonly used in BCI systems, these findings highlight the importance of baseline selection in achieving reliable and robust motor imagery decoding, especially in online BCI operation.

ACKNOWLEDGEMENTS

This work was supported by the Portuguese Foundation for Science and Technology under Grants UpReGain (<https://doi.org/10.54499/2023.13809.PEX>), BCI4ALL (COMPETE2030-FEDER00842800, doi: <https://doi.org/10.54499/2023.17977.ICDT>), and <https://doi.org/10.54499/UID/PRR2/00048/2025>.

REFERENCES

- [1] Baniqued PDE et al. Brain-computer interface robotics for hand rehabilitation after stroke: A systematic review. *Journal of neuroengineering and rehabilitation*. 2021;18(1):15.
- [2] Ntetska A, Mimou A, Tzimourta KD, Angelidis P, Tsipouras MG. A deep learning approach to classifying user performance in bci gaming. *Electronics*. 2025;14(24):4974.
- [3] Sharma N, Baron JC. Does motor imagery share neural networks with executed movement: A multivariate fmri analysis. *Front. Hum. Neurosci*. 2013;7:564.
- [4] Pfurtscheller G, Da Silva FL. Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clinical neurophysiology*. 1999;110(11):1842–1857.
- [5] Tangwiriyasakul C, Verhagen R, Van Putten MJ, Rutten WL. Importance of baseline in event-related desynchronization during a combination task of motor imagery and motor observation. *Journal of Neural Engineering*. 2013;10(2):026009.
- [6] Piazza C, Visintin E, Reni G, Montirosso R. The effect of baseline on toddler event-related mu-rhythm modulation. *Brain Sciences*. 2021;11(9):1159.
- [7] Rimbart S, Trocellier D, Lotte F. Impact of the baseline temporal selection on the erd/ers analysis for motor imagery-based bci. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2023, 1–4.
- [8] Jones TA. Motor compensation and its effects on neural reorganization after stroke. *Nature Reviews Neuroscience*. 2017;18(5):267–280.
- [9] Ho N et al. An emg-driven exoskeleton hand robotic training device on chronic stroke subjects: Task training system for stroke rehabilitation. In: 2011 IEEE international conference on rehabilitation robotics. 2011, 1–5.
- [10] Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*. 2000;8(4):441–446.
- [11] Cruz A, Kuzmanoski M, Pires G. Optimizing bci rehabilitation protocols for stroke: Exploring task design and training duration. In: 2025 IEEE 8th Portuguese Meeting on Bioengineering (ENBENG). 2025, 57–60.
- [12] Pires G et al. A new error-monitoring brain-computer interface based on reinforcement learning for people with autism spectrum disorders. *Journal of neural engineering*. 2022;19(6):066032.

Decoding Meaning: Cluster-Based Analysis of Semantic Processing Patterns in EEG

Joshua James Podmore¹, Moein Radman¹, Riccardo Poli¹, Silke Paulmann², Ian Daly¹

¹ Brain-Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, England, UK

² Department of Psychology, University of Essex, England, UK

E-mail: jp24194@essex.ac.uk

ABSTRACT: Conventional brain-computer interfaces (BCIs) typically rely on slow character-by-character selection mechanisms. Semantic BCIs (sBCIs) aim to overcome this limitation by decoding entire concepts directly from neural activity. Here, we introduce GRASP (Game of Retrieval and Semantic Processing), a novel auditory paradigm designed to isolate volitional semantic processing in EEG recordings. Participants ($n = 9$) identified concrete concepts during a gamified word-guessing task and provided response-locked markers (spacebar press). Importantly, 25% of trials consisted of phonologically matched pseudo-word sentences, serving as an active baseline condition. Time-frequency representations were computed using Morlet wavelets and statistically evaluated using threshold-free cluster enhancement permutation testing to characterise the spatio-spectro-temporal signatures of semantic processing within individual participants. As an initial proof-of-concept analysis, we focus on the body *vs.* dummy contrast within the GRASP paradigm. Our findings provide insight into the timing, scalp regions, and frequency bands associated with EEG-based semantic processing and may support the development of targeted preprocessing and machine-learning approaches for sBCIs.

INTRODUCTION

Progressive motor neurone diseases such as Amyotrophic Lateral Sclerosis (ALS) can ultimately produce a Locked-In State (LIS), in which cognitive function is preserved but volitional movement is lost [1]. Individuals living with LIS, as well as those affected by severe congenital motor impairments such as cerebral palsy, often require augmentative and alternative communication (AAC) systems to maintain any channel for expressing needs and intent [2]. Most existing BCI-based AAC solutions, such as P300 spellers and SSVEP [3] systems, share the fundamental constraint of iterative character-by-character selection. These place substantial demands on residual sensory and motor functions that may themselves be compromised in target user populations [4]. A single sentence can require dozens of sequential selections distributed across minutes of interaction, imposing severe limits on information transfer rate [5].

Semantic brain-computer interfaces (sBCI) offer a qualitatively different approach where a single decoding operation communicates a complete concept rather than

a single character, offering potential gains in communication efficiency and, in principle, independence from the intact sensory pathways on which conventional BCIs depend [6].

The process of recovering semantic content from distributed neural activity has been highlighted in several key papers [6]. Notably, Mitchell and colleagues [7] successfully predicted fMRI activation patterns for 60 concrete nouns from distributional semantic features, achieving cross-validated decoding accuracies of 0.68-0.85 and generalising to stimuli words unseen during classifier training. This is made possible by the distributed and category-specific organisation of semantic representations across the cortex, a principle formalised by Patterson and team [8].

Advances towards harnessing the mechanics of these distributions with more practical brain imaging were illustrated by Murphy et al [9], who demonstrated above-chance performance for binary classification (animals *vs.* tools) with Electroencephalography (EEG). While single-trial decoding yielded a mean accuracy of 72%, performance increased to 98% when predictions were aggregated across multiple stimulus repetitions. This is primarily attributable to the boost in data quality provided by anchoring EEG epochs to a spacebar press at the moment of semantic retrieval. This dramatically reduced the inclusion of uninformative samples and removed the need for computationally expensive sliding-window approaches. Despite this, the epoching was stimulus-locked to data post-image cue presentation. This undoubtedly introduced significant perceptual confounds and diminishes any claim of true semantic processing [10].

This form of direct cueing causes early visual cortical activations, attentional orienting, and lexical access processes to co-occur with the intended semantic signal [10]. Attempts to circumvent this via a delayed ideation paradigm in concert with fNIRS have shown promising results [11]. However, attempts to utilise these same stimuli and associated classification methods for EEG time-series data have been less fruitful [10]. This is likely due to weak task-engagement, low target-concept relevance, and the absence of trial-level behavioural validation.

Kivisaari and team [12] addressed prior paradigm limitations by having participants derive a target from three increasingly specific orthographic clue sentences, never presenting it directly. Their fMRI findings demonstrated that semantic representations emerge progressively as evidence accumulates across clues, paralleling natural language comprehension. Ghazaryan and collaborators (2024) [13] adapted this paradigm for MEG using 60 concrete Finnish nouns and zero-shot ridge regression to predict word2vec [14] semantic vectors from brain activity. Cross-subject averaged data reached significance after the third clue, however, participant-level decoding failed, attributed to MEG's low signal-to-noise ratio for written stimuli and declining engagement across the 12-18 repetitions required for signal averaging.

The present study introduces GRASP (Game of Retrieval and Semantic Processing). We aim to address the shortcomings of previous experimental paradigms by combining the best elements of each aforementioned approach. The GRASP paradigm is a word-guessing task. It features response-locked temporal indexing (spacebar press), functionally motivated and meta-analytically verified semantic categories (food/body parts/musical instruments) [15]. Additionally, GRASP employs an auditory stimulus presentation format to control for eye-movement artefacts and reading speed variation. Task engagement is boosted by ensuring all clues are unique, and the spacebar interrupt feature allows participants to conduct the task at their own pace. Further, GRASP introduces a novel within-session pseudo-word control condition in which one quarter of trials present acoustically matched sentences in which all content words are replaced by phonologically valid nonsense words. This provides an empirical acoustic baseline against which semantic responses can be contrasted.

This is key, as prior work in the EEG-based semantic decoding space has focused primarily on concept *vs.* concept binary decoding [6]. This assumes that the top-down labels assigned to any given set of samples are discrete. It should be plainly understood that the distance in semantic space between, e.g. foot and banana, is dramatically larger than, e.g. tongue and banana. Despite this, in a traditional semantic decoding experiment, the concepts of foot and tongue would typically be binned into a generic “body part” group. The same could be noted of cross-category comparisons between ear and piano, or fingers and violin. In other words, treating all these items as belonging to pure, orthogonally separable categories fails to accommodate the presence of a connected semantic topology [16]. Notably, these obstacles, which are inherent to concept *vs.* concept decoding, could underpin the low semantic-decoding performance reported in the corresponding literature.

Importantly, this cannot be overcome via the simplistic use of baseline or resting period data in a so-called concept *vs.* resting state setup. Since the resting brain is

intrinsically and continuously active in semantically structured ways [17, 18], any conventional rest period constitutes an uncontrolled, heterogeneous comparator, rendering these undirected experimental periods unsuitable for the characterisation of semantic decoding-related signals.

The utility of our “active-resting state” comes from the explicit positioning of each subject towards a specific low-semantic state across the semantic processing continuum. We hypothesise that this homogenises the active-rest class and allows us to investigate the semantic signals unique to specific concepts in a more discrete and structured manner with greater sensitivity and selectivity. Concretely, we believe that it is vital to first determine the typical spatio-spectro-temporal properties of semantic processing in EEG. We argue that this can only be done cleanly using a controlled, active-rest period. Any attempt to perform this task via concept *vs.* concept decoding is necessarily problematic given the continuous, non-discrete geometry of the cortical semantic topology.

The aims of this study include: (i) establishing whether the pseudo-word control condition can isolate semantic from acoustic processing contributions; and (ii) to characterise the spatial, spectral, and temporal profiles of EEG activity associated with the body *vs.* dummy contrast at the single-subject level as an initial proof-of-concept analysis within the GRASP paradigm. Additionally, we also intend to interpret these results to inform the development of targeted preprocessing stages and machine learning approaches that more precisely exploit the frequency bands, scalp regions, and response-locked time windows most discriminative of semantic processing.

MATERIALS AND METHODS

Participants: 9 neurologically healthy adults participated (avg: 23 years, 6 female). Recruitment was restricted to British English first-language speakers. Ethical approval was granted by the University of Essex Ethics Committee. Participants received £30 compensation irrespective of performance.

Stimulus Generation: Three semantic domains were employed: food (9 fruits/vegetables), musical instruments (9 items), and body parts (9 limbs/sense organs), yielding 27 concept items in total (see Figure 1). For each item, a three-part sentence-level definition was composed, with each sub-sentence increasing in specificity towards the ground truth term. Each definition was produced in two forms: an original sentence (real words) and a pseudo-sentence in which all content words were replaced by phonologically matched pseudo-words. Note, all function words such as stopwords, modals, conjunctions, and pronouns were left unchanged.

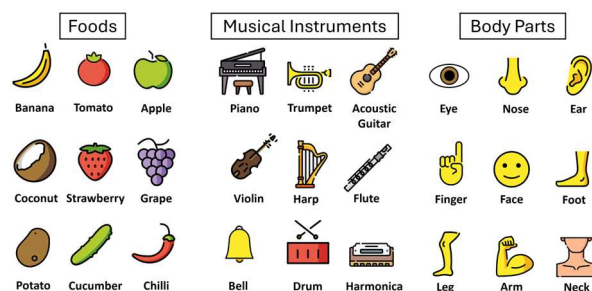


Figure 1: A pictographic of all 27 items for guessing in the GRASP clue paradigm. Each of the three categories comprises 9 items. Note, each block is constructed from 36 items total, with the remaining 9 filled by the so-called “dummy” trials to serve as a low-semantic comparator set.

Pseudo-word variants were generated in two stages. Firstly, automatic precaching used a constrained variant of the Wuggy pseudo-word generator. Each candidate was validated for syllable-count match, character-length, confirmed non-word (dictionary lookup), and text-to-speech (TTS) duration within 200ms of the original word.

All audio was synthesised using Microsoft Edge TTS to produce mono 22,050 Hz 16-bit WAV files. Word-level time boundaries were obtained for every utterance via the Montreal Forced Aligner (MFA) using a UK English pronunciation dictionary and acoustic model. The experimental structure comprised 12 blocks of 36 trials each, 27 original and 9 pseudo-trials per block. A fixed-subset randomisation strategy was applied, ensuring all participants received the same 144 pseudo-sentences with only trial order randomised.

All clue sentences were unique, and comprise 3 subcomponents increasing in specificity to the ground-truth item at each step, e.g. “This is known for its universal appeal, it is often served with butter, and is frequently baked with a cheese topping.”, answer: potato. Here is the same sentence transformed into a pseudo-sentence: “This is known for its unbizzal asteal, it is often sained with beller, and is frequently roked with a shoose flipping.”

Experimental Paradigm: The experiment was implemented in PsychoPy with audio playback via PyGame. Each session required approximately 3 hours (inc. setup & breaks). At the start of each block, a pictographic was presented of all 27 target items for 5 s (see Figure 1). Each trial commenced by displaying a central fixation cross, followed by the playback of the sentence WAV audio (see Figure 2). Participants were instructed to maintain fixation but could press the spacebar at any point to interrupt the audio and signal early identification. Following the audio, or its interruption, a text-entry box appeared instructing participants to type their best guess (30 s timeout). For pseudo-word trials, participants were also required to perform the spacebar interrupt upon identification and enter the term ‘dummy’ to verify attentive listening. If the audio played to completion, a 1.5 s dynamic Gaussian mask was displayed, followed by a 0.5 s pause.

Each trial concluded with a jittered inter-trial interval (ITI) of 0.5-1.25 seconds. Post-block responses were autocorrected using an item-prioritised strategy and a lemmatisation stage. The average trial-level semantic accuracy was then quantified as the mean cosine similarity between word2vec embeddings of the corrected responses and the ground-truth items. For dummy trials, the latency from the first pseudo-word onset to the spacebar response was also computed.

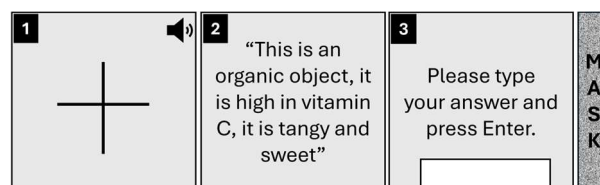


Figure 2: A representation of the stimulus presentation order. The “speaker” icon on the 1st tile is only displayed here to visually indicate the onset of the audio clue. Further, the 2nd tile is here only for demonstration purposes. Only the clue sentence audio is presented to participants. At no point are the clues provided orthographically. The 3rd tile is a faithful representation of the input screen users either see after the clue audio has completed, or immediately after the onset of a spacebar press. The final narrow panel illustrates the onset of the visual mask stage displayed at the end of each trial.

EEG Recording and Preprocessing: EEG data were recorded using a 64-channel BioSemi ActiveTwo system at 2,048 Hz with a standard 10-20 montage and four external EOG channels (EXG1-4), with all subsequent preprocessing performed in MNE-Python [19]. Continuous data were preprocessed via a 1-80 Hz zero-phase bandpass finite-impulse response filter, downsampling to 512 Hz, and inter-block cropping to retain only task-relevant data. Following a 50 Hz notch filter, bad channels were detected via PyPrep [20] (RANSAC) and corrected using spherical spline interpolation.

Data were then re-referenced to an approximate infinity reference. To protect the fitting of the Extended Infomax ICA (32 components, 10,000 iterations), transient and muscle artefacts were annotated and temporarily excluded. Automated artefact rejection was then performed via z-score correlation ($z = 2.0$) with EOG and ECG channels. An additional behavioural filter was also applied to ensure the exclusive retention of exact matches (user input vs. ground truth item label) and trials within the typical range of the subject-specific typing speed.

Time-Frequency Analysis: EEG was analysed in a response-locked framework, with time zero defined as the spacebar press, examining spectral power differences between conditions over the 1.5 s pre-response window (−1.5 to 0 s). A response-locked alignment was preferred over stimulus-locking because the timing of conceptual identification varies across trials and participants. All analyses were conducted at the single-subject level, in recognition of the substantial inter-individual variability in EEG spectral profiles and scalp topographies.

The analyses were implemented in MNE-Python and SciPy [21]. Epochs were resampled to 256 Hz and restricted to the 64 EEG channels under the standard 10-20 montage. For each contrast, single-trial time-frequency representations (TFRs) were computed independently for each condition using a Morlet wavelet transform over a logarithmically spaced frequency range of 3-35 Hz (32 frequencies). Condition differences were tested simultaneously across the full scalp and time-frequency plane using a non-parametric permutation test with threshold-free cluster enhancement (TFCE; [22]). At each channel-frequency-time voxel, an independent-samples Student's t-statistic was computed.

A three-dimensional adjacency structure was defined jointly over channels (electrode neighbourhood graphs derived from the standard 10-20 montage), frequency bins (adjacent bins along the frequency axis), and time points (adjacent samples along the time axis). This ensured that cluster enhancement captures the natural smoothness of EEG spectral data in all three dimensions simultaneously rather than treating each dimension independently.

The permutation null distribution was constructed from 1,000 random permutations of condition labels. Statistical inference was performed using threshold-free cluster enhancement (TFCE) with a permutation-based cluster test. A TFCE statistic map was computed across the channel $\times$ frequency $\times$ time space. Significant voxels were defined as those belonging to clusters whose TFCE-based cluster-level p-values were below $\alpha = 0.05$, thereby controlling the family-wise error rate across the full spatio-spectro-temporal space.

The main aim of this analysis was to define the spatial, spectral, and temporal structure of semantic processing at the single-subject level rather than to perform binary classification. For each participant, the analysis yielded a three-dimensional TFCE statistic together with a voxel-level significance mask derived from TFCE-significant clusters, capturing the neural differentiation between body-part semantic trials and dummy trials.

RESULTS

In the present report, we focus on the body *vs.* dummy contrast, which provides a controlled comparison between semantically meaningful trials and pseudo-word trials lacking interpretable semantic content. This, therefore, offers a representative case for examining the 3-D spatio-spectro-temporal structure of semantic processing within the GRASP paradigm. Figure 3 illustrates the corresponding TFCE statistic map across all 3 dimensions for one subject (07) within a response-locked analysis window (-1.5 s, 0 s).

The topographic TFCE summary (Fig. 3A) reveals a broad left-lateralised spatial pattern, with the strongest TFCE statistics concentrated over fronto-central sites, extending into left temporal and centro-parietal electrode locations. The temporal TFCE profile (Fig. 3B) shows a

gradual increase in statistical strength beginning approximately 1.4 s prior to the spacebar press. TFCE values remain elevated across much of the pre-response interval and reach a maximum at around -0.6 s. Following this peak, TFCE values decrease but remain above early-window levels. The spectral TFCE summary (Fig. 3C) shows that elevated TFCE values are primarily concentrated in the low-frequency range, with a dominant peak around 7-9 Hz, with a smaller secondary elevation present in the 13-16 Hz range.

Taken together, these results indicate that semantic pseudo-word differentiation in Subject 7 is characterised by a left-lateralised fronto-central network expressed predominantly in the theta and alpha range and emerging during the late stages of the pre-response interval.

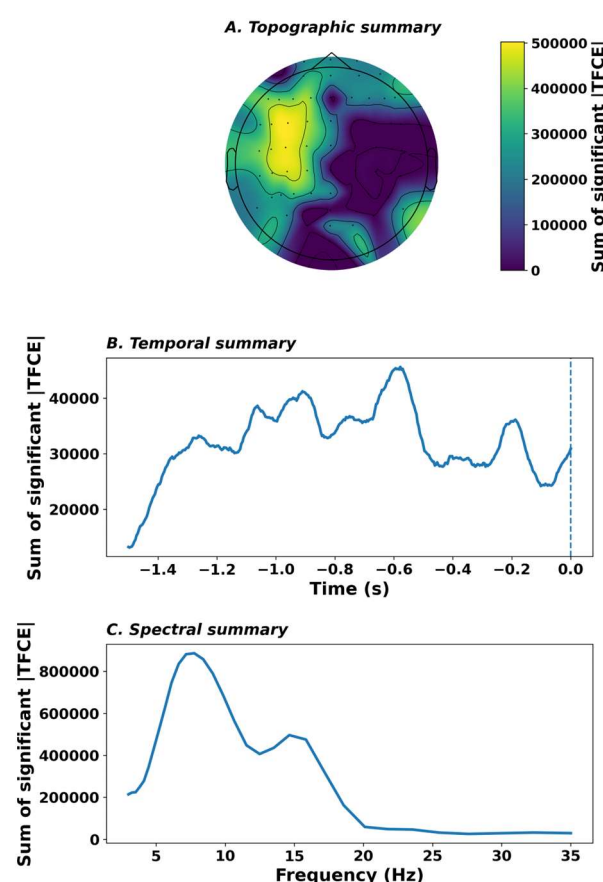


Figure 3: Spatio-spectro-temporal TFCE summary for Subject 07 in the body *vs.* dummy contrast. TFCE values are shown without normalisation to preserve single-subject interpretability.

Further, to illustrate the directionality of this effect at its dominant scalp locus, Figure 4 shows the mean log₁₀ Morlet-derived power difference between body and dummy trials at channel FC3, with all TFCE-significant clusters overlaid as black contours. The plot reveals a power increase in the theta-alpha range, extending all the way up to 20 Hz in some instances. Note also that the primary significant cluster spans most of the pre-response window. Additionally, isolated higher-frequency clusters are visible at approximately 20-25 Hz and 30-35 Hz, however, these should be understood as potentially subject-specific.

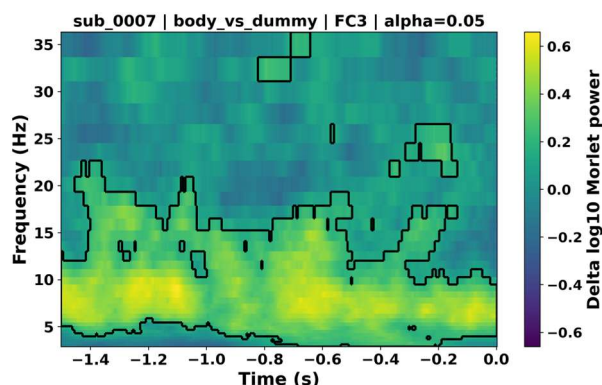


Figure 4: Mean log10 Morlet-derived power difference (body vs. dummy) at channel FC3 for Subject 07. FC3 was the highest-ranked channel in the single-subject TFCE summary (Fig. 3A). Black contours indicate TFCE-corrected significant clusters at $\alpha = 0.05$.

Further, to assess the consistency of the observed effects across participants, TFCE results were summarised descriptively across *all nine subjects*. Each participant exhibited at least one TFCE-significant cluster in the body vs. dummy contrast (see Figure 5).

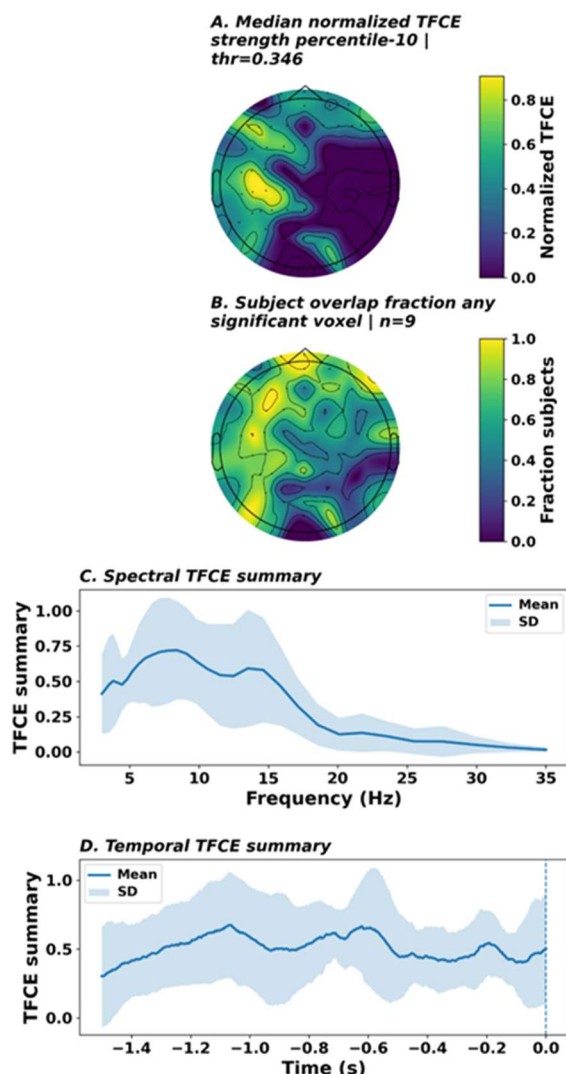


Figure 5: Cross-participant consistency of TFCE effects across spatial, spectral, and temporal domains (normalised within each participant prior to aggregation to account for inter-individual variation in absolute TFCE magnitude).

The spatial TFCE summary (Fig. 5A) shows the median normalised TFCE strength across participants projected onto the scalp topography, illustrating the cross-participant spatial distribution of the effect. The resulting map reveals a broadly left-lateralised pattern, with elevated TFCE values over fronto-central, central, and temporal electrodes of the left hemisphere. In contrast, right-hemisphere electrodes show comparatively lower TFCE values, indicating that the observed TFCE pattern is predominantly left-lateralised across participants.

The subject overlap map (Fig. 5B) shows the proportion of participants exhibiting TFCE-significant voxels at each scalp location. The highest overlap is observed over left fronto-central, central, and temporal electrodes, indicating that these regions represent the most consistent spatial locus of the effect across participants. The spectral TFCE summary (Fig. 5C) illustrates the distribution of TFCE values across frequencies summarised across participants. Elevated TFCE values are primarily concentrated within the theta to alpha range, with a smaller secondary contribution extending from the alpha band into the low-beta range. The temporal TFCE summary (Fig. 5D) illustrates the cross-participant temporal profile of TFCE values. TFCE values increase during the pre-response interval and exhibit two local maxima. One around -1.1 s and another around -0.6 s prior to the behavioural response. Following these peaks, TFCE values gradually decline.

DISCUSSION

This article provides an initial characterisation of the spatio-spectro-temporal signatures of semantic processing in EEG within the GRASP paradigm. Across participants, TFCE-based differentiation between semantic and pseudo-word trials was expressed as a predominantly left-lateralised spatial pattern spanning fronto-central, central, temporal, and centro-parietal scalp regions (see Figs. 3A, 5A, 5B). This spatial distribution broadly aligns with established models of semantic processing, which emphasise the role of left temporal cortices in semantic representation and frontal regions in semantic retrieval [16].

Spectrally, the strongest TFCE values were concentrated within the theta to alpha range, with smaller contributions extending into the low-beta band (see Figs. 3C & 5C). Oscillatory activity in this frequency range has previously been linked to semantic integration and lexical retrieval during language comprehension, suggesting that the signals identified here likely reflect neural processes involved in conceptual identification.

Temporally, the TFCE profiles indicated that these differentiating signals emerged well before the overt behavioural response, with local maxima occurring approximately 1 s and 0.6 s prior to the spacebar press (see Figs. 3B & 5D). This is consistent with cognitive processes related to semantic retrieval rather than the motor execution of the response.

As seen in Figure 4, the primary TFCE-significant cluster, situated in the theta-alpha range, spans most of the pre-response window. The fact that it does not appear to resolve within this window is of interest, however, the epoching onset could not be extended further back to observe potential propagation events without introducing overlap with preceding trials and compromising trial quality. We believe that this represents a potentially fruitful avenue of investigation worth pursuing in future work.

Taken together, these findings suggest that the GRASP paradigm can reliably elicit EEG signatures of semantic processing. As an initial proof-of-concept, the present results highlight candidate scalp regions, frequency bands, and response-locked temporal windows that may be exploited in future machine-learning approaches for semantic brain-computer interfaces.

CONCLUSION

Beyond characterising these effects, the present findings demonstrate the value of response-locked analysis for isolating semantic processing from motor execution. Here, we highlight the spatial, spectral, and temporal features that may be informative for future decoding or machine learning applications. As a proof of concept, this work lays the foundation for developing sBCIs and motivates further studies with larger samples, richer stimulus sets, and cross-participant analyses.

ACKNOWLEDGEMENTS

This work was supported by funding from the Leverhulme Trust [Grant Number: RPG-2023-207].

REFERENCES

- 1] Brown RH, Al-Chalabi A. Amyotrophic lateral sclerosis. *N. Engl. J. Med.* 2017;377:162–172
- 2] Linse K, Aust E, Joos M, Hermann A. Communication matters: pitfalls and promise of high-tech communication devices in palliative care of severely physically disabled patients with amyotrophic lateral sclerosis. *Front. Neurol.* 2018;9:603
- 3] Abiri R, Borhani S, Sellers EW, Jiang Y, Zhao X. A comprehensive review of EEG-based brain-computer interface paradigms. *J. Neural Eng.* 2019;16:011001
- 4] Wolpaw JR, Wolpaw EW. *Brain-Computer Interfaces: Principles and Practice.* Oxford University Press, Oxford, UK (2012)
- 5] Waller A. Telling tales: unlocking the potential of AAC technologies. *Int. J. Lang. Commun. Disord.* 2019;54:159–169
- 6] Rybář M, Daly I. Neural decoding of semantic concepts: a systematic literature review. *J. Neural Eng.* 2022;19:021002
- 7] Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, et al. Predicting human brain activity associated with the meanings of nouns. *Science* 2008;320:1191–1195
- 8] Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? *Nat. Rev. Neurosci.* 2007;8:976–987
- 9] Murphy B, Talukdar P, Mitchell T. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain Lang.* 2011;117:12–22
- 10] Rybář M, Poli R, Daly I. Using data from cue presentations results in grossly overestimating semantic BCI performance. *Sci. Rep.* 2024;14:28003
- 11] Rybář M, Poli R, Daly I. Decoding of semantic categories of imagined concepts of animals and tools in fNIRS. *J. Neural Eng.* 2021;18:046035
- 12] Kivisaari SL, van Vliet M, Hultén A, Lindh-Knuuttila T, Faisal A, Salmelin R. Reconstructing meaning from bits of information. *Nat. Commun.* 2019;10:927
- 13] Ghazaryan G, van Vliet M, Saranpää A, Lammi L, Lindh-Knuuttila T, Hultén A, et al. Trials and tribulations when attempting to decode semantic representations from MEG responses to written text. *Lang. Cogn. Neurosci.* 2024;39:1149–1160
- 14] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* 2013;26:3111–3119
- 15] Radman M, Podmore JJ, Poli R, Paulmann S, Daly I. Decoding semantic categories: insights from an fMRI ALE meta analysis. *J. Neural Eng.* 2025;22:061006
- 16] Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 2016;532:453–458
- 17] Raichle ME. The restless brain. *Brain Connect.* 2011;1:3–12
- 18] Jackson RL, Hoffman P, Pobric G, Lambon Ralph MA. The semantic network at work and rest: differential connectivity of anterior temporal lobe subregions. *J. Neurosci.* 2016;36:1490–1501
- 19] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 2013;7:267
- 20] Bigdely-Shamlo N, Mullen T, Kothe C, Su KM, Robbins KA. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front. Neuroinform.* 2015;9:16
- 21] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 2020;17:261–272
- 22] Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 2009;44:83–98

Investigating Human and Classifier-Level Uncertainties: An EEG Study Using Grasp Imitation Task

Osman Cagri Oguz¹, Hamidreza Kasaei¹, Steffen Frey¹, Andreea Ioana Sburlea¹

¹ Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

E-mail: o.c.oguz@rug.nl

ABSTRACT:

In BCI research, uncertainty quantification recently attracted researchers to enhance human-machine collaboration and transparency of the systems. Grasping constitutes one of the most crucial interactions with the environment, for which the associated risks of overconfident but inaccurate classification are high. However, there is yet to be an explicit paradigm design that allows investigating if classifier uncertainties align with that of humans. By developing a grasp imitation paradigm with visually degraded videos during the observation stage, we aimed to incur uncertainties on the grasping decisions and execution. EEG and EMG data from 8 participants is collected. The results demonstrated that among the grasp types, 5-finger precision grasp is the best classified one, with 47% within-class accuracy according to an LDA classifier using movement related cortical potentials as features. Interestingly, for the participant having highest ROC-AUC score, subjectively reported and the LDA classifier uncertainty aligns. As the data collection is still ongoing, future directions are also discussed.

INTRODUCTION

We perceive and act within noisy sensory environments, where the affordances and consequences are not always immediately clear. Imagine grasping a cup with a beverage with not yet known temperature. Humans are more likely to combine multiple information sources to estimate the risks and reduce uncertainties. Even if the interaction is supposed to happen in a very short manner, accommodating the safest grasp alternative would be wise to prevent harmful consequences, such as spill and burn. Humans can be remarkably good at accommodating different motor decisions as a result of constantly taking sensory and perceptual uncertainties into account [1]. One might also imagine this scenario where the motor intentions from brain signals are decoded by a classifier, which is then translated into executable motor commands for the robotic hand. If the classifier is insensitive to how humans represent uncertainties or overconfident regardless of its classification accuracy, generated response may likely cause injuries. In this research, we aim to see if uncertainty for brain decoding indeed aligns with how humans represent their uncertainties in a motor task, more

specifically grasp-imitation. The novel aspect of the research is twofold. First, we contribute to the field by introducing a paradigm inducing uncertainty in motor actions. Second, as a result of this adjustment, studying for the first time if uncertainty represented by BCI classifiers actually aligns to that of humans themselves.

In the field of brain-computer interfaces (BCIs), grasping attracts growing interest from the researchers due to its importance for daily functioning and interactions with the environment. To provide a comprehensive model of how grasps are represented in the brain, previous research showed that different stages of motor actions carry information about different characteristics. For instance, properties of the object to-be-grasped can be observed during the hand pre-shaping stage, while muscle activity is mostly evident in holding stages [2]. Using low-frequency time-domain signals (e.g., movement-related cortical potentials (MRCPs)) the researchers demonstrated a decoding performance above the chance level among three different grasps, especially during the planning and object release stages [3]. Unimanual and bimanual grasps have also been shown to be decoded if MRCPs are used as features [4].

However, merely focusing on the classification accuracy for BCI decoders may not provide the most comprehensive picture. In a similar line of reasoning, there is also a significant caveat for evaluating the conclusions reached by brain decoding studies. That is, demonstrating accurate classification based on the input signal (e.g., EEG features) may not naturally follow conclusions that those features indeed carry the relevant information for the task. This risk is often associated with greedy and non-transparent nature of algorithms used for classification tasks. To overcome this phenomenon of decoder's dictum, it is proposed to connect behavior in a psychologically plausible manner to have better grounded inferences for underlying neural representations [5]. In this respect, second-order representation of the informational content, such as uncertainty given the classification output, is a candidate to produce more grounded inferences. However, the previous research does not incorporate uncertainty regarding motor or perceptual cues for instructing movement execution to participants. Therefore, missing the opportunity to connect classifier and human level uncertainties.

It is also known that BCI classifiers are often overconfident even for inaccurate classifications, resulting in uncertainty quantification (UQ) techniques to attract researchers in the field [6]. More specific to EEG-based BCIs, UQ is arguably even more important given the poor signal-to-noise ratio compared to invasive recordings, especially for multiclass decoding scenarios, such as for different grasps. Additionally, for BCIs on decoding motor signals, overconfidence may lead to unwanted consequences in real-life when the potential risks of choosing the wrong action is high. Promisingly, recent research show improvements through uncertainty-based early classification and rejection [7], [8].

Previous research on UQ for EEG-based BCIs revealed that even though deep-learning based methods, such as Deep Ensembles, produced best classification accuracies, more traditional methods may represent uncertainty better through post-hoc calibration to match predicted class probabilities to accuracy [7]. In addition, researchers proposed frameworks based on Bayesian networks and dropout-based methods to estimate uncertainty for decoding EEG signals. They showed promising results for the classification tasks on public datasets with comparable performance to computationally costly ensemble-based methods [9].

To the best of our knowledge, if uncertainty represented by BCI classifiers actually aligns to that of humans themselves is yet to be studied. With this goal in mind, we use a paradigm inspired by previous perceptual decision-making research, in which the sensory stimulus strength is manipulated to adjust difficulty [10]. With this manipulation, it is expected to incur uncertainty regarding the perceptual and hence the motor decisions determining subsequent grasp execution. In this paper, we use a traditional classification algorithm to decode brain activity during grasp execution, namely the linear discriminant analysis (LDA), due to previous reports on proper uncertainty representation.

MATERIALS AND METHODS

Paradigm: The current research investigates multiple grasps by utilizing a grasp imitation task. The participants watched grayscale videos depicting three different grasps on four different objects to decide which type of grasp they will perform subsequently (see Fig. 1). The grasps include pincer (2-finger grip), precision (5-finger) and power grasp (with the palm also touching the object). The grasping videos follow after fixation cross on the screen and a variable-timed random noise overlaid on the first frame. Pre-stimulus noise is used in line with the previous studies on perceptual decision-making, to eliminate activity bursts in decision-making areas through introducing sudden visual stimuli [10]. The difficulty of the videos is manipulated by adjusting the opacity level, in other words, via the blending coefficient with a random noise overlay (see Fig. 2). The stimulus presentation is done via Psychtoolbox in Matlab [11]. The optimal

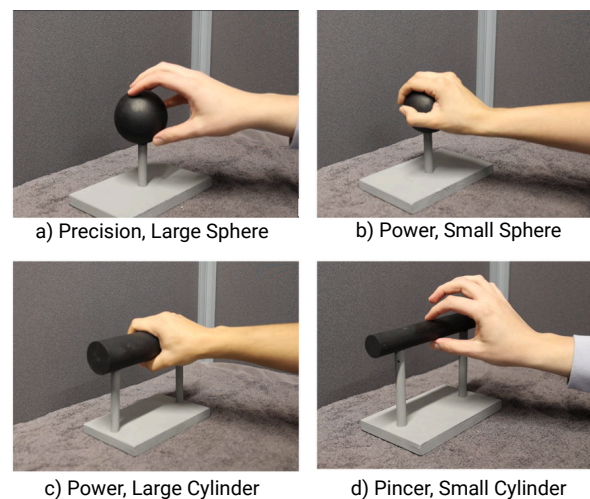


Figure 1: Three different grasps and four different objects used during the experiment, with grasp and object types are indicated below each figure.

opacity level is determined during the practice session by using staircase procedure utilizing Quest+, to ensure balanced distribution of confidence ratings and to train the participants in terms of the timing of the grasps [12].

By presenting ambiguous video clips for motor action selection and planning, uncertainty is induced at the level of perceptual decision-making before any movement is initiated. By inducing uncertainty on the sensory stimuli itself that participants base their decisions, we expect to induce uncertainty on subsequent grasping action. The second stage is characterized by translating the motor intent established through perceptual decision into a grasping alternative among 3 options. Displaying the motor execution cue right after the displays prevents additional (extensive) re-evaluation of the stimuli, thus the internal uncertainties are reflected in grasp execution. The BCI decoder takes the EEG signal from this stage as an input with the goal of returning ground-truth grasp category (i.e. actually realized grasp alternative), and uncertainty for its output is calculated. The third stage is characterized by participants' metacognitive evaluation regarding stage two, which is expected to reveal trial-by-trial fluctuations in the internal (un)certainties for grasp execution. This evaluation is realized through explicit instruction for reporting the certainty while executing the grasp rather than during video displays and with a scale ranging from 1-to-9. Participants are explicitly instructed prior to the experimental session to report the feeling of certainty while executing the grasps, rather than judgements on the accuracy of their decisions. Since the BCI classifiers are trained on EEG data during grasping itself, reporting the uncertainty in that same stage provides a more comparable measure. However, we still expect indirect coupling between perceptual decision and motor execution uncertainty. On the other hand, single-trial LDA classifier uncertainty for the executed grasp type is quantified as its distance to the linear decision boundary constrained from 0 to 1.

At the end of each trial, participants are also asked about which grasp they performed to have the ground truth labels for decoding. There were 384 trials in the actual experiment, equally balanced over objects, actors, and grasp types. In addition, participants are also given the option to report 'none' for the trials they simply did not perform any grasp in case of attentional lapses.

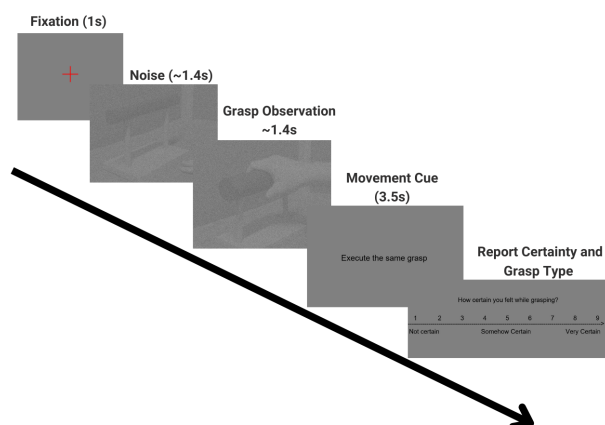


Figure 2: The experimental paradigm and the timeline of a trial. First, following a fixation cross and pre-stimulus noise, participants view the grasping videos where the actors reach for the objects and grasp them accordingly. Following this, they are asked to execute the same grasp type. In the end of each trial, they are asked about which grasp type they executed and also how certain they felt while executing the movements.

Data Collection: The EEG data is collected from 8 participants (age range: 21 - 29) with 500Hz of sampling rate, AFz channel as ground and FCz as online reference via the LiveAmp 64 (Brain Products GmbH, Germany). 50 Ag/Cl electrodes are placed on the scalp to have a good coverage of the fronto-central, central, and centroparietal areas and hence motor and decision-related activity areas. Some electrodes for the anterior and temporal areas are repurposed for EOG and EMG recordings. Specifically, we used 3 EOG channels to capture horizontal and vertical eye movements for facilitating artifact rejection. 2 EMG channels are placed on the right forearm of the participants to measure the activity during grasping and to capture movement onset. The remaining electrodes in temporal and anterior areas are discarded.

EEG Preprocessing and Feature Extraction: EEG data analysis is performed in EEGLAB [13]. Data is first cleaned from the break periods, then band-pass filtered from 0.1 to 40Hz using a Butterworth 4-th order, zero-phase filter. Then, data is downsampled from 500Hz to 100Hz. ICA decomposition followed this step, for which only EEG channels are considered. Following the decomposition, individual components are rejected based on SASICA, with an ADJUST procedure and correlation with horizontal / vertical EOG channels [14], visual inspection and alignment with ICLabel [15]. Remaining components are projected back and additional filtering from 0.1 to 3Hz using Butterworth 4th order zero-phase filter is performed to compute MRCPs. For EMG chan-

nels, the data is band-pass filtered from 20 to 250Hz, notch-filtered at 50Hz, and averaged across the two electrodes. The onset detection is done by epoching the continuous data from -1 to 3.5 seconds with respect to 'execute' prompt onset. From there, Teager-Kaiser Energy Operator (TKEO) is calculated due to previous results suggesting enhanced onset detection [16]. The first time samples after the movement cue that crosses the z-threshold of 10 is marked as the movement onset, while 1 second period before the movement cue is taken as a baseline. The movement onset provided an additional informative feature to epoch the EEG data for MRCPs.

Classification of Grasp Type: To classify different grasp types, regularized LDA with shrinkage is used. The MRCPs starting from 1 seconds prior to and lasting until 1 seconds after the movement onset is taken as an input for an LDA classifier. Data is further reduced to channels around supplementary and primary motor areas, and decimated by a factor of 5 to have manageable amount of features for each trial. Data from 2 subjects are excluded from this analysis due to unreliable detection or failure to detect trial-wise EMG onsets. The trials for which the participants did not execute any grasp or if the movement onset is not reliably calculated are discarded. The decoding accuracy and confusion matrix is obtained for each participant using a stratified 5-fold cross validation, and in the end predictions are aggregated over the participants.

RESULTS

Behavioral Results: To analyze the behavioral data, we first removed the trials in which the participant reported that they could not perform any grasp within the instructed time limits. On average, 12.75 trials (std: 16.65) are removed. This resulted in eliminating a total of 102 trials out of total 3072 trials from 8 participants. We refer to accuracy as executing the same grasp type among the three alternatives with respect to the grasp video. We observed that accuracy was highly variable between the subjects, ranging from 0.38 to 0.71. Next, we examined if the reported certainties at the end of each trial can be predicted by accuracy or executed grasp type. The summary of the certainty scores across accuracy and grasp types can be seen in Table 1. For that purpose, we used a linear mixed model with subject as the grouping variable. Results showed a significant main effect of accuracy ($z = 8.74, p < .001$), for which correct trials are associated with higher certainties ($\beta = 0.61$). This pattern suggests that participants can, at least to some extent, monitor the accuracy of their perceptual decisions hence being more confident in their subsequent movement. Additionally, a significant effect of Grasp Type was also observed ($z = 2.91, p = .004$). From Table 1, it can be seen that this significance may be explained by higher certainty associated with executing the power grasp compared to other two alternatives. This stems potentially by the fact that power grasp requiring higher force and lower precision,

Table 1: Certainty Levels Across Accuracy and Grasp Types

	Mean(SD)	Count
Accuracy		
Correct	5.69 (2.45)	1674
Incorrect	4.78 (2.59)	1296
Grasp Type		
Pincer	5.23 (2.51)	821
Precision	5.14 (2.51)	1141
Power	5.51 (2.63)	1008

leading to higher certainty estimates than pincer and precision grasps.

Neurophysiology: For neurophysiological investigations, determining the precise movement onset is important since onsetting with respect to the motor cue may not provide the most informative time window. Movement execution time may differ significantly between participants and between trials of the same subject following such a noisy visual stimuli. For that purpose, we used EMG electrodes and have found that mean movement onsets are indeed variable among participants, ranging from 333 to 698 milliseconds. Using Spearman correlation, we also investigated if there is a significant relation with movement onset and reported certainties. The results showed a significant negative correlation consistently for each subject (all $p < .05$, with r_s values ranging from -0.20 to -0.49). This pattern indicates that if the participants are uncertain, they may use additional time to potentially decide between the motor alternatives and start grasp execution. We also investigated the activity in centro-parietal areas, which was previously linked with perceptual decision-making and confidence for a given decision. The patterns we observed also supports the previous interpretation regarding using additional time (see Fig. 3a). Specifically, the rising activity is observed until movement onset (i.e., detected by the EMG electrodes) rather than cue onset (i.e., when participants are instructed to execute the same grasp type). This finding suggests that decision-making may still continue even after the visual stimuli is no longer present on the screen. [17].

Another observed pattern from EEG data suggested an early motor planning with negative MRCP peak in the contralateral side of the brain with respect to the grasping hand. That is, following the presentation of grasp videos but before the movement cue, negative going potential is peaked in C3 channel (see Fig. 3b). This might be the result of participant's anticipation regarding starting the movement as soon as the videos end, and might be further contributed by action observation during video displays. Therefore, investigation of EEG data points towards the interpretation that motor planning starts before or concurrently as the evidence accumulation for perceptual decision determining the details of the subsequent action.

Classification Results: To classify between different grasps, shrinkage regularized LDA is utilized with MRCPs from -1 to 1 seconds of trial-wise movement onsets as features. The results over 6 subjects is summa-

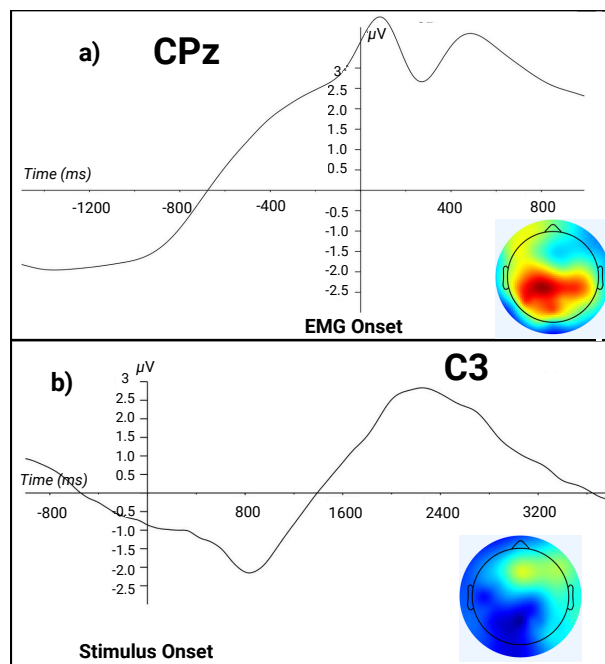


Figure 3: CPz and C3 waveforms when averaged over the subjects. a) CPz channel activity in grand-averaged data over 6 subjects, along with the scalp map representing the mean activity for -200 to $+200$ ms from movement onset. The activity rises until the movement onset. b) C3 channel activity in grand-averaged data over 8 subjects is shown with the scalp map representing the mean activity around 600 to 800 ms from stimulus (grasp video) onset. The peak negativity is observed around 800 ms following the stimulus onset.

rized in Fig. 4. The ROC-AUC scores across participants range from 0.5 to 0.62 , suggesting poor decodability using these features. This might be due to combination of the limitations regarding experiment design that will be detailed in discussion, EEG signal-to-noise ratio, or class imbalance as can be seen from Tab. 1). Indeed, predictions tend to favor precision grasp compared to other grasp types, which is the most reported category. The relationship between classification and human uncertainties are investigated within-subject manner. The class probability for the ground-truth label can be interpreted as its distance to the decision boundary constrained in the range from 0 to 1 . When analyzed this way, Spearman correlation test was only significant for the subject having the highest ROC-AUC score of 0.62 ($\rho = 0.14$, $p < 0.05$). Although it is promising to see this pattern for the subject with the highest classification metric, the lack of statistical significance for the other subjects necessitates collecting data from more participants before reaching any conclusions regarding the relationship between classifier and human-level uncertainties.

DISCUSSION

In this study, we developed a paradigm inspired by previous perceptual decision-making literature to manipulate difficulty for the grasp videos. This design allowed us to investigate uncertainties from human reports and BCI

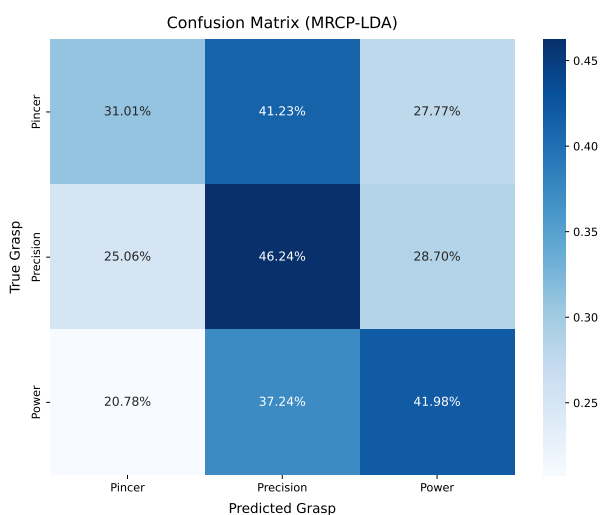


Figure 4: Confusion matrix when the predictions and true labels are aggregated over 6 subjects. Among the grasp types, precision grasp is the best classified category followed by power grasp. The LDA predictions are biased towards precision grasp, likely due to class imbalance.

classifiers for the first time in the field. While the data collection is still ongoing, the decoding performances are variable across participants. Poor decoding performances can be partly explained by low signal-to-noise ratio of EEG, especially for decoding multiclass grasp scenarios where there is no immediately distinct underlying activity. However, we observed a significant relationship between LDA classifier uncertainty and participant reports for the subject with highest ROC-AUC metric. This pattern indicates that as the classification performance improves, classifiers and subjective reports might be utilizing the similar neural representations. However, it is not yet possible to reach any final conclusions with the current state of data collection and analysis as we require more subjects and improved decoding performance for the already collected data from other participants.

Interestingly, while previous research demonstrated higher performance for decoding pincer grasps [3], we did not see this pattern in our data. In addition to class imbalance, methodological differences might play a role in this difference, since the participants are not given any objects while performing the grasps. This choice is deliberately made to use videos with same grasps but different objects to increase variability and prevent participants to decide based on the low-level cues. However, it might lead to decreased decodable information from the brain. Indeed, for that study, decoding accuracy for pincer grasp is at its highest when participants were releasing the objects [3].

Limitations: The choice of not using the objects while participants are executing the grasps may lead to more unnatural feeling accompanying the movements. It might therefore influence reporting the certainty of the movement. Additionally, we explicitly instructed participants for reporting the uncertainties during grasping rather than during video observation. However, this might be overly

complex for some people that finds it difficult to disentangle the two and reporting metacognitive evaluations. The feeling of certainty might also be affected significantly by the grasp type in general that normally requires more force, such as power grasp. Therefore, this needs to be kept in mind for the future analysis.

Future Direction: The prominent models for perceptual decision-making are based on the idea of evidence accumulation until some decision bound is reached. This formulation views the decision process as gathering sufficient sensory evidence from the environment to come up with a choice [18]. In humans, a signal showing these characteristics is found in EEG studies. An event-related potential, called centro-parietal positivity (CPP), is demonstrated to rise following the stimulus displays and until the response. Previous studies reported that especially the rising slope of CPP aligns with task difficulty [10]. In addition to decision-making related characteristics, there is a growing body of research connecting CPP to decision confidence. Since CPP is generally thought to index decision-making process itself, using CPP as a candidate signal challenges the traditional views of confidence as emerging post-decision. Researchers provided supporting evidence for this view and suggested that choice confidence might be represented in the very same mechanisms underlying decision-making [19]. In a recent formulation of decisions involving motor reports, researchers proposed a leaky accumulation process between perceptual and motor systems, where the leakage is controlled by confidence [20]. Following these lines of research, it might be plausible to connect this EEG signals from centro-parietal areas for indexing confidence as the decisions evolve to guide subsequent motor actions. Indeed, we have observed that CPz activity rises until the movement. Therefore, using peak amplitudes before the movement onset might provide an additional uncertainty metric to compare with human-level or classifier-level uncertainties.

Previous studies for uncertainty quantification have traditionally categorized it into two main sources, namely epistemic and aleatoric uncertainty [21]. The first one refers to uncertainty that can be improved with more training since it resulted from model incompetence, while the latter is the uncertainty that cannot be reduced and due to inherent stochasticity of the input. Despite the abundant use of this formulation, it should be treated cautiously for how well the two sources can be disentangled [6]. Evidential deep learning (EDL), on the other hand, views uncertainty in vacuity and dissonance terms. The first refers to incomplete evidence for the decision alternative while the latter refers to conflicting evidence among decision alternatives. By learning Dirichlet Density, predictions are modeled as distributions over potential softmax outputs [22]. More specific to our task design, it might be that the uncertainties that participants report resonates more with EDLs approach. That is, participants either have incomplete or conflicting evidence for different grasp alternatives. Therefore, future research

by calibrating the loss functions of shallow neural networks, such as EEGNet [23], to return Dirichlet parameters might be promising.

CONCLUSION

Our study introduces a new paradigm for BCI research on grasp decoding to connect uncertainties returned from the classifiers, from relevant EEG indices for perceptual decisions, and from human metacognitive judgements. Such a holistic approach to investigate the relationship between human behavior and machine learning algorithms is crucial for improving trust and comprehensive understanding for BCI systems. We think that this approach will also be fruitful to develop adaptive BCI systems that are informed by human-aligned uncertainty representation inherent in behavior and decisions.

REFERENCES

- [1] Wolpert DM, Landy MS. Motor control is decision-making. *Current Opinion in Neurobiology*. 2012;22(6):996–1003.
- [2] Sburlea AI, Müller-Putz GR. Exploring representations of human grasping in neural, muscle and kinematic signals. *Scientific Reports*. 2018;8(1):16669.
- [3] Sburlea AI, Wilding M, Müller-Putz GR. Disentangling human grasping type from the object's intrinsic properties using low-frequency eeg signals. *Neuroimage: Reports*. 2021;1(2):100012.
- [4] Schwarz A, Pereira J, Kobler R, Muller-Putz GR. Unimanual and bimanual reach-and-grasp actions can be decoded from human eeg. *IEEE transactions on biomedical engineering*. 2020;67(6):1684–1695.
- [5] Ritchie JB, Kaplan DM, Klein C. Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*. 2019;70(2):581–607.
- [6] Jong IPd, Sburlea AI, Valdenegro-Toro M. Uncertainty quantification in machine learning for biosignal applications – a review. 2025;(arXiv:2312.09454).
- [7] Suurmeijer J, Jong IPd, Valdenegro-Toro M, Sburlea AI. Uncertainty quantification for motor imagery bci – machine learning vs. deep learning. 2025;(arXiv:2507.07511).
- [8] Hassan M. Confidence-based early classification for motor imagery brain-computer interfaces. 2024.
- [9] Duan T, Wang Z, Liu S, Yin Y, Srihari SN. Uner: A framework for uncertainty estimation and reduction in neural decoding of eeg signals. *Neurocomputing*. 2023;538:126210.
- [10] Kelly SP, O'Connell RG. Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*. 2013;33(50):19434–19441.
- [11] Brainard DH. The psychophysics toolbox. *Spatial Vision*. 1997;10(4):433–436.
- [12] Watson AB. Quest+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision*. 2017;17(3):10.
- [13] Delorme A, Makeig S. Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*. 2004;134(1):9–21.
- [14] Chaumon M, Bishop DVM, Busch NA. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of Neuroscience Methods*. 2015;250:47–63.
- [15] Pion-Tonachini L, Kreutz-Delgado K, Makeig S. Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*. 2019;198:181–197.
- [16] Carvalho CR, Fernández JM, Ama AJ del, Oliveira Barroso F, Moreno JC. Review of electromyography onset detection methods for real-time control of robotic exoskeletons. *Journal of NeuroEngineering and Rehabilitation*. 2023;20:141.
- [17] Grogan JP, Rys W, Kelly SP, O'Connell RG. Confidence is predicted by pre- and post-choice decision signal dynamics. *Imaging Neuroscience*. 2023;1:imag-1–00005.
- [18] Ratcliff R, McKoon G. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation*. 2008;20(4):873–922.
- [19] Gherman S, Philiastides MG. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*. 2015;106:134–143.
- [20] Balsdon T, Philiastides MG. Confidence control for efficient behaviour in dynamic environments. *Nature Communications*. 2024;15(1):9089.
- [21] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*. 2021;110(3):457–506.
- [22] Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. 2018.
- [23] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.

A Comparative Study of EEG Representations and Channel Settings for EEG-to-Text Decoding during Reading Tasks

Eduardo V. Pérez-Hernández, Alejandro A. Torres-García, Luis Villaseñor-Pineda

Dept. of Computational Science
Biosignals Processing and Medical Computing Laboratory
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Puebla, México

E-mail: {eduardo.perezhera, alejandro.torres, villasen}@inaoep.mx

ABSTRACT: Decoding natural language from electroencephalography (EEG) signals remains a major challenge in brain-computer interface research for assistive communication. This work presents a comparative study of EEG representations and channel configurations for EEG-to-text decoding during reading tasks. We evaluate Hilbert–Huang Transform (HHT) features, power spectral density (PSD), and temporal statistical descriptors across three spatial configurations: full-head EEG, a reduced 10–20 system, and a Broca–Wernicke subset. Experiments are conducted on the ZuCo v1 and v2 datasets using fully autoregressive decoding without teacher forcing. Results show that decoding performance varies across EEG representations and channel configurations. Temporal statistical features achieve the best overlap-based performance under the full-head configuration (BLEU-1 = 11.45, ROUGE-1 F1 = 11.52), while reduced channel subsets remain competitive despite lower spatial dimensionality. These findings highlight the importance of representation design, spatial selection, and task consistency for EEG-to-text decoding under short and variable EEG segments.

INTRODUCTION

Human communication is a fundamental component of learning, social interaction, and personal autonomy. However, this ability can be severely impaired in individuals with neurological disorders affecting movement and speech, such as Amyotrophic Lateral Sclerosis (ALS). Brain-Computer Interfaces (BCIs) have emerged as a promising technology for restoring communication by directly decoding neural activity. Early BCI communication systems primarily relied on paradigms such as P300-based spellers, which allow users to select characters through visually evoked responses [1]. Although these systems can achieve reliable performance, they typically require structured interfaces and sequential character selection, resulting in slow communication rates. Another widely explored approach is imagined speech recognition, where users mentally rehearse words to induce discriminative neural patterns. However, such systems are

typically limited to closed vocabularies or predefined linguistic categories [2], and their performance degrades as vocabulary size increases or when generalizing to unseen words [3]. Recent advances in large language models have enabled new paradigms for open-vocabulary neural decoding, where brain activity is mapped directly to natural language sequences. In this context, electroencephalography (EEG) represents an attractive modality due to its non-invasive nature, relatively low cost, and high temporal resolution. The EEG2Text framework proposed by [4] introduced the first approach integrating word-level EEG features with a pre-trained language model to reconstruct sentences. This framework relies on the Zurich Cognitive Language Processing Corpus (ZuCo) datasets [5, 6], which combine EEG and eye-tracking recordings collected during natural reading. Although ZuCo provides a unique resource for studying neural language processing, it was not originally designed for text decoding tasks. The dataset aligns EEG signals with individual words using gaze duration information, producing short and highly variable word-level EEG segments. Earlier language decoding studies were primarily conducted using neuroimaging modalities such as fMRI [7]. Several recent studies have extended the EEG2Text framework by improving the transformation of EEG features into language-model embeddings and introducing semantic alignment strategies [8–13]. A common characteristic of many of these approaches is the use of the Hilbert–Huang Transform (HHT) as the primary EEG representation, motivated by its suitability for analyzing non-stationary signals. However, these methods typically do not consider how adaptive spectral decomposition behaves when applied to short-duration EEG segments. In word-level EEG extracted from natural reading, gaze-based segmentation produces highly variable temporal windows, often containing a limited number of samples. Under such conditions, decomposition-based representations may become unstable due to insufficient temporal support. Despite this limitation, the impact of EEG feature representation and spatial channel configuration on EEG-to-text decoding under short and variable-length conditions has not been systematically analyzed.

The main contributions of this work are:

- A controlled comparison of EEG representations for word-level EEG-to-text decoding, analyzing Hilbert–Huang Transform (HHT) features used in prior work, Power Spectral Density (PSD), and temporal statistical descriptors under short and variable EEG segments.
- A systematic evaluation of spatial electrode configurations, comparing full-head EEG (105 channels), a reduced 10–20 montage (20 channels), and a task-relevant subset covering Broca and Wernicke areas (10 channels).
- An analysis of how reading task composition influences decoding performance under realistic short-segment conditions.

Rather than proposing a new architecture, this study focuses on understanding representation–space interactions to inform the design of robust EEG-to-text systems.

MATERIALS AND METHODS

Task Definition: As illustrated in Fig. 1, EEG and eye-tracking signals are simultaneously recorded during reading tasks. Eye-tracking information is used to temporally align EEG recordings with individual words, enabling the extraction of word-level EEG segments. These raw EEG segments are subsequently preprocessed and transformed into structured EEG representations, which serve as inputs to a sequence-to-sequence language model for sentence generation.

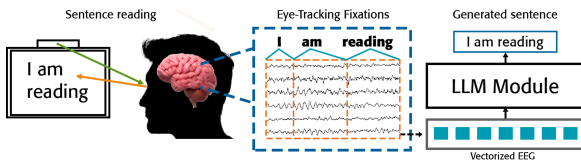


Figure 1: Overview of the EEG-to-text decoding task

The EEG-to-text decoding task aims to generate a sentence S from a sequence of word-level EEG feature vectors $E = [e_1, e_2, \dots, e_K]$, where each e_k corresponds to the EEG activity associated with the k -th word. The decoding process follows a sequence-to-sequence formulation that models the conditional probability of the target sentence given the EEG sequence:

$$p(S|E) = \prod_{k=1}^K p(s_k|E, s_{<k}) \quad (1)$$

where K denotes the length of the target text sequence and $s_{<k}$ represents previously generated tokens.

Database: This study uses the Zurich Cognitive Language Processing Corpus (ZuCo 1.0) [5] and ZuCo 2.0 [6], which provide synchronized EEG and eye-tracking recordings collected during multiple reading tasks: Sentiment Reading (SR), Normal Reading (NR),

and Task-Specific Reading (TSR). In the SR and NR tasks, participants were instructed to read sentences naturally for comprehension, which was verified through randomly presented questions. In the TSR task, participants were required to determine whether a sentence corresponded to a predefined semantic category.

Table 1 summarizes the task-level statistics of the benchmark datasets. Here, N denotes the number of subjects, and %Fix represents the percentage of words containing at least 50 EEG samples. Across all tasks, approximately 62% of words are represented by fewer than 50 EEG samples. The average word-level EEG segment length is 141 samples, with a standard deviation of 88 samples, indicating substantial variability in segment duration.

Table 1: Task-level summary statistics. Values are reported as mean $\pm$ standard deviation across subjects.

Task	N	Sentences	% Fix	Av. Samples
SR v1	12	400	67.3 $\pm$ 8.6	145.0
NR v1	12	300	66.3 $\pm$ 11.8	138.7
TSR v1	12	407	50.3 $\pm$ 13.1	131.7
NR v2	18	349	66.6 $\pm$ 10.2	149.3

EEG signals were recorded using a 128-channel Geodesic Hydrocel system at a sampling frequency of 500 Hz. Twenty-three channels were removed due to artifact contamination, primarily from neck, chin, and electrooculogram electrodes. The artifact-corrected EEG signals provided by Hollenstein et al. [5] were used in all experiments.

Word-level EEG segments extracted from gaze duration exhibit highly variable duration, with many segments containing fewer than 50 samples (100 ms at 500 Hz). Such short windows can limit the stability of spectral estimates, particularly for low-frequency components.

EEG Channel Configurations: To analyze the effect of spatial channel selection, three EEG channel configurations were evaluated, as illustrated in Fig. 2: (i) full-head EEG with 105 channels, (ii) a reduced 10–20 system comprising 20 channels, and (iii) a task-relevant subset of 10 channels covering Broca and Wernicke areas.

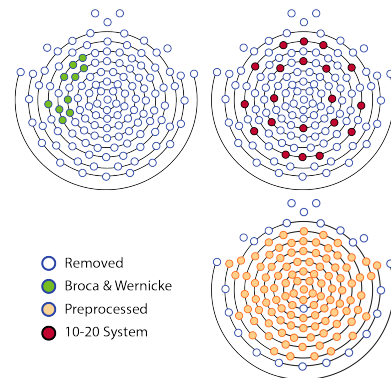


Figure 2: EEG channel configurations evaluated in this study: (i) full-head (105 channels), (ii) reduced 10–20 system (20 channels), and (iii) Broca–Wernicke subset (10 channels).

These configurations allow controlled analysis of spatial information reduction.

EEG Representations: Word-level EEG segments are defined using gaze duration information obtained from eye-tracking data. For each word, consecutive fixations are grouped, and the longest fixation sequence is selected. The EEG samples corresponding to this fixation group are concatenated to form a continuous EEG segment associated with the word. This procedure ensures temporal alignment between neural activity and linguistic units.

Temporal Statistical Features: Eight time-domain features are computed per channel: mean, standard deviation, skewness, kurtosis, Hjorth mobility, Hjorth complexity, sample entropy, and root mean square (RMS) [14]. This results in 840 features for the full-head configuration (105 channels), 160 for the reduced 10–20 configuration (20 channels), and 80 for the Broca–Wernicke configuration (10 channels).

Due to the short duration of word-level EEG segments, low-frequency bands that require longer temporal windows may produce unstable estimates. Therefore, spectral features are computed only within frequency bands that remain robust under short-window conditions.

Power Spectral Density (PSD) Features: Frequency-domain features are extracted using Welch’s method applied independently to each word-level EEG segment and channel. A 50-sample window (100 ms at a sampling rate of 500 Hz) with 50% overlap and a 256-point FFT is used. The mean spectral power is computed within five frequency bands: alpha2 (10.5–13 Hz), beta1 (13.5–18 Hz), beta2 (18.5–30 Hz), gamma1 (30.5–40 Hz), and gamma2 (40–49.5 Hz).

PSD features are computed for the same three electrode configurations as the temporal features. Each channel contributes one feature per frequency band, resulting in 525 features for the full-head configuration, 100 features for the reduced 10–20 configuration, and 50 features for the Broca–Wernicke configuration.

For comparability with prior EEG-to-text frameworks, the baseline Hilbert–Huang Transform (HHT) representation follows the original configuration used in previous work, which includes additional lower-frequency bands (theta1: 4–6 Hz, theta2: 6.5–8 Hz, and alpha1: 8.5–10 Hz).

Proposed Architecture: Following prior EEG-to-text decoding approaches [4, 12], a sequence-to-sequence architecture is employed to map word-level EEG representations to natural language text. Let $E = [e_1, e_2, \dots, e_K]$ denote the sequence of word-level EEG feature vectors extracted from a sentence. These features are first processed by a dedicated Transformer-based EEG encoder composed of six layers, each using five attention heads. The output is passed through a linear projection layer with ReLU activation to a 1024-dimensional space, matching the embedding dimension of the downstream language model. The projected embeddings are provided as input to a pre-trained BART-large encoder–decoder model [15], replacing the standard textual embeddings

at the BART encoder input. The overall architecture is illustrated in Fig. 3.

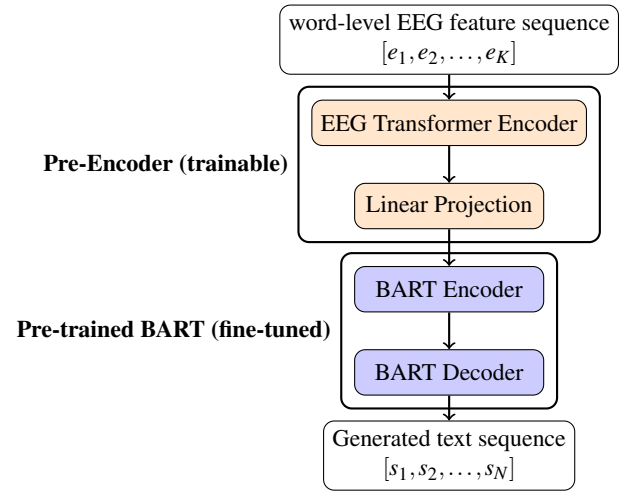


Figure 3: Overview of the EEG-to-text decoding architecture, illustrating the word-level EEG feature sequence, projection to the language model embedding space (Pre-Encoder), and pre-trained BART.

All model components are trained jointly for 30 epochs using the Adam optimiser with a learning rate of 2×10^{-5} and gradient clipping with a maximum ℓ_2 norm of 1.0. The ZuCo dataset is partitioned using an 80%–10%–10% split (train–development–test) with a fixed random seed of 42. The split is performed at the sentence level across the entire dataset. This setup corresponds to an intra-subject evaluation protocol, where samples from the same subjects may appear in both training and test sets, but no sentence appears in more than one split. The objective of this work is controlled representation-level comparison under a fixed evaluation setting. The complete set of hyperparameters is summarised in Table 2.

Table 2: Model configuration and experimental settings.

Parameter	Value
EEG Input Dimension	Representation-dependent
EEG Encoder	6-layer Transformer (5 heads)
Projection Layer	Linear + ReLU (1024 dimensions)
Language Model	BART-large (1024-dim embeddings)
Training Epochs	30
Batch Size	32
Beam Search Size	5
Repetition Penalty	5.0
Max Sequence Length	56 tokens

Inference and Evaluation: Model evaluation is conducted using autoregressive generation without teacher forcing during inference. No ground-truth tokens are provided to the decoder at test time. Model performance is evaluated using complementary automatic metrics commonly adopted in EEG-to-text studies. BLEU scores with n -grams ranging from 1 to 4 are reported to measure lex-

Table 3: Feature representation comparison under the SR+NR+NRv2 setting using the full-head configuration (sentence-level mean $\pm$ standard deviation).

Representation	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1 F1	ROUGE-1 P	ROUGE-1 R	WER
Baseline HHT	9.9 $\pm$ 5.9	2.2 $\pm$ 1.9	1.2 $\pm$ 0.8	0.9 $\pm$ 0.5	10.0 $\pm$ 7.0	11.3 $\pm$ 8.1	9.8 $\pm$ 7.4	119.9 $\pm$ 48.7
PSD	10.0 $\pm$ 6.3	2.3 $\pm$ 2.3	1.3 $\pm$ 1.3	0.9 $\pm$ 0.9	9.9 $\pm$ 7.2	12.0 $\pm$ 9.3	9.4 $\pm$ 7.5	112.7 $\pm$ 41.6*
Statistical	11.4 $\pm$ 6.6*	2.8 $\pm$ 3.1*	1.5 $\pm$ 1.6*	1.1 $\pm$ 1.1*	11.5 $\pm$ 7.8*	12.8 $\pm$ 9.0*	11.3 $\pm$ 7.8*	113.3 $\pm$ 35.8*

* $p < 0.05$ with respect to HHT features. Bold values indicate the best result for each metric.

ical overlap between generated and reference sentences. ROUGE scores are used to assess recall- and precision-oriented overlap at the sequence level. Word Error Rate (WER) is additionally employed to quantify token-level transcription errors. For readability, BLEU and ROUGE scores are reported as percentages. WER values may exceed 100% because autoregressive generation can produce insertion, deletion, and substitution errors whose cumulative count surpasses the number of reference tokens.

RESULTS

The experimental analysis compares EEG representations, channel configurations, and task composition under identical decoding conditions. The Hilbert–Huang Transform (HHT) representation is used as the reference baseline because it corresponds to the feature extraction strategy adopted in prior EEG-to-text frameworks [4, 12]. Results are reported as mean $\pm$ standard deviation across sentences, and statistical significance was assessed using paired bootstrap resampling with 10,000 iterations at the sentence level. Asterisks indicate statistically significant differences ($p < 0.05$). The reported standard deviations capture sentence-level variability caused by differences in segment duration, sentence complexity, and error propagation in autoregressive generation.

Impact of EEG Representation Type: To isolate representational effects, we compare three feature types under a controlled setting: (i) HHT features as a baseline representation, (ii) PSD features, and (iii) temporal statistical features proposed in this study.

All representations are evaluated under identical decoding conditions. Table 3 reports results using the full-head spatial configuration and the SR+NR+NRv2 task combination, allowing the effect of EEG representation to be analyzed while keeping spatial configuration and dataset composition fixed.

Temporal statistical features achieve the highest BLEU and ROUGE scores, indicating more stable lexical reconstruction under short EEG segments. In contrast, PSD features obtain the lowest WER, suggesting slightly lower token-level reconstruction error. The HHT baseline consistently produces lower overlap-based performance under the evaluated conditions.

Impact of Spatial Configuration: To assess the impact of spatial electrode selection, we analyze the SR+NR+NRv2 configuration across all spatial settings while keeping task composition fixed. Table 4 summarizes the results obtained using PSD and statistical feature

representations.

For PSD features, reduced spatial configurations achieve comparable or slightly higher overlap-based metrics than the full-head setup, with the 10–20 configuration obtaining the highest BLEU-1 and ROUGE-1 F1 scores and the Broca–Wernicke subset yielding the lowest WER. A similar trend appears for temporal statistical features: the 10–20 configuration produces the highest BLEU-1 score, the full-head configuration achieves the best ROUGE-1 F1, and the Broca–Wernicke subset again obtains the lowest WER. Overall, performance differences between spatial configurations remain small, suggesting that reducing the number of electrodes does not systematically degrade decoding performance.

Impact of Task Composition: We further analyzed the impact of task composition across feature representations using the full-head spatial configuration. This experimental design allows evaluation of the effect of progressively increasing dataset size and task heterogeneity while maintaining identical spatial conditions. Table 5 summarizes the obtained results.

Under the SR setting, temporal statistical features achieve the highest BLEU-1 and ROUGE-1 scores, while PSD features produce the lowest WER within this task configuration. When NR is incorporated, temporal statistical features achieve the strongest overall overlap-based performance, whereas both PSD and HHT representations produce slightly lower scores.

The inclusion of TSR changes the relative behavior across representations. Under this setting, PSD features achieve the highest ROUGE scores and the lowest WER, while temporal statistical features exhibit lower overlap-based performance compared with the SR+NR configuration.

Finally, the SR+NR+NRv2 configuration yields the strongest overall performance for statistical representations, with improvements across most BLEU and ROUGE metrics relative to HHT and PSD features.

Overall, the results indicate that decoding performance depends not only on dataset size, but also on the cognitive consistency of the reading tasks. Combining tasks with similar reading behavior generally produces more stable performance, whereas incorporating TSR introduces additional variability that may affect the temporal consistency of word-level EEG segments.

Table 4: Impact of spatial configuration under the SR+NR+NRv2 setting using PSD and temporal statistical features (sentence-level mean $\pm$ standard deviation).

PSD features								
Spatial Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1 F1	ROUGE-1 P	ROUGE-1 R	WER
Full-head	10.0 $\pm$ 6.3	2.3 $\pm$ 2.2	1.3 $\pm$ 1.3	0.9 $\pm$ 0.9	9.9 $\pm$ 7.2	12.0 $\pm$ 9.3	9.4 $\pm$ 7.5	112.7 $\pm$ 41.6
10–20	11.1 $\pm$ 6.9*	2.9 $\pm$ 3.4*	1.5 $\pm$ 2.2*	1.2 $\pm$ 1.7*	11.9 $\pm$ 8.3*	14.0 $\pm$ 10.4	11.9 $\pm$ 8.3*	111.7 $\pm$ 42.2
Broca–Wernicke	11.0 $\pm$ 6.1*	2.5 $\pm$ 2.4*	1.3 $\pm$ 0.9*	1.0 $\pm$ 0.5	10.8 $\pm$ 7.1*	12.3 $\pm$ 8.4	10.5 $\pm$ 7.8*	111.5 $\pm$ 36.5
Temporal statistical features								
Spatial Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1 F1	ROUGE-1 P	ROUGE-1 R	WER
Full-head	11.4 $\pm$ 6.6	2.8 $\pm$ 3.1	1.5 $\pm$ 1.6	1.1 $\pm$ 1.1	11.5 $\pm$ 7.8	12.8 $\pm$ 9.0	11.3 $\pm$ 7.8	113.3 $\pm$ 35.8
10–20	11.9 $\pm$ 6.5*	2.7 $\pm$ 2.3	1.4 $\pm$ 1.1	1.1 $\pm$ 0.8	11.2 $\pm$ 7.5	12.5 $\pm$ 8.7	11.0 $\pm$ 8.0	111.2 $\pm$ 33.9*
Broca–Wernicke	11.2 $\pm$ 6.7	2.8 $\pm$ 2.8	1.5 $\pm$ 1.4	1.1 $\pm$ 1.0	10.0 $\pm$ 8.0	12.8 $\pm$ 9.6	10.36 $\pm$ 7.9	107.6 $\pm$ 31.3*

* $p < 0.05$ with respect to the full-head configuration. Bold values indicate the best result for each metric within each feature representation.

Table 5: Impact of task composition across EEG feature representations using the full-head spatial configuration (sentence-level mean $\pm$ standard deviation).

Task	Representation	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1 F1	ROUGE-1 P	ROUGE-1 R	WER
SR	HHT	12.9 $\pm$ 6.8	3.0 $\pm$ 2.9	1.6 $\pm$ 1.8	1.2 $\pm$ 1.0	12.1 $\pm$ 7.4	13.9 $\pm$ 9.7	12.1 $\pm$ 8.1	128.0 $\pm$ 66.9
	PSD	12.3 $\pm$ 6.2	2.8 $\pm$ 2.4	1.5 $\pm$ 0.9	1.1 $\pm$ 0.5	11.5 $\pm$ 6.9	13.6 $\pm$ 9.2	11.1 $\pm$ 7.2	121.3 $\pm$ 57.8
	Statistical	13.5 $\pm$ 6.8*	3.3 $\pm$ 3.0	1.7 $\pm$ 1.6	1.2 $\pm$ 0.8	12.8 $\pm$ 7.3	14.8 $\pm$ 9.7	12.7 $\pm$ 7.9	124.8 $\pm$ 67.3
SR+NR	HHT	10.6 $\pm$ 6.4	2.4 $\pm$ 2.1	1.3 $\pm$ 0.9	0.9 $\pm$ 0.6	10.7 $\pm$ 7.4	13.0 $\pm$ 9.0	10.2 $\pm$ 7.9	120.8 $\pm$ 47.6
	PSD	11.2 $\pm$ 6.0	2.2 $\pm$ 1.7	1.2 $\pm$ 0.8	0.9 $\pm$ 0.5	10.5 $\pm$ 6.9	11.0 $\pm$ 8.4	11.2 $\pm$ 7.3*	115.0 $\pm$ 41.9*
	Statistical	11.5 $\pm$ 6.8*	2.8 $\pm$ 2.3	1.5 $\pm$ 1.1	1.1 $\pm$ 0.8	11.6 $\pm$ 7.7*	13.4 $\pm$ 9.5	11.1 $\pm$ 8.2	111.5 $\pm$ 37.6*
SR+NR+TSR	HHT	10.9 $\pm$ 6.1	2.6 $\pm$ 2.3	1.4 $\pm$ 1.1	1.0 $\pm$ 0.6	10.9 $\pm$ 7.1	13.2 $\pm$ 9.0	10.3 $\pm$ 7.9	109.0 $\pm$ 29.2
	PSD	9.9 $\pm$ 8.6	2.8 $\pm$ 3.5	1.4 $\pm$ 1.6	1.0 $\pm$ 1.0	12.4 $\pm$ 8.3*	16.3 $\pm$ 11.5*	11.0 $\pm$ 8.0*	107.0 $\pm$ 38.0*
	Statistical	10.4 $\pm$ 5.9	2.4 $\pm$ 2.0	1.3 $\pm$ 0.9	1.0 $\pm$ 0.5	10.6 $\pm$ 7.9	12.1 $\pm$ 9.8	10.5 $\pm$ 8.4	112.7 $\pm$ 36.9
SR+NR+NRv2	HHT	9.9 $\pm$ 5.9	2.2 $\pm$ 1.9	1.2 $\pm$ 0.8	0.9 $\pm$ 0.5	10.0 $\pm$ 7.0	11.3 $\pm$ 8.1	9.8 $\pm$ 7.4	119.9 $\pm$ 48.7
	PSD	10.0 $\pm$ 6.3	2.3 $\pm$ 2.3	1.3 $\pm$ 1.3	0.9 $\pm$ 0.9	9.9 $\pm$ 7.2	12.0 $\pm$ 9.3	9.4 $\pm$ 7.5	112.7 $\pm$ 41.6*
	Statistical	11.4 $\pm$ 6.6*	2.8 $\pm$ 3.1*	1.5 $\pm$ 1.6*	1.1 $\pm$ 1.1*	11.5 $\pm$ 7.8*	12.8 $\pm$ 9.1*	11.3 $\pm$ 7.8*	113.4 $\pm$ 35.8*

* $p < 0.05$ with respect to the HHT baseline within the corresponding task configuration. Bold values indicate the best result for each task configuration.

DISCUSSION

The results show that EEG representation, channel configuration, and task composition all influence EEG-to-text decoding performance, although the absolute scores remain modest. This is consistent with the difficulty of the task: word-level EEG segments are short, heterogeneous, and only indirectly aligned with linguistic structure. Under these conditions, the relative trends across experimental settings become more informative than the absolute metric values themselves.

One of the main findings of this study is that temporal statistical features perform consistently well across experimental conditions, suggesting that compact time-domain descriptors may provide more stable representations than decomposition-based spectral methods for gaze-aligned EEG segments.

Another important observation concerns spatial configuration. Reducing the number of EEG channels does not systematically degrade decoding performance, and reduced channel subsets remain competitive with the full-head configuration across several metrics.

Task composition also plays an important role. The best performance is not always obtained by simply increasing the amount of training data. Instead, combinations of tasks with similar reading behavior tend to produce more stable decoding patterns, whereas Task-Specific Reading introduces additional cognitive variability. This variability likely affects fixation dynamics and temporal align-

ment between EEG activity and linguistic processing.

The relatively low BLEU and ROUGE scores observed across all experiments should be interpreted in the context of the evaluation protocol and the intrinsic difficulty of EEG-to-text decoding. Unlike constrained decoding settings, the present work evaluates fully autoregressive open-vocabulary generation without teacher forcing. Under this setting, decoding errors propagate across generated tokens, substantially increasing task difficulty. In addition, the ZuCo datasets provide short and highly variable EEG segments aligned at the word level rather than sentence-level neural representations, which further limits decoding stability.

Although direct quantitative comparison with previous EEG-to-text studies remains difficult due to differences in inference protocols, evaluation settings, and dataset composition, the HHT representation used in this work follows the same feature extraction strategy adopted in previous EEG-to-text frameworks [4, 12]. Under this controlled setting, the proposed temporal statistical representations achieved competitive or superior overlap-based performance relative to the HHT baseline across several experimental configurations.

Importantly, the observed limitations of HHT representations should not be interpreted as a general limitation of adaptive spectral decomposition methods. Instead, the results suggest that these methods may become less stable when applied to gaze-aligned word-level EEG recordings with limited temporal context.

The present study should be interpreted as a controlled comparative analysis of EEG representations and spatial configurations under a fixed autoregressive decoding framework rather than as a direct state-of-the-art comparison. Taken together, the results suggest that future EEG-to-text systems should not focus exclusively on model complexity, since representation design, spatial parsimony, and cognitively consistent datasets may be equally important for improving neural language decoding performance.

CONCLUSIONS

This study investigated the influence of EEG representation, spatial channel configuration, and task composition on word-level EEG-to-text decoding using the ZuCo v1 and v2 datasets.

The results show that temporal statistical features provide competitive and stable decoding performance across several experimental settings, suggesting that compact time-domain descriptors may be more robust than decomposition-based spectral representations for short EEG segments.

The spatial analysis further demonstrates that reduced electrode subsets can achieve performance comparable to full-head EEG configurations. In particular, the 10–20 and Broca–Wernicke configurations remain competitive while substantially reducing spatial dimensionality.

The experiments also highlight the importance of task composition. Combining reading paradigms with similar cognitive characteristics generally improves decoding stability, whereas heterogeneous tasks such as Task-Specific Reading introduce additional variability that can negatively affect performance.

ACKNOWLEDGMENT

The first author of this work acknowledges the Secretaría de Educación, Ciencia, Tecnología e Innovación (SECIHTI) of Mexico for the postgraduate scholarship they received. This support made it possible to successfully complete this work.

REFERENCES

- [1] Pan J, Chen X, Ban N, He J, Chen J, Huang H. Advances in p300 brain–computer interface spellers: Toward paradigm design and performance evaluation. *Frontiers in Human Neuroscience*. 2022;16.
- [2] Kheddar H, Hemis M, Himeur Y. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*. 2024;109:102422.
- [3] Lopez-Bernal D, Balderas D, Ponce P, Molina A. A state-of-the-art review of EEG-based imagined speech decoding. *Frontiers in Human Neuroscience*. 2022;16:867281.
- [4] Liu H, Hajialigol D, Antony B, Han A, Wang X. EEG2text: Open vocabulary EEG-to-text translation with multi-view transformer. In: *Proc. 2024 IEEE Int. Conf. Big Data (BigData)*. IEEE: Washington, DC, USA, 2024, 1824–1833.
- [5] Hollenstein N, Rotsztein J, Troendle M, Pedroni A, Zhang C, Langer N. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*. 2018;5(1):180291.
- [6] Hollenstein N, Troendle M, Zhang C, Langer N. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In: *Proc. 12th Int. Conf. Language Resources and Evaluation (LREC)*. 2020.
- [7] Tang J, LeBel A, Jain S, Huth AG. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*. 2023;26:858–866.
- [8] Feng X, Feng X, Qin B, Liu T. Aligning semantic in brain and language: A curriculum contrastive method for EEG-to-text generation. *IEEE Trans. Neural Syst. Rehabil. Eng.*. 2023;31:3874–3883.
- [9] Zhou J, Duan Y, Chang YC, Wang YK, Lin CT. BELT: Bootstrapped EEG-to-language training by natural language supervision. *IEEE Trans. Neural Syst. Rehabil. Eng.*. 2024;32:3278–3288.
- [10] Amrani H, Micucci D, Napoletano P. Deep representation learning for open vocabulary EEG-to-text decoding. *IEEE J. Biomed. Health Inform.*. 2024.
- [11] Wang J, Song Z, Ma Z, Qiu X, Zhang M, Zhang Z. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL: Bangkok, Thailand, 2024, 7278–7292.
- [12] Jo H, Yang Y, Han J, Duan Y, Xiong H, Lee WH. Evaluating EEG-to-text models through noise-based performance analysis. *Scientific Reports*. 2025.
- [13] Zheng H, Wu Y, Qian T, Yue W, Wang X. Guiding LLMs to decode text via aligning semantics in EEG signals and language. *Expert Systems with Applications*. 2025;299:130300.
- [14] Jin Z, Li D, Huang S. A systematic review of eeg-based imagined speech decoding. *Applied Soft Computing*. 2025;183:113563.
- [15] Lewis M et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL: Online, Jul. 2020, 7871–7880.

Evaluating the Impact of Electrode Type and Montage on Classification Performance in Assistive P300-based BCIs

V. Caracci^{1,2,3}, A. Riccio², S. Petrinotti², A. Bigioni^{2,3}, S. Boesso²,
J. Toppi^{1,2}, F. Cincotti^{1,2}, D. Mattia²

¹Dep. of Computer, Control and Management Engineering, Sapienza University of Rome, Italy

²Neuroelectrical Imaging and BCI Laboratory, Fondazione Santa Lucia IRCCS, Rome, Italy

³Fondazione Neurone Onlus, Rome, Italy

E-mail: valentina.caracci@uniroma1.it

ABSTRACT: Severe motor disabilities from neurological conditions often prevent individuals from accessing modern Assistive Technologies (AT) for communication. Brain-Computer Interfaces (BCIs) bypass motor pathways, translating brain activity into device control. Among non-invasive BCIs, P300-based systems, exploiting the P300 Event-Related Potential, are widely investigated for communication: in the P300-speller users select letters by focusing on target stimuli, enabling communication for individuals with severe motor impairments. BCIs are therefore promising, but usability, setup complexity, and gel-based EEG caps limit integration. Traditional gel-based caps provide high-quality signals but are uncomfortable and slow to set up, while semi-dry headsets improve comfort and usability, but typically have suboptimal montages for P300-applications. This raises questions about the effects of electrode type and montage. This study compared a commercial semi-dry 7-channel headset for out-of-the-lab applications and a traditional wet EEG cap during P300-speller tasks. Classification performance was evaluated online and offline, providing insight into how electrode design and configuration influence the usability and effectiveness of BCIs.

INTRODUCTION

Severe motor disability is a common consequence of a wide range of neurological conditions, including traumatic brain injury, stroke, cerebral palsy, spinal cord injury, and neurodegenerative diseases. These conditions often result in deeply impaired communication and interaction, affecting quality of life and social participation [1]. Recent advances in Assistive Technologies (AT) have enabled people with disability to access mainstream digital communication through advanced input devices (i.e., eye-trackers, switches, joysticks, alternative keyboards [2]) and customizable software platforms to facilitate access to communication and interaction [3]. However, despite the maturity of commercial AT, for individuals with severe motor impairments these solutions often remain inaccessible due to the complexity of control, limited operability, or requirement for residual motor abilities [4]. Brain-Computer Interfaces (BCIs), particularly non-

invasive EEG-based systems, offer a promising alternative by establishing a direct channel for communication and control from the brain to external devices, bypassing muscular and peripheral nervous system pathways [5, 6]. Among non invasive BCIs, P300-based systems, exploiting the P300 Event-Related Potential (ERP) as control feature, are widely investigated for communication [1, 7]. The P300 ERP is a positive deflection in the EEG, occurring approximately 300 ms after a rare and task-relevant stimulus, reflecting the user's attentional orientation and cognitive processing [8]. The typical paradigm to elicit the P300 is the oddball, where frequent non-target and rare target stimuli are presented [8]. The oddball paradigm sets the basis for more sophisticated interfaces such as the P300-speller [9], which is one of the most extensively validated BCIs for communication. The user is presented with a matrix of letters flashing in rows and columns, and focuses on a specific character. Each flash containing the desired letter to be spelled represents a "rare target" eliciting the P300, while flashes of other letters act as "frequent non-targets". By evaluating the P300 following each flash, the system identifies the target letter as the intersection of the row and column with the largest P300 responses [9].

The acquisition of the EEG signal is a crucial step in BCI systems, as signal quality directly influences subsequent processing and overall performance. Traditional gel-based, or wet, electrodes are the gold-standard in hospitals and laboratories. Typically based on Ag/AgCl sensors with conductive gel, they ensure low and stable electrode-skin impedance and high-quality signals. However, their use requires skin preparation and conductive gel application, resulting in time-consuming setups performed by technicians, while users often experience discomfort and must wash their hair after recordings. These practical limitations have restricted gel-based EEG to laboratory settings [10, 11]. Dry electrodes represent an alternative, as they do not require gel or skin preparation, allowing faster and more user-friendly setups [11]. However, they often present higher impedances and lower signal quality, limiting their reliability [10]. Semi-dry electrodes have been proposed as a compromise between these approaches. They require only a small amount of

electrolyte fluid and combine the signal reliability of wet electrodes with the faster setup and improved comfort of dry systems. This enables stable recordings while reducing preparation and post-recording cleaning times [10]. Nonetheless, semi-dry caps often feature fixed montages, which may not be optimal for P300-based applications. Indeed, electrodes' placement plays a critical role in the classification performance of the system, and the optimal configuration depends on the specific use. P300-based BCIs benefit from including electrodes at positions with a prominent P300 response, namely the midline positions Fz, Cz and Pz [8, 12]. Also posterior and occipital electrodes provide supplementary discriminative information that improves target from non-target detection [13]. *Spencer et al.* (2001, [14]) reported that frontal, central, and posterior electrodes capture the majority of P300-related variance, identifying a component from Cz to Oz and P7 to P8 as accounting for the largest portion of variance, and a component from Cz to Fz accounting for the second largest portion [14].

Despite the potential of P300-based BCIs [1], they still cannot seamlessly integrate with commercial AT software. Their real-world use remains limited by performance variability, low selection speed and practical constraints related to EEG acquisition systems. Users and AT experts have emphasized the need for setups that are fast, reliable, comfortable, and easy to mount, supporting everyday use outside laboratory environments [15]. In response, portable EEG headsets based on dry and semi-dry electrodes can represent an alternative to improve usability and reduce preparation time. However, these systems often rely on predefined montages, limiting flexibility in electrode placement according to the specific requirements of the intended use.

The trade-off between the reliability and quality levels of the gold-standard gel caps, and the portability, comfort and ease of use of recent commercial headsets raises an important question: how do electrode type and montage affect BCI performance? In this study, the question was addressed by comparing two EEG headsets. Healthy participants performed a P300-speller task twice: first using a semi-dry commercial headset with a fixed 7-channel montage, emphasizing comfort and ease of setup but with an electrode configuration not specifically designed for P300 applications; then using a traditional wet EEG cap with an 8-channel montage commonly used in P300-BCI communication research [16], representing the gold-standard. This comparison allowed to evaluate whether the practical advantages of portable semi-dry systems can be achieved without substantially compromising BCI performance. The results provide preliminary insights into how headset design and electrode configuration may influence BCI performance and usability, helping to inform future device development and customization.

MATERIALS AND METHODS

Participants: EEG data were collected from a popula-

tion of 10 healthy volunteers (7F-3M, 28.8 ± 4.3 years old) according to the following inclusion criteria: *i*) no history of neurological or psychiatric diseases; *ii*) age > 18 years old. All the research was conducted in accordance with the declaration of Helsinki. The study protocol was approved by the Independent Ethics Committee of Fondazione Santa Lucia in Rome.

Experimental procedure: Subjects were administered with a traditional P300-speller paradigm [9], repeated twice during two experimental sessions acquired within the same day, 30 minutes apart. The two sessions comprised the same P300-speller protocol, administered through the BCI2000 framework [17], but differed in terms of EEG headset employed to acquire the signal (see section below). The order of the sessions was counter-balanced across participants. The stimulation protocol consisted of a 5x6 matrix with 30 alphabetic characters (Fig.1) sequentially intensified in rows and columns in pseudo-random order. Each stimulation flash had a duration of 120 ms, with an inter-stimulus interval (ISI) of 120 ms, obtaining a stimulus onset asynchrony (SOA) of 240 ms. The stimulation was presented through a 24" screen located 60 cm away from the subject's eyes. Participants were required to spell 9 predefined words of common use (i.e., 9 runs), selected from the Italian dictionary so that the target characters covered the most of the spelling matrix. All words were 5-characters long. Subjects were instructed to focus their attention on the letter to write (i.e., the target stimulus) and mentally count how many times the row and the column in which it was contained were intensified. Each target was intensified 16 times (i.e., 8 rows and 8 columns), corresponding to 8 stimulation sequences. The number of stimuli presented was kept constant during the whole experiment for all subjects, to ensure the consistency of stimulus presentation and allow for direct online and offline comparisons between participants. At the beginning of each trial (i.e., set of flashes to write a letter), the target character was cued at the top of the screen.

Each session consisted in a calibration phase with no feedback provided, and a spelling phase in online copy-mode. Within the copy-mode, the subject was allowed a fixed number of selections and was asked to write predefined words, with no possibility to correct any error. The EEG data acquired during the calibration process, comprising 4 words (20 selections: "BACIO", "POSTO", "METRO", "ZEBRA"), was used to compute the BCI classifier parameters of each subject, by applying a Stepwise Linear Discriminant Analysis (SWLDA) [16]. The classifier parameters were used to provide the online feedback in the copy-mode, consisting of 5 words (25 selections: "DANZA", "GUSTO", "MENTE", "VIOLA", "FIORE"). The spelling of the five words was completed in its entirety, and no criterion was established to stop the task based on performance.

EEG recordings: For each participant, two different EEG systems were used: *i*) the 7-channel commercial semi-dry X.on headset (Brain Products, Germany: F3,



Figure 1: P300-speller 5x6 matrix used for the stimulation protocol. A row or column intensifies for 120 ms every 120 ms. The word at the top of the window is the text to be spelled.

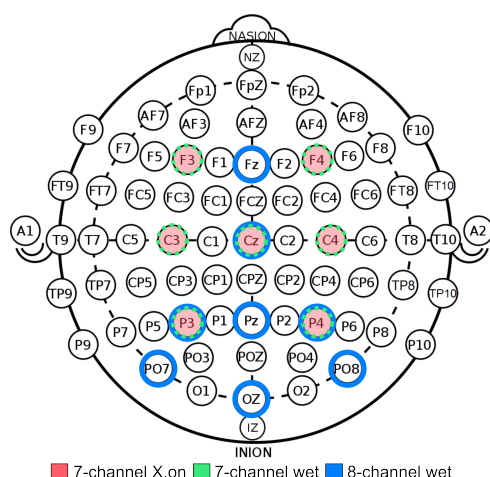


Figure 2: Electrode configurations stored and analyzed. The 12-channel wet cap was the union of the 8-channel wet and 7-channel wet subsets. The online analysis involved the 7-channel X.on and the 8-channel wet montages, while for the offline analysis also the 7-channel wet montage was considered.

F4, C3, Cz, C4, P3, P4, reference and ground on the right ear lobe; sampling rate 250 Hz); *ii*) a 12-channel wet EEG cap (EasyCap-LiveAmp, Brain Products, Germany: F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, PO7, PO8, Oz, reference and ground on the right ear lobe; sampling rate 250 Hz) based on the International 10–20 system. The 12 channels used in case (*ii*) represent the union of the X.on montage with the 8-channel configuration commonly adopted in the literature for P300-spellers [16]. Despite 12 channels being acquired and stored, only the 8 channels' subset from *Krusienski et al.* (2006, [16]) was considered during the online protocol and transmitted to the BCI2000 framework in the calibration and the copy-mode phases. The wet 7-channel's configuration was then only evaluated through an offline analysis by considering the proper subset of channels from those acquired. In Fig.2, the three configurations are highlighted. To obtain a wet montage fully comparable with the fixed X.on configuration for the offline analysis, the ground electrode was moved from the right mastoid to the right ear lobe.

Data Analysis: The performance of the two BCI systems was evaluated in terms of *online accuracy* and *of-*

line information transfer rate (ITR) [18]. The online accuracy was defined as the percentage of correct selections in the copy-mode, expressed as the ratio between the total number of correct selections across the five copy-mode runs. The EEG data acquired and managed online by the BCI2000 framework was band-pass filtered [0.1, 70] Hz, with a notch filter at 50 Hz, decimated by a factor 10 and segmented into 800 ms epochs from stimulus onset. In the copy-mode runs, the processed and segmented signal was then averaged for each stimulus type prior to the SWLDA classification. The same procedure was adopted for data acquired with both the 7-channel semi-dry X.on montage and the 8-channel wet montage [16], extracted as a subset of the 12-channel configuration. Two distributions of the online accuracy index were therefore obtained: *a*) the 7-channel semi-dry X.on, *b*) the 8-channel wet from the EasyCap-LiveAmp system.

The ITR metric was evaluated via an offline analysis of the four calibration runs, carried out in MATLAB environment (MATLAB R2023b). The ITR (bits/min) is an offline metric defined as the amount of information transferred per unit of time, incorporating information on both speed and accuracy. In fact, maintaining a constant number of stimulation sequences in the calibration and copy-mode phases improves the standardization of the experimental task between participants, but does not account for individual differences in their ability to process the stimuli, not fully reflecting realistic conditions of use in terms of speed and accuracy. The ITR combines the accuracy level, time and optimal number of sequences to reach it [7]. Three configurations (i.e., datasets) were obtained from the recordings through the two systems: *a*) the 7-channel semi-dry X.on; *b*) the 7-channel wet subset reproducing the X.on electrode's placement; *c*) the 8-channel wet subset reported in the literature [16]. To avoid feedback-related effects influencing the evaluation of the ITR, the offline analysis was performed on data from the calibration phase. In fact, during the online copy-mode, feedback was computed using the 8-channel wet montage transmitted to the BCI2000 framework. Therefore, using the same data to evaluate the performance of the 7-channel wet subset offline would have introduced a bias in the comparison. For each dataset and stimulus sequence (i.e., $n=1, 2, \dots, 8$), an offline accuracy value was computed using a leave-one-run-out cross-validation technique applied to the four calibration runs: at each iteration of the cross-validation process, three runs were used as the training set, and one run was used as the test set. Within the training phases, data segments from the three runs were concatenated by channel for each intensification, generating a single feature vector for training the classifiers and extract the SWLDA weights [16]. The mean offline accuracy for each subject and stimulation sequence was obtained by averaging the values resulting from the four iterations. Starting from the offline accuracy values, the ITR was computed by applying the bit-rate definition [18] and multiplying the bit-rate by the selection speed (selection/minute). Fi-

nally, the highest individual ITR value for each subject, with an offline accuracy equal or above the 70% level, was considered as a representation of the subject's overall performance. As outlined in the literature, the 70% accuracy has been identified as the threshold sufficient for effective communication with a BCI [19].

Three ITR distributions were obtained: one for each of the aforementioned configurations (i.e., *7-channel semi-dry X.on*, *7-channel wet*, *8-channel wet*). The data used for the offline analysis underwent the same processing steps performed online by the BCI2000 framework.

Statistical Analysis: Online accuracy distributions from the two copy-mode phases (i.e., *7-channel semi-dry X.on*, *8-channel wet*) were statistically compared to evaluate differences between the two recording systems. A non-parametric Wilcoxon-matched pair test was performed (factor: RECORDING SYSTEM, 2 levels). To investigate the effects of the electrode type (factor: ELECTRODE TYPE, 2 levels) and positioning (factor: MONTAGE, 2 levels) separately, the offline ITR distributions underwent statistical analysis. The ELECTRODE TYPE factor was evaluated through a Wilcoxon-matched pair test between the *7-channel semi-dry X.on* and the *7-channel wet* distributions. This controlled for the montage effect, since both configurations consisted of the same channels, thus the electrode type could be evaluated independently. Conversely, the MONTAGE factor was investigated applying the Wilcoxon-matched pair test to the *7-channel wet* and the *8-channel wet* distributions. This controlled for the electrode type, since both configurations were acquired from the same electrodes at the same time, enabling the evaluation of the montage without confounds. All statistical analyses were performed in the MATLAB environment (MATLAB R2023b), and the level of significance was set to $\alpha = 0.05$.

RESULTS

Results of the online task are displayed in Fig.3a, presenting the online accuracy boxplots from the copy-mode phase. The distribution obtained for the *7-channel semi-dry X.on* exhibited a greater dispersion, with a median value around 85%, while the accuracy levels from the *8-channel wet* system showed a narrower distribution centered close to 98%. The statistical test revealed a significant difference between the two distributions ($p = 0.0039$), with the 7-channel semi-dry system exhibiting lower performance than the 8-channel wet system.

Concerning the offline analysis conducted on the four calibration runs to separately assess the effect of the electrode type and the montage, the results are summarized in Fig.3b, which shows the distributions of the maximum ITR in cross-validation. The ITR values were computed using EEG data from the *7-channel semi-dry X.on*, the *7-channel wet*, and the *8-channel wet* configurations. The 8-channel wet condition exhibited a wide distribution, with the median value centered at 55 bit/min. The distribution obtained from the 7-channel wet configuration

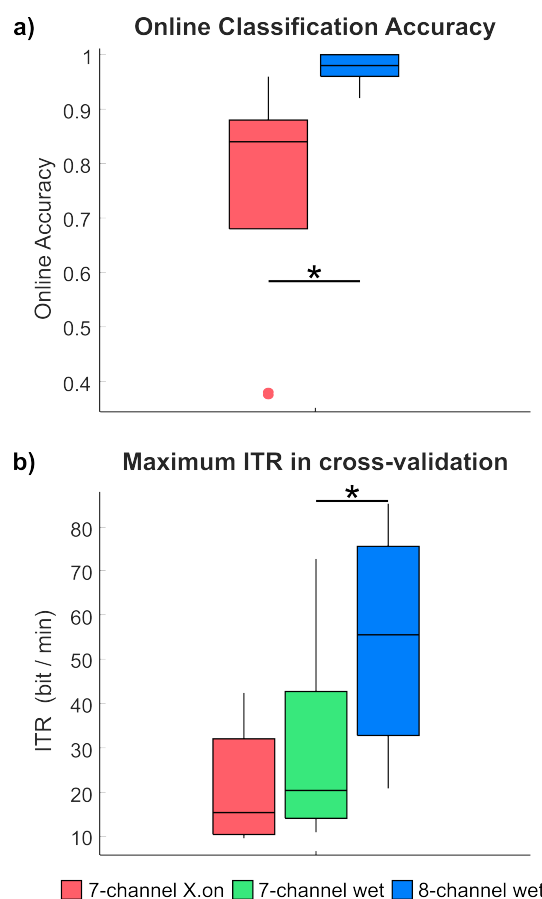


Figure 3: P300-speller classification performance: a) Online accuracy distributions computed for the five copy-mode runs for the *7-channel semi-dry X.on* and the *8-channel wet* configurations; b) Maximum offline ITR distributions, computed in cross-validation considering the four calibration runs for the *7-channel semi-dry X.on*, the *7-channel wet* and the *8-channel wet* configurations. For each subject, the highest ITR value, with an offline accuracy equal or above 70%, was taken into account. The symbol * highlights significant differences between two levels ($\alpha = 0.05$).

was narrower and shifted toward lower ITR values, with the median value around 20 bit/min. Finally, the values obtained from the data acquired through the 7-channel semi-dry system, exhibited a distribution that was similar to the previous condition with respect to dispersion and range. However, the median value was shifted slightly downward, centered around 15 bit/min. With regard to the ELECTRODE TYPE factor (2 levels: *7-channel semi-dry X.on*, *7-channel wet*), no significant differences were identified ($p = 0.46$). Conversely, a significant effect was highlighted for the MONTAGE factor (2 levels: *8-channel wet*, *7-channel wet*), with the 8-channel configuration exhibiting higher ITR values compared to the 7-channel condition ($p = 0.0039$).

DISCUSSION

The present study examined the impact of different EEG recording systems on the performance of a P300-based BCI in healthy subjects. A P300-speller task was pre-

sented twice. First, the X.on semi-dry commercial headset was used, featuring a fixed 7-channel montage that prioritizes comfort and ease of setup, but not specifically designed for P300 applications. Subsequently, a traditional 12-channel wet EEG cap was employed, transmitting the 8-channel montage widely regarded as the gold-standard to the BCI framework [16]. The influence on performance was evaluated through a dual approach: an online and an offline analysis to investigate multiple effects. From the online accuracy general information regarding the recording system was obtained. This hindered the independent evaluation of the electrode type and the montage effects, which were separately investigated within the offline analysis, where EEG datasets acquired from the semi-dry headset and subsets of channels from the wet headset were considered. This approach enabled the comparison of two types of electrode (i.e., semi-dry vs wet) while maintaining the same montage, and between two montages (i.e., 7-channel vs 8-channel) under the same electrode type condition.

The findings indicated that both recording systems allowed the attainment of an average across-subjects 70% online accuracy level, deemed sufficient for effective communication with a BCI [19]. However, recording the EEG signal through the 8-channel montage from *Krusienski et al.* [16], representing the gold-standard for P300-oriented protocols in the literature, yielded higher performance levels in both online and offline analyses. When assessing whether the performance drop in the 7-channel semi-dry X.on cap was due to the electrode type and/or positioning, no significant effect was found for the electrode type factor (i.e., *7-channel semi-dry X.on* vs *7-channel wet*). The result is consistent with previous literature comparing wet and semi-dry electrodes: semi-dry electrodes achieve impedances similar to wet electrodes [10, 20], and deliver comparable signal quality [11], thereby bridging the gap between high-fidelity laboratory setups and portable, user-friendly commercial devices [10, 20]. Semi-dry electrodes avoid hair contamination and require no post-recording hair washing, highlighting their applicability for frequent or prolonged use in home and clinical environments [10]. This supports the reliability of semi-dry systems and their potential for real-world BCI applications, where comfort, rapid setup, and ease of use are essential. In particular, their usability advantages may facilitate the adoption of BCIs in AT contexts, supporting practical communication solutions for individuals with severe motor disabilities.

In contrast, a significant effect emerged for the montage factor (i.e., *7-channel wet* vs *8-channel wet*), with better performance observed when selecting the 8-channel gold-standard subset compared to the X.on-like 7-channel subset. This finding aligns with previous studies that evaluated the effects of spatial channel selection on classification performance [13]. Indeed, supplementing the midline P300 recording locations (i.e., Fz, Cz, Pz [8]) with posterior regions significantly improves the classification performance of the P300-speller [13].

Posterior-occipital electrodes (i.e., PO7, PO8, Oz) capture additional information reflecting visual processing in both focal and peripheral fields. This contributes to P300 response detection and improves classification performance when used in conjunction with the classical P300 feature space at Fz, Cz, Pz [13]. The posterior features are hypothesized to contribute to classification due to the differences observed in visual evoked responses depending on the target location (i.e., in the focal or peripheral field): a prominent event-related negativity between 200 and 250 ms can be found in posterior-occipital target responses reflecting the focal field [12]. In the 7-channel X.on configuration, Fz was replaced by F3 and F4, only Cz was retained among midline electrodes, and posterior-occipital sites (i.e., PO7, PO8, Oz) were omitted, explaining the observed reduction in classification performance compared to the 8-channel montage. These results highlight that electrode spatial configuration can be as critical as electrode type for effective BCI performance.

Taken together, these findings provide a comprehensive perspective on the interplay between electrode technology and montage design. Semi-dry electrodes have been confirmed as a reliable alternative to wet electrodes, offering comfort, ease of setup, and practical suitability for extended use. However, the limitations associated with fixed montages not specifically designed for P300-applications, as demonstrated with the 7-channel X.on headset, indicate that adequate spatial coverage, particularly including posterior-occipital sites, is essential for maximizing P300 classification performance. This highlights the need for customizable headset designs that can combine the practical advantages of semi-dry technology with flexible electrode placement, supporting both research applications and real-world AT deployments. Collaboration between researchers and commercial developers will be key to achieving these solutions and ensuring that BCIs provide reliable, user-friendly communication channels for people with severe motor disabilities. This will expand access to BCI communication and promote autonomy, social participation, and inclusion.

CONCLUSION

This study investigated the effects of electrode type and montage on the performance of a P300-based BCI. Semi-dry electrodes were confirmed to provide signal quality comparable to traditional wet systems, while offering increased comfort, faster setup, and ease of use, making them a practical option for both laboratory and real-world applications. In contrast, electrode montage significantly influenced classification performance, with the 8-channel gold-standard configuration outperforming the 7-channel fixed semi-dry montage, highlighting the importance of adequate spatial coverage, particularly including posterior-occipital sites. These findings suggest that future BCI systems for AT should combine semi-dry electrodes with montages optimized for P300-applications, to balance usability, comfort, and classification accuracy.

Such designs could enable reliable, efficient communication for individuals with severe motor impairments, supporting daily interaction and social participation. Optimizing both electrode type and montage is crucial for translating laboratory BCI research into practical, user-friendly assistive solutions, helping to reduce barriers and enhance autonomy for users with severe disabilities.

ACKNOWLEDGMENT

This work was partially supported the Italian Ministry of Health (PNRR POC-2023- 12377627) and Sapienza University of Rome (LEAF 2023- RM123188F229EC72; Avvio alla Ricerca 2025- AR125199BE5D0A07).

REFERENCES

- [1] Riccio A et al. Interfacing brain with computer to improve communication and rehabilitation after brain damage. *Progress in brain research*. 2016;228:357–387.
- [2] Nisbet P. Integrating assistive technologies: Current practices and future possibilities. *Medical engineering & physics*. 1996;18(3):193–202.
- [3] Andrich R, Mathiassen NE, Hoogerwerf EJ, Gelderblom GJ. Service delivery systems for assistive technology in europe: An aaate/eastin position paper. *Technology and Disability*. 2013;25(3):127–146.
- [4] Pitt KM, Brumberg JS, Pitt AR. Considering augmentative and alternative communication research for brain-computer interface practice. *Assistive technology outcomes and benefits*. 2019;13(1):1.
- [5] Wolpaw JR, Millán JdR, Ramsey NF. Brain-computer interfaces: Definitions and principles. *Handbook of clinical neurology*. 2020;168:15–23.
- [6] Galiotta V et al. P300-based brain-computer interface for communication in assistive technology centres: Influence of users' profile on bci access. *Journal of Neural Engineering*. 2025;22(3):036044.
- [7] Riccio A et al. On the relationship between attention processing and p300-based brain computer interface control in amyotrophic lateral sclerosis. *Frontiers in human neuroscience*. 2018;12:165.
- [8] Polich J. Updating p300: An integrative theory of p3a and p3b. *Clinical neurophysiology*. 2007;118(10):2128–2148.
- [9] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*. 1988;70(6):510–523.
- [10] Li GL, Wu JT, Xia YH, He QG, Jin HG. Review of semi-dry electrodes for eeg recording. *Journal of Neural Engineering*. 2020;17(5):051004.
- [11] Yuan H et al. State of the art of non-invasive electrode materials for brain-computer interface. *Micromachines*. 2021;12(12):1521.
- [12] Sellers E, Krusienski D, McFarland D, Wolpaw J. Non-invasive brain-computer interface research at the wadsworth center. In: *Toward Brain-Computer Interfacing*. Jul. 2007.
- [13] Krusienski D, Sellers E, McFarland D, Vaughan T, Wolpaw J. Toward enhanced p300 speller performance. *Journal of Neuroscience Methods*. 2008;167(1):15–21.
- [14] Spencer KM, Dien J, Donchin E. Spatiotemporal analysis of the late erp responses to deviant stimuli. *Psychophysiology*. 2001;38 2:343–58.
- [15] Zickler C et al. A brain-computer interface as input channel for a standard assistive technology software. *Clinical EEG and neuroscience*. 2011;42(4):236–244.
- [16] Krusienski DJ et al. A comparison of classification techniques for the p300 speller. *Journal of neural engineering*. 2006;3(4):299–305.
- [17] Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR. Bci2000: A general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*. 2004;51(6):1034–1043.
- [18] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*. 2002;113(6):767–791.
- [19] Kübler A, Neumann N, Wilhelm B, Hinterberger T, Birbaumer N. Predictability of brain-computer communication. *Journal of Psychophysiology*. 2004;18(2/3):121–129.
- [20] Li G, Wang S, Duan YY. Towards conductive-gel-free electrodes: Understanding the wet electrode, semi-dry electrode and dry electrode-skin interface impedance using electrochemical impedance spectroscopy fitting. *Sensors and Actuators B: Chemical*. 2018;277:250–260.

First Real-Time Use of Quasi-Movements for Exoskeleton Control: Comparison with MI-BCI in Healthy Participants and Pilot Application in Stroke Rehabilitation

A.S. Yashin¹, L.A. Ivanov², A.A. Kondur², D.A. Bobrov³, E.V. Slyunkova², G.G. Gabuzov³, E.S. Lukyanov³, Y.G. Shevtsova^{1,4}, A.N. Vasilyev^{1,4}, P.D. Bobrov³, S.V. Kotov², S.L. Shishkin¹

¹ MEG Center, Moscow State University of Psychology and Education, Moscow, Russia

² Vladimirsky Moscow Regional Research and Clinical Institute, Moscow, Russia

³ Pirogov Russian National Research Medical University, Moscow, Russia

⁴ Department of Human and Animal Physiology, Lomonosov Moscow State University, Moscow, Russia

E-mail: yashinart1996@gmail.com

ABSTRACT: Quasi-movements (QM) are movement attempts minimized to produce no overt movement or muscle activation. Unlike kinesthetic motor imagery (MI), they may not require mastering an abstract mental skill; unlike conventional attempted movements, they may reduce risks of abnormal motor synergies. QM-based BCI training may therefore be a useful clinical alternative for sensorimotor BCI rehabilitation. We evaluated QM for real-time control of a hand-exoskeleton BCI. First, 19 healthy participants completed QM- and MI-based control sessions with online three-class EEG classification and post-session ratings. QM achieved performance comparable to MI in online accuracy and subjective ratings, although it required closer muscle-tension monitoring. Preferences for QM versus MI were split evenly, indicating inter-individual variability. Second, four chronic stroke patients completed repeated QM-BCI training over multiple days. Patients quickly understood the instructions and produced classification accuracies consistently above a nominal three-class chance reference. Clinical changes were heterogeneous but generally favorable. QM may therefore provide a promising additional strategy for exoskeleton-based BCI rehabilitation.

INTRODUCTION

Motor imagery (MI) with feedback from the brain-computer interfaces (BCIs) has repeatedly been shown to be a promising supplement to post-stroke rehabilitation. However, despite encouraging results, MI-BCI training is still rarely used in clinical practice [1-3]. One major limitation is the difficulty of teaching patients to perform MI reliably enough to activate an MI-BCI within the limited training time available in clinical settings [1,3]. Attempted movements or movement attempts (MA) represent a potentially more natural task for paralyzed patients. Attempting to perform a familiar motor action may require less training than MI, or no training at all;

thus, reliable BCI control can be achieved faster. Previous studies reported that MA can produce more pronounced EEG effects than MI [4,5] and may enable higher BCI classification accuracy [5]. Moreover, BCI training based on MA has been associated with improved clinical outcomes compared with MI-based training [6]. However, it could be argued that encouraging patients with paresis or plegia to repeatedly attempt overt movements may lead to undesirable side effects, including increased spasticity, activation of pathological motor synergies, or frustration caused by repeated failure to produce movement.

One possible alternative is the use of quasi-movements (QM). QM are movement attempts which are so weak that they produce no overt movement and no detectable electromyographic (EMG) activity [7]. In this task, participants attempt to perform the instructed movement while intentionally minimizing the strength of the attempt. Importantly, QM elicit EEG changes that were found to be even stronger than MI-related effects in offline EEG studies [7,8], suggesting that patients may be able to activate BCIs through such minimized movement attempts.

Because QM do not involve overt movement or detectable muscle activation, they may reduce the risk of side effects associated with overt movement attempts.

However, to our knowledge, QM have not yet been applied to online BCI control of rehabilitation devices.

In this study we, for the first time, investigate the feasibility of using quasi-movements for real-time BCI control of a rehabilitation hand exoskeleton:

- First, we compare QM-based and MI-based BCI control in healthy participants.
- Second, we perform a pilot clinical evaluation of QM-BCI in chronic stroke patients.

MATERIALS AND METHODS I

Participants: The experimental comparison of QM and MI was conducted with 19 healthy volunteers (15 f,

4 m; 14 right-handed; age range 19-33 years, median = 22 years) who provided written informed consent. All experimental procedures had local ethical approval.

Equipment and signal acquisition: EEG and EMG signals were acquired at a sampling rate of 500 Hz with 24-bit resolution, using a 0-300 Hz bandwidth, via a DC-coupled NVX136 EEG amplifier (Medical Computer Systems, Moscow, Russia).

EEG was recorded from 64 built-in gel-based Ag/AgCl electrodes arranged according to the international 10-10 system. For online classification, data from 52 electrodes were used. Signals were recorded in a unipolar montage, with the ground electrode positioned at Fpz. Surface EMG electrodes were placed on both forearms. On each forearm, one electrode was positioned over the belly of the *m. extensor digitorum communis*, and the second was placed 2-3 cm distal to it, along the longitudinal axis of the muscle. The signals were converted into two bipolar EMG channels and processed using the Teager-Kaiser energy operator for online visualization. EEG and EMG electrode impedances were kept below 20 k Ω .

To provide kinesthetic feedback, we used a hand exoskeleton system for both hands with two degrees of freedom and external software control (ExoHand-M; Fig. 1A). This system is an extension of a setup previously employed in post-stroke rehabilitation using kinesthetic motor imagery [9]. The exoskeletons were mounted on supports positioned on either side of a comfortable chair in which the participant was seated (Fig. 1B). Participants' hands were secured in the exoskeletons.

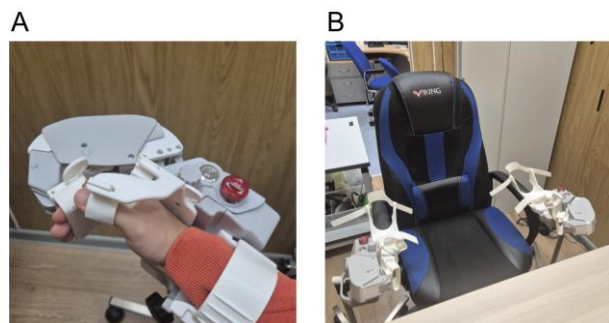


Figure 1: (A) Fixation of the hand in the hand exoskeleton; (B) the experimental chair with hand exoskeletons mounted on both sides.

Design: After electrode placement and before the main part of the experiment, participants completed a training phase in both quasi-movements (QM) and kinesthetic motor imagery (MI). The main part of the experiment comprised three conditions: (1) controlling the hand exoskeletons using QM; (2) controlling the exoskeletons using MI; and (3) a passive-movement condition, in which participants performed no task and the exoskeletons opened automatically with simulated classifier operation. The passive-movement condition was included for EEG-analysis outside the scope of the present paper. The order of the three conditions was randomized across participants. Each condition comprised two sessions, each lasting approximately 6 minutes. After each session, participants completed a

questionnaire.

Procedure: The QM training protocol followed procedures previously described in the literature [7,8]. Relative to earlier studies, we used a different target movement – hand opening (extension of four fingers) – and participants performed QM with both hands. When interacting with participants, QM were referred to as “micro-movements.” This terminology was adopted to reduce suspicion: participants were not informed that, under ideal performance, the EMG signal during micro-movements is indistinguishable from rest.

Training comprised three stages: (i) gradually reducing the amplitude of the overt movement in the dominant hand while monitoring EMG, (ii) refining small movements into QM under experimenter supervision, and (iii) transferring the skill to the non-dominant hand. Training in MI was also conducted in multiple stages: (i) working with a finger-extension resistance device to elicit and label kinesthetic descriptors using the dominant hand, (ii) synthesizing a kinesthetic image based on these descriptors, and (iii) repeating the procedure with the non-dominant hand. Total QM training lasted approximately 15 minutes as did MI training.

Each session in each condition consisted of four blocks of trials. The first block was used for initial classifier training, whereas the remaining three blocks involved online BCI operation. After each block, the classifier was updated using the data accumulated up to that point, so online control did not rely solely on the first block. Each block contained eight trials: in half of the trials participants were instructed to rest, and in the other half they were instructed to interact with one of the two exoskeletons (motor trials). Trial type was cued visually using three arrows presented on the screen (Fig. 2): an upward arrow indicated a rest trial, a leftward arrow indicated the left exoskeleton, and a rightward arrow indicated the right exoskeleton. Each rest trial was always followed by a motor trial; within each block, the order of left- versus right-hand commands was randomized. Arrow color indicated the current task phase: blue corresponded to preparation and green corresponded to execution. The preparation phase lasted 2 s and the execution phase lasted 9 s.



Figure 2: Visual arrow cues used during the sessions. (A) a blue arrow indicating preparation to perform the task with the left hand; (B) a green arrow indicating active task execution; (C) Illumination of a circle indicating successful BCI classification.

During online BCI blocks, successful detection of a command triggered opening of the corresponding exoskeleton. Opening speed depended on the estimated

probability of the target class: opening occurred during those time intervals within a trial in which the classifier output favored the target class among the three classes. In addition, a circle presented at the center of the screen provided visual feedback; its brightness reflected classification confidence for the current command. At the end of each trial, the exoskeleton closed and the on-screen circle disappeared. Participants were instructed to perform micro-movements or MI regularly, without compromising task quality and without inducing excessive muscle tension. Muscle activity was monitored using EMG signals.

EEG classifier: Classifier training used EEG data band-pass filtered from 5 to 30 Hz. The signal was segmented into non-overlapping 250 ms epochs; epochs with at least 10% artifacts were excluded. Covariance matrices were estimated from the remaining data separately for each class.

For each pair of classes, Common Spatial Pattern (CSP) spatial filters were computed from the corresponding covariance matrices. Four CSP components were selected per class pair. Components were first selected if the variance ratio of the corresponding CSP component for the given class pair fell within 0.65-0.95. If fewer than four components were obtained by this criterion, the remaining components were selected from the unused set by the absolute value of the log variance ratio. Spatial filters selected across all class pairs were then combined and used to construct feature vectors.

Each feature vector was computed from a 1 s window advanced in steps of 250 ms; features were the log-variances of the selected CSP components. Feature vectors were classified using regularized Linear Discriminant Analysis (shrinkage LDA). The classifier was updated after each block. During online feedback, classifier outputs were also generated from 1 s windows advanced in 250 ms steps.

Questionnaires: Participants completed a set of questionnaires throughout the experiment. After each session in the QM and MI conditions, participants completed several scale items. Using a 0-1 scale, they rated Task Success (“How successful were you at performing micro-movements [imagining movements]?”), from “Could not do it at all” to “Could do it well”), Difficulty (“How difficult was it for you to perform micro-movements [imagine movements]?”), from “Not difficult at all” to “Extremely difficult”), Tension Monitoring (“How closely did you monitor muscle tension while the exoskeletons were opening?”), from “Did not monitor it at all” to “Monitored it constantly”), Compatibility (“How well did micro-movements [motor imagery] fit with the opening of the exoskeletons?”), from “Did not fit at all” to “Fit well”), and their sense of control over the left and right exoskeleton movements (“Did not control it at all” to “Controlled it completely”). For follow-up questionnaires within the same condition, participants were shown their previous responses on the scales, enabling them to evaluate the new session relative to earlier sessions in the same mode.

At the end of the experiment, participants completed comparative scale items: PerformDiff (“In which condition were you better able to open the exoskeletons?”, from “Micro-movements” to “Motor imagery”), ComfortDiff (“Which condition felt more comfortable?”), and CompatDiff (“Which task fit better with opening the exoskeletons?”). They also provided open-ended responses describing their preferred condition and the perceived advantages of the two control conditions. Questionnaires were presented on a computer screen using PsychoPy [10].

Statistical analyses: The following quantitative dependent variables were analyzed: (1) EEG-classification accuracy during interface operation and (2) numerical questionnaire ratings. Classification accuracy was computed for each session as the proportion of correct recognitions of the target command class (left/right command and “rest”) based on online blocks, with the distribution of misclassifications across classes additionally inspected using the confusion matrix. When interpreting recognition quality, we used 1/3 (≈ 0.33), the theoretical chance level for a balanced three-class task, as a nominal reference for evaluating whether the feedback loop produced task-contingent responses. Questionnaire outcomes were values on continuous rating scales with two verbal anchors.

For the two sessions in the MI condition and the first two sessions in the QM condition, a two-way repeated-measures ANOVA was conducted with factors Condition and Session number; computations were performed in Python using the SciPy library. Post-hoc analyses used t-tests with FDR correction for multiple comparisons; statistical significance was determined after correction (p-values are reported as p_{FDR}). Open-ended responses were treated as qualitative data to support interpretation of the quantitative findings.

RESULTS I

Classification accuracy: Mean three-class EEG classification accuracy was comparable between the QM and MI conditions. In the MI condition, mean accuracy was 0.49 ± 0.10 in Session 1 and 0.50 ± 0.09 in Session 2; in the QM condition, the corresponding values were 0.49 ± 0.10 and 0.54 ± 0.10 (Fig. 3). The two-way repeated-measures ANOVA for the first two sessions showed no statistically significant effects of Condition, $F(1, 18) = 3.138$, $p = 0.093$, Session, $F(1, 18) = 3.635$, $p = 0.073$, or their interaction, $F(1, 18) = 1.884$, $p = 0.187$. Overall, classifier performance remained broadly stable across tasks and sessions.

Questionnaire analyses: For **Task Success** (Fig. 4A), the ANOVA revealed no main effect of Condition, $F(1, 18) = 0.487$, $p = 0.494$, and no Condition $\times$ Session interaction, $F(1, 18) = 0.004$, $p = 0.953$. The main effect of Session was significant, $F(1, 18) = 6.287$, $p = 0.022$, reflecting higher perceived success in S2 than in S1. However, FDR-corrected post-hoc tests did not yield significant pairwise differences. For **Task Difficulty** (Fig. 4B), there was no main effect of Condition, $F(1, 18)$

= 0.160, $p = 0.694$, and no Condition $\times$ Session interaction, $F(1, 18) = 0.001$, $p = 0.977$. The main effect of Session was significant, $F(1, 18) = 6.655$, $p = 0.019$, reflecting lower perceived difficulty in S2 than in S1. No FDR-corrected post-hoc comparison reached significance. For **Tension Monitoring** (Fig. 4C), the ANOVA revealed a significant main effect of Condition, $F(1, 18) = 7.878$, $p = 0.012$, reflecting higher tension monitoring in the QM condition than in the MI condition. The main effect of Session and the Condition $\times$ Session interaction were not significant, $F(1, 18) = 0.841$, $p = 0.371$, and $F(1, 18) = 3.300$, $p = 0.086$, respectively. FDR-corrected simple-effects analyses showed that the between-condition difference was significant in S2, $t(18) = -3.763$, $p_{\text{FDR}} = 0.006$, but not in S1. For **Task-Device Compatibility** (Fig. 4D), no significant effects were observed: Condition, $F(1, 18) = 0.003$, $p = 0.957$; Session, $F(1, 18) = 0.101$, $p = 0.755$; and Condition $\times$ Session, $F(1, 18) = 0.394$, $p = 0.538$. For **Perceived Control of the Exoskeleton**, averaged across the two hand-specific scales (Fig. 4E), no significant effects were observed: Condition, $F(1, 18) = 0.031$, $p = 0.863$; Session, $F(1, 18) = 0.726$, $p = 0.406$; and Condition $\times$ Session, $F(1, 18) = 0.016$, $p = 0.902$.

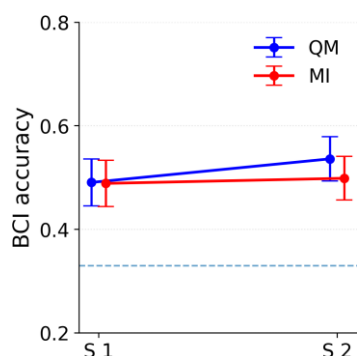


Figure 3: Group average three-class EEG classification accuracy for QM and MI in Sessions 1 and 2 (M with 95% CI). The dashed line indicates the nominal chance reference for a balanced three-class task (1/3).

In the final questionnaire, participants compared the two active conditions on three comparative scales: **PerformDiff**, **ComfortDiff**, and **CompatDiff** (Fig. 4F). On all three scales, 0.5 indicated no preference, values below 0.5 favored QM, and values above 0.5 favored MI. Mean PerformDiff was 0.45 ± 0.31 (M $\pm$ SD) and did not differ from 0.5, $t(18) = -0.653$, $p = 0.522$. Mean ComfortDiff was 0.47 ± 0.34 and also did not differ from 0.5, $t(18) = -0.420$, $p = 0.679$. Mean CompatDiff was 0.37 ± 0.28 and did not significantly differ from 0.5, $t(18) = -2.030$, $p = 0.057$.

Open-ended responses indicated marked inter-individual variability: 9 participants preferred QM, 9 preferred MI, and 1 reported no preference. Participants who favored QM often described them as easier to execute ($n = 6$), whereas four participants who preferred MI noted that it reduced the need to regulate muscle tension compared to QM.

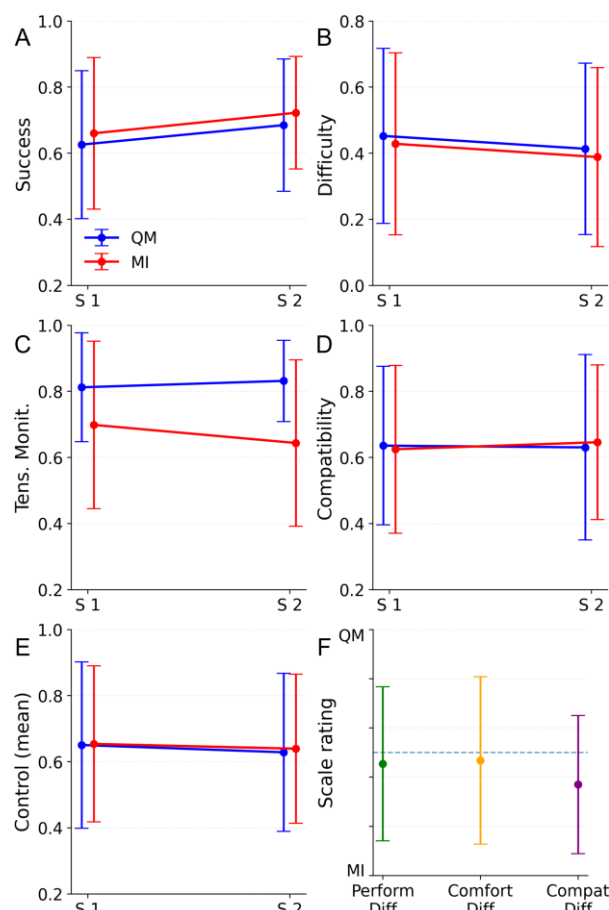


Figure 4: Group average questionnaire ratings (M $\pm$ SD).

MATERIALS AND METHODS II

Participants: The clinical sample comprised four patients with ischemic stroke (2 f, 2 m; all right-handed; age range 46-73, median 57) admitted to the Vladimirsky Moscow Regional Research and Clinical Institute (MONIKI). Time since stroke was 48, 11, 7, and 4 months for Patients 1-4, respectively. All patients were right-handed and had right-hemisphere lesions; two exhibited severe paresis and two moderate paresis. A local ethical committee approved the procedures. All patients signed informed consent.

Equipment and signal acquisition: Signal acquisition was identical to that used in the healthy sample, except that EEG was recorded with an NVX-52 amplifier from 32 scalp electrodes arranged according to the extended international 10-20 system (all used for classification). EMG was recorded from both the extensor muscles and the finger flexors (*m. flexor digitorum communis*).

Procedure: QM training in patients was shortened and generally adapted to reduce fatigue. Patients first performed overt hand-opening movements with the non-paretic hand and were asked to gradually reduce movement amplitude while observing the EMG signal. The EMG feedback was then hidden, and patients were instructed to further minimize muscle tension, although they were not required to suppress EMG activity completely. This was done to avoid excessive fatigue. Patients then performed the same minimized motor task

with the paretic hand, initially with EMG feedback available. The entire QM training procedure took only a few minutes. We did not use a repeated-measures QM–MI design in patients to avoid task-related confusion and potential disruption of the rehabilitation procedure. Although patients understood the instructions, they remained cognitively vulnerable, and introducing multiple motor tasks could have increased cognitive load. Patients completed QM sessions over 6–7 days (3–5 sessions per day depending on patient condition). Before and after the course of the procedures, patients underwent a clinical assessment that included the Mini-Mental State Examination (MMSE), the Montreal Cognitive Assessment (MoCA), the upper-extremity subsection of the Fugl-Meyer Assessment (FMA-UE), the Hospital Anxiety and Depression Scale (HADS), and the Barthel Index of Activities of Daily Living.

RESULTS II

Classification accuracy: Three-class classification accuracy across patient sessions is shown in Fig. 6. Mean accuracy across all sessions was 0.53 ± 0.12 for Patient 1, 0.50 ± 0.09 for Patient 2, 0.53 ± 0.10 for Patient 3 and 0.47 ± 0.09 for Patient 4.

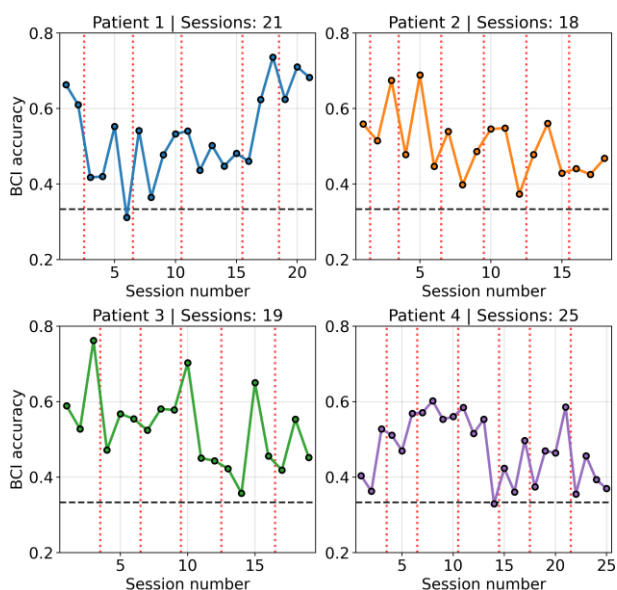


Figure 6: Three-class EEG classification accuracy across QM sessions for each patient. Points indicate session-wise BCI accuracy; the dashed horizontal line indicates the nominal chance reference for a balanced three-class task (1/3). Red vertical lines separate training days.

Clinical assessment: Descriptive inspection of the clinical assessments in the four patients indicated a heterogeneous but generally favorable pattern from admission to discharge (see Table 1). The most noticeable changes were observed in upper-limb motor function, while cognitive and functional measures showed only modest shifts overall. Notably, cognitive performance improved by an average of 2.75 points on the MoCA, allowing patients to move out of the

dementia range. Emotional status changed little, with no marked dynamics on the HADS scale.

Table 1: Clinical assessment results.

Assessment	At admission Mdn (range)	At discharge Mdn (range)
MMSE	25 (22-28)	n/a
MoCA	23.5 (18-25)	25 (23-28)
FMA-UE	29 (11-54)	35 (12-58)
HADS anxiety	8 (3-19)	7.5 (3-17)
HADS depression	5.5 (2-9)	5 (2-13)
Barthel Index	95 (70-100)	97.5 (95-100)

DISCUSSION

The present study had two main goals. First, we sought to determine, for the first time, whether quasi-movements (QM) can support real-time BCI control of a rehabilitation hand exoskeleton and how this control compares with conventional kinesthetic motor imagery (MI) in healthy participants, both in terms of classification accuracy and subjective experience. Second, we conducted a pilot clinical evaluation to examine whether QM-based exoskeleton training is feasible in chronic stroke patients.

Note that QM in this study did not follow the strict “classical” definition [7] that requires that no EMG activity was evoked. Interaction with a moving exoskeleton seemed to provoke some EMG activation even when participants did not report any muscle tension. However, we previously did not find a difference between EEG effects of such “quasi-quasi-movements” and “true” QM [8]. In the current study, therefore, we considered any movement attempt intentionally minimized up to the limit as QM. The relation between EMG and EEG in the new data will be further investigated and reported elsewhere.

Our results provide initial evidence that QM can serve as a viable sensorimotor task for exoskeleton-based BCI rehabilitation. In healthy participants, QM yielded online three-class classification accuracy and subjective ratings broadly comparable to those obtained with MI, suggesting that this task can support closed-loop BCI interaction without an evident loss of performance relative to the standard MI approach. At the same time, the higher ratings for monitoring of muscle tension indicate that QM is not simply interchangeable with MI at the phenomenological level, but places somewhat different self-regulatory demands on the user.

The comparative ratings and open-ended responses further point to substantial inter-individual variability. Preferences were divided almost equally between QM and MI, and participants described opposite advantages of the two modes: QM was often experienced as easier to execute, whereas MI reduced the need to regulate residual muscular effort. This pattern suggests that the main practical value of QM may lie not in universal superiority over MI, but in expanding the set of sensorimotor strategies available for BCI training. For

some users, especially those who struggle with purely imagery-based control, QM may provide a more accessible route to effective interaction with the exoskeleton.

The pilot clinical evaluation strengthens this interpretation. Although the patient sample was small and does not allow statistical inference, the descriptive findings indicate that repeated QM-based training was feasible in chronic stroke patients across multiple daily sessions and successive days of rehabilitation. The clinical scales showed a heterogeneous but generally favorable admission-to-discharge pattern, with the clearest changes observed in upper-limb motor function. At the same time, these clinical changes should be interpreted with caution, since the present pilot was not designed to isolate the specific therapeutic contribution of QM-BCI training. Importantly, patients understood the task instructions at the start of the course and consistently produced classification accuracies above the nominal three-class chance reference.

The healthy-participant study, however, support a personalized rather than one-size-fits-all view of choosing the patient's task in the BCI rehabilitation. One possible approach to be explored in future studies is to suggest patients, from time to time, choose between QM and MI according to their own preference. MI has certain advantages over MA, for example, it is possible to imagine complex movements at all their stages, while attempts are linked to the start of a movement, and this is applied to QM, which are a subtype of MA. Moreover, it is yet unknown if repeated training with QM-BCI is as beneficial for the post-stroke recovering as MI-BCI procedures. Nevertheless, our studies already indicate that QM is a good candidate for a task from which BCI-based rehabilitation can be started.

A key limitation of the clinical part of the study is that patients performed only QM tasks; future work will compare QM- and MI-based rehabilitation protocols in patient groups. Another limitation is that non-sensorimotor contributions to EEG classification cannot be fully excluded: the paradigm used abrupt visual arrow cues, no dedicated analysis separated sensorimotor activity from cue-onset visual responses, and ocular artifacts were handled only by the general amplitude- and distribution-based rejection procedure rather than by specific correction. Because CSP can in principle exploit any class-related variance, residual visual-evoked or ocular contributions remain possible. However, classification was restricted to the 5–30 Hz range during the execution phase, cues were centrally presented, and no gaze-dependent response was required.

CONCLUSION

Quasi-movements appear to be a feasible new control strategy for exoskeleton-based sensorimotor BCI rehabilitation. In healthy participants, they performed similarly to motor imagery in online accuracy and most subjective measures. The pilot clinical evaluation suggested that repeated QM-based training can be

implemented in chronic stroke patients and can support BCI control. While MI was not studied in patients, pronounced variability in performance and preferring QM vs. MI observed in healthy participants justify further patient-centered studies of QM as part of an individualized rehabilitation protocol rather than as a universal replacement for motor imagery.

REFERENCES

- [1] Simon C, Bolton DAE, Kennedy NC, Soekadar SR, Ruddy KL. Challenges and opportunities for the future of brain-computer interface in neurorehabilitation. *Front. Neurosci.* 2021;15: 699428
- [2] Colucci A, Vermehren M, Cavallo A, Angerhöfer C, Peekhaus N, Zollo L, Kim W-S, Paik N-J, Soekadar SR. Brain-computer interface-controlled exoskeletons in clinical neurorehabilitation: ready or not. *Neurorehabil. Neural Repair* 2022;36(12): 747-756
- [3] Mokienko OA, Lyukmanov RKh, Bobrov PD, Suponeva NA, Piradov MA. Brain-computer interfaces for upper limb motor recovery after stroke: current status and development prospects (review). *Sovrem. Tehnol. Med.* 2023;15(6): 63
- [4] Blokland Y, Spyrou L, Lerou J, Mourisse J, Scheffer GJ, van Geffen GJ, Farquhar J, Bruhn J. Detection of attempted movement from the EEG during neuromuscular block: proof of principle study in awake volunteers. *Sci. Rep.* 2015;5: 12815
- [5] Chen S, Shu X, Wang H, Ding L, Fu J, Jia J. The differences between motor attempt and motor imagery in brain-computer interface accuracy and event-related desynchronization of patients with hemiplegia. *Front. Neurobot.* 2021;15: 706630
- [6] Mansour S, Ang KK, Nair KPS, Phua KS, Arvaneh M. Efficacy of brain-computer interface and the impact of its design characteristics on poststroke upper-limb rehabilitation: a systematic review and meta-analysis of randomized controlled trials. *Clin. EEG Neurosci.* 2022;53(1): 79-90
- [7] Nikulin VV, Hohlefeld FU, Jacobs AM, Curio G. Quasi-movements: a novel motor-cognitive phenomenon. *Neuropsychologia* 2008;46(2): 727-742
- [8] Vasilyev AN, Yashin AS, Shishkin SL. Quasi-movements and “quasi-quasi-movements”: does residual muscle activation matter? *Life* 2022;13(2): 303
- [9] Frolov AA, Mokienko O, Lyukmanov R, Biryukova E, Kotov S, Turbina L, Nadareyshvily G, Bushkova Y. Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface (BCI)-controlled hand exoskeleton: a randomized controlled multicenter trial. *Front. Neurosci.* 2017;11: 400
- [10] Pierce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 2019;51: 195-203

Volitional and Non-Volitional Motor Tasks Encode Distinct EEG Dynamics: Implications for Paradigm-Specific BCI Decoder Design

F. Morales-Magallón¹, E. Bojorges-Valdez¹, D.C. González-González²

¹ Engineering Studies for Innovation Department, Universidad Iberoamericana Ciudad de México, Mexico City, Mexico

² Geriatric Evaluation Center, Universidad Iberoamericana Ciudad de México, Mexico City, Mexico

E-mail: fernandomoralesm01@gmail.com, erik.bojorges@ibero.mx

ABSTRACT: This work explores differences in brain dynamics during the execution of voluntary and non-voluntary knee-joint movements (using an isokinetic dynamometer) by analyzing distance matrices constructed from a d -dimensional reconstructed phase-space of instantaneous band-power time series (iBPTs). Two condition-specific decoders were developed: a volitional (**Vol**) model, calibrated exclusively on active motor execution, and a non-volitional (**NVol**) model trained on machine-assisted movement. The nonlinear features derived from the phase-space distance matrix successfully encode condition-specific dynamical structures, yielding classification metrics consistent with current BCI standards, even for data acquired across a several-day interval. The **Vol** model achieved higher performance when classifying voluntary data (AUC: 0.772 ± 0.128) than non-voluntary data (AUC: 0.613 ± 0.159), whereas the **NVol** model demonstrated consistent performance across both conditions (AUC: 0.694 ± 0.119 for voluntary and AUC: 0.697 ± 0.100 for non-voluntary). A Linear Discriminant Analysis (LDA) classifier, combining the outputs of both models, revealed condition-dependent decoder contributions, confirming that volitional and non-volitional motor states elicit separable dynamical regimes in the EEG signal. These results indicate that although transfer learning from overt motor execution enables reliable classification of non-voluntary data, the underlying dynamical differences lead to degraded performance. Consequently, while calibration on overt movement may result in reduced accuracy in covert scenarios, the use of covert-movement data may provide more representative performance metrics for online implementations.

INTRODUCTION

A large body of EEG-based brain-computer interface (BCI) research relies on the extraction of features derived from Event-Related Desynchronization (ERD) and Event-Related Synchronization (ERS), quantified across frequency bands associated with motor activity [1, 2]. The μ (8-13 Hz) and β (13-30 Hz) bands, localized predominantly over the sensorimotor cortex, are the most widely employed [2–6], given their well-established sensitivity to both motor and somatosensory processes. The

γ band has additionally been implicated in movement onset facilitation [7], broadening the spectral repertoire relevant to motor decoding.

A common assumption underlying many BCI systems is that the neural dynamics of motor imagery (MI) and motor execution (ME) are sufficiently similar—exhibiting comparable data distributions and dynamical processes, encompassing both cortical regions and frequency band dynamics—to support cross-condition generalization. That is, models trained on ME data can be transferred to MI tasks without substantial performance loss [8]. However, this assumption is increasingly challenged. ERD/ERS-based classifiers consistently plateau at approximately 70% accuracy [8], even without accounting for the BCI “illiteracy” population. This suggests a non-trivial gap between overt and imagined movement that spectral features alone cannot bridge. Assistive technologies, such as therapeutic robotics or human-assisted rehabilitation, allow for the execution of non-volitional limb movements that enhance neuroplasticity by promoting the activation of afferent pathways [9]. This scenario provides an MI-like environment where the user observes limb motion without corresponding efferent pathway activation; however, the mechanisms underlying overt and covert motion differ substantially [10]. We argue that these inter-condition differences constitute a fundamental dynamical distinction that linear spectral methods are structurally incapable of resolving.

Nonlinear dynamical approaches have been proposed as a means to capture motor-related EEG information that remains undetectable via conventional frequency-domain analyses [11, 12]. Methods grounded in phase-space reconstruction—such as Recurrence Quantification Analysis (RQA)—have demonstrated sensitivity to subtle, condition-specific changes in neural dynamics that elude linear features [11, 12].

Within this framework, the instantaneous band-power time series (iBPTs) provides a tractable representation that preserves the temporal structure of band-specific power fluctuations while enabling estimation of nonlinear features via distance matrices in reconstructed phase-space.

In this work, we evaluate the performance of paradigm-specific decoders, trained for volitional (**Vol**) and non-

volitional (**NVol**) limb control, in decoding movement-related cortical activity. Two independent CNN-Transformer models are trained on mutually exclusive datasets corresponding to each motor condition. The relative contribution of each decoder is subsequently quantified via a Linear Discriminant Analysis (LDA) model, providing a direct empirical measure of representational separability between **Vol** and **NVol** neural states. This framework offers both a principled test of the MI/ME equivalence assumption and a practical foundation for designing condition-aware decoders in BCI systems.

MATERIALS AND METHODS

Experimental Paradigm:

The experimental paradigm and data acquisition followed the protocol described in a previous work [13]. Fifteen healthy participants (eight males, seven females; mean age 23 ± 3 years) were recruited. The experimental protocol was approved by the Ethics Committee of Universidad Iberoamericana (Folio 221), and all participants provided written informed consent prior to enrollment.

Participants performed three knee movement exercises using an isokinetic dynamometer (System 4 Quick-Set, Biodex, USA): two volitional conditions at different load levels (10% and 5% of maximum voluntary torque) and one non-volitional (assisted) movement condition. Maximum torque was determined individually for each subject during isometric contractions in both flexion and extension prior to the experimental sessions.

Each participant completed three sessions, with each session adhering to a structured experimental protocol: Assessment of maximum isometric torque in flexion and extension; 4–6 cycles of volitional movement at 10% of maximum torque (**VM10**): 40 s of activity followed by 20 s of rest; 4–6 cycles of volitional movement at 5% of maximum torque (**VM5**): 40 s of activity followed by 20 s of rest; 4–6 cycles of non-volitional movement (**NVM**): 40 s of activity followed by 20 s of rest.

Signal Acquisition:

Electrophysiological and kinematic signals were simultaneously recorded using two synchronized g.USBamp amplifiers (g.tec Medical Engineering, Austria).

EEG signals were acquired with a 16-channel montage following the extended 10–20 international system, using the following electrode positions: FC3, FCz, FC4, CP3, C3, C1, Cz, C2, C4, CP4, P3, Pz, P4, O1, Oz, and O2 as seen in Fig 1, A. The ground electrode was placed at Fpz and the right earlobe served as the common reference.

The second amplifier was dedicated to kinematic signals from the isokinetic dynamometer, including joint angular velocity, position, and torque. This channel also recorded surface electromyography (EMG) signals to assess muscular activity and to verify the absence of voluntary muscle activation during the **NVol** condition.

All signals were digitized at a sampling rate of 4800 samples per second. Initial signal conditioning included a band-pass filter from 0.1 to 2000 Hz and a 60 Hz notch

filter to attenuate power-line interference. EEG signals were preprocessed using Independent Component Analysis to remove artifacts using the mne package [14, 15]. To increase the signal-to-noise ratio of the EEG signals, it was used a “common average reference”. EEG signals were subsequently downsampled to 500 Hz for further processing.

Feature Extraction:

EEG signals were processed through a filter bank of fourth-order Butterworth band-pass filters to extract three physiologically relevant frequency bands: μ , β , and γ . The filtered signals were subsequently squared sample-wise to obtain the Instantaneous Band Power Time Series (iBPTs) shown in Fig 1 B.

Phase-Space distance matrices:

Recordings were segmented into windows of 2 seconds length with no overlapping. Within each window, the band power time series of the seven selected channels (FC3, FCz, C3, C1, C2, Cz, CP3) were embedded using a time-delay embedding with $\tau = 100$ ms and embedding dimension $d = 3$, yielding a 21-dimensional phase-space representation according to Takens theorem [16]. Each space was built from the projection of the sample n , $n - \tau$ and $n - 2\tau$ of each channel. In total, three 21-dimensional spaces were built one per frequency band. Euclidean distances between all pairs of embedded points within each window were computed to produce an unthresholded weighted distance matrix. The matrices derived from the β , μ , and γ bands were then stacked to form a multichannel representation of brain dynamics, which served as the input for a deep learning architecture. Figure 1B illustrates representative examples of the phase-space and distance matrices derived via this method.

Model Architecture:

Each paradigm-specific decoder utilized a CNN-Transformer architecture designed to extract dynamical structures from the distance-pattern image representation.

The input image was processed by a convolutional patch extractor comprising three successive Conv2D–Batch Normalization blocks. The first two Conv2D layers applied 32 filters each with 3×3 kernels, same padding, and linear activations; the third applied 15 filters with a 3×3 kernel and ReLU activation. A MaxPooling2D layer with pool size 4×4 then down-sampled the feature maps, projecting them into a sequence of non-overlapping patch embeddings with projection dimension 15.

A positional encoding layer was applied to the patch sequence to preserve spatial order. The encoded sequence was then passed through three stacked Transformer blocks, each comprising a Layer Normalization layer, a Multi-Head Self-Attention (MHSA) module with $h = 2$ attention heads. This design enables the model to capture global dependencies across distant regions of the distance pattern matrix, reflecting the complex temporal dynamics inherent in the EEG signal.

Following the final Layer Normalization, a Global Average Pooling layer aggregated the sequence into a fixed-

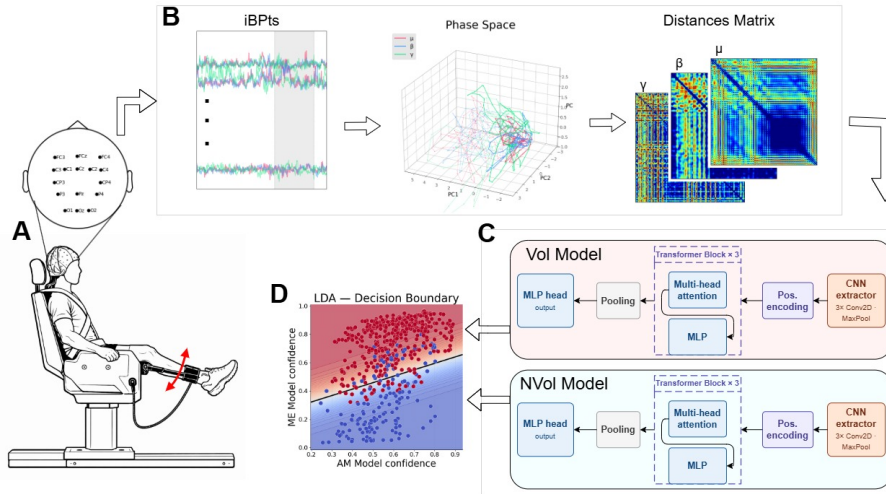


Figure 1: Schematic of the BCI pipeline. **A.** Experimental setup showing the isokinetic machine, right leg movement, and EEG cap with electrode configuration. **B.** Signal processing and feature extraction: cleaned EEG signals are used to compute iBPTs for the μ , β , and γ bands; the center panel shows the phase-space representation of the d -dimensional embedding used to extract Euclidean distances and Distances Matrix of each band. **C.** Two paradigm-specific models: Volitional (**Vol**, top) and Non-volitional (**NVol**, bottom), sharing the same architecture but trained on different datasets. **D.** Linear Discriminant Analysis (LDA) showing the contribution of both models to the true label.

length feature vector. This vector was passed through two fully connected Dense-Dropout blocks before a final Dense layer with softmax activation for binary classification (Activity vs. Rest).

Paradigm-Specific Training:

Two independent decoders were trained on mutually exclusive datasets: the **Vol** model was trained exclusively on volitional motor task recordings (**VM10** condition), while the **NVol** model was trained exclusively on non-volitional passive movement recordings (**NVM** condition). This paradigm-specific training strategy was adopted to preserve the distinct neural representations associated with each motor condition, rather than conflating them into a single shared representation.

Data from sessions S1 and S2 were used exclusively for model development. Files were partitioned at the window level into training (65%), validation (15%), and internal test (20%) subsets using stratified random sampling, with the validation subset used for early stopping and hyperparameter monitoring. Session S3 was held out entirely prior to any model development and served as the independent evaluation set for all reported performance metrics, ensuring no overlap between training data and final assessment.

Both the **Vol** and **NVol** decoders were trained once using pooled data across all subjects, yielding subject-independent models. This design choice reflects the assumption that the dynamical structures captured by the distance-pattern representation are consistent across individuals.

Assessment of Model Contribution:

The scalar output of each paradigm-specific decoder was normalized via z-score standardization on the training data and the held-out validation session, data was used as a two-dimensional feature vector for a Linear Discrimi-

nant Analysis (LDA) fusion classifier, trained per subject on sessions training data for S1 and S2. The resulting discriminant coefficients provide a direct measure of the relative contribution of each decoder to the final classification decision, computed as:

$$w_k = \frac{|\beta_k|}{|\beta_{\text{Vol}}| + |\beta_{\text{NVol}}|} \times 100\% \quad (1)$$

where β_k denotes the LDA coefficient associated with decoder k . LDA was evaluated on the held-out session S3, consistent with the evaluation protocol of the paradigm-specific decoders.

Statistical Analysis:

Classification performance was assessed using three metrics computed per subject on the held-out session S3: Area Under the ROC Curve (AUC), accuracy (ACC), and precision. These metrics were evaluated independently for the **Vol** model and **NVol** model.

Pairwise comparisons between **Vol** and **NVol** were performed using the Wilcoxon signed-rank test for paired samples ($\alpha = 0.05$), with subjects as the pairing unit. Four comparisons were conducted: **Vol** vs **NVol** within each motor condition (**NVM**, **VM10**), and cross-condition comparisons for each model separately. A Bonferroni correction was applied to all p-values to account for multiple comparisons.

Cross-condition comparisons within each model (**VM10** vs **NVM** for **Vol**, and **VM10** vs **NVM** for **NVol**) were performed independently using the Wilcoxon signed-rank test ($\alpha = 0.05$), with no correction for multiple comparisons applied.

Relative LDA weights on each model (**Vol** and **NVol**) were compared using the Wilcoxon signed-rank test for paired samples. Comparison was applied independently for **NVM** and **VM10** conditions.

RESULTS

Figures 2a and 2b illustrate the model predictions on the validation session of a representative subject, for the **NVM** and **VM10** conditions respectively. In both cases, the condition-specific model yielded higher performance on its corresponding task, while exhibiting reduced metrics when applied to the alternative condition.

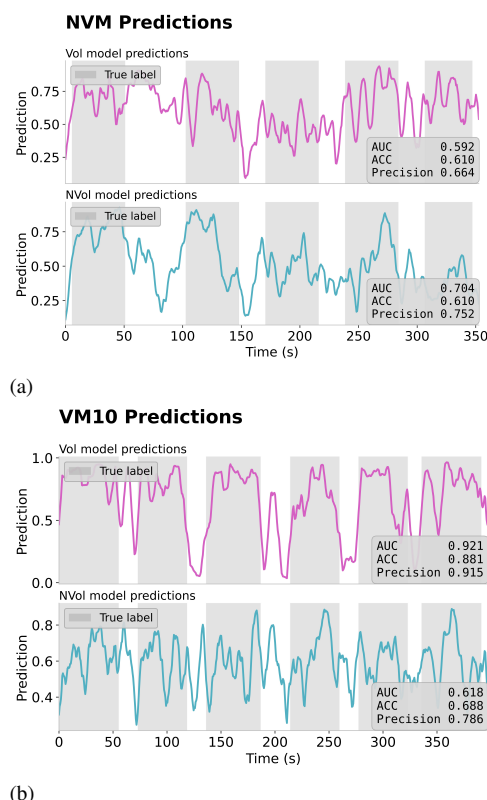


Figure 2: Model predictions on validation recordings. Dark magenta corresponds to the **Vol** model predictions and dark cyan to the **NVol** model predictions. In each plot, the gray shaded region represents the true activity label. AUC, Accuracy, and Precision scores are shown at the bottom.

Figure 3 presents the within-condition validation metrics for the **NVM** and **VM10** recordings across all subjects using both models. Across all metrics, the **NVol** model achieves superior performance in classifying **NVM** data, whereas the **Vol** model performs better on **VM10** data. This behavior demonstrates that models constructed for a specific task are well-suited for similar data, even when those data were not acquired on the same day. The figure also shows the cross-dataset comparison of model performance; p-values are reported after Bonferroni correction for multiple comparisons, where the **Vol** model approaches statistical significance in AUC, with higher mean values observed for the corresponding model dataset pairs. However, no statistically significant differences were found.

Table 1 presents a cross-condition numerical comparison, evaluating how each model generalizes when applied to a recording outside its training condition. The **NVol** model

exhibited consistent performance across both **VM10** and **NVM** recordings; the only statistically significant difference was found in Precision, reflecting slightly higher precision on its native **NVM** task. In contrast, the **Vol** model showed a more pronounced performance gap between conditions: while it performed well on its native **VM10** recording, its metrics declined notably on **NVM** data, with statistically significant differences found in AUC and ACC metrics. This suggests limited transferability of the learned parameters to the non-volitional condition.

Table 1: Comparison of model performance across **VM10** and **NVM** recordings. Values are mean $\pm$ SD. P-values from Wilcoxon signed-rank test; no correction for multiple comparisons applied.

Model	Metric	VM10	NVM	pval
Vol	AUC	0.772 ± 0.128	0.613 ± 0.159	0.0166
	ACC	0.692 ± 0.105	0.588 ± 0.099	0.0245
	Precision	0.825 ± 0.097	0.789 ± 0.088	0.3258
NVol	AUC	0.694 ± 0.119	0.697 ± 0.100	0.9032
	ACC	0.644 ± 0.091	0.630 ± 0.087	0.8077
	Precision	0.785 ± 0.073	0.833 ± 0.060	0.0494

Figure 4 displays the relative LDA weights, reflecting the contribution of each model to the fusion classifier across both recording conditions. A statistically significant difference between model contributions was observed in both conditions, with the **VM10** recording yielding a higher level of significance ($p < 0.05$) than the **NVM** recording. Notably, for the **VM10** data, all weights from the **Vol** model achieve higher values than those of the **NVol** model. Similarly, for the **NVM** data, the **NVol** weights are larger than those of the opposing model, although this trend is not absolute like that observed in the **Vol-VM10** pairing.

DISCUSSION

The results suggest that the two paradigm-specific decoders capture distinct dynamical signatures of motor-related cortical activity derived from phase-space analysis. As illustrated in Figure 3 and Table 1, model-data comparisons yield superior performance when each model is applied to its respective training condition—specifically, the **NVol** model on **NVM** data and the **Vol** model on **VM10** data. In general, the **NVol** model exhibits less dispersion across all evaluated metrics, suggesting that it maintains a more stable performance when capturing both volitional and non-volitional motor dynamics.

The cross-condition comparison in Table 1 further supports this interpretation: **NVol** predictions remained largely consistent across both **VM10** and **NVM** conditions, implying that the dynamical patterns captured during passive movement overlap substantially with those present during active execution. In contrast, the **Vol** model showed a more pronounced performance decline when transferred to **NVM**, showing statistical signifi-

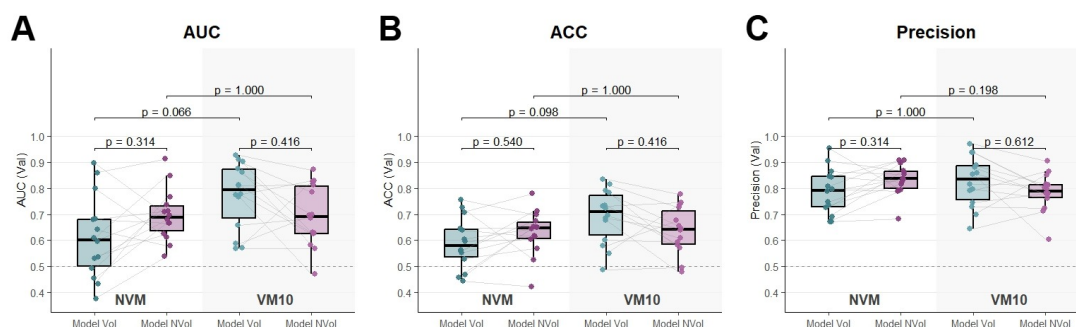


Figure 3: Validation metrics for **Vol** (cyan) and **NVol** (magenta) models evaluated on **NVM** and **VM10** datasets. P-values for between-group comparisons are reported after Bonferroni correction. Panels A, B, and C illustrate AUC, Accuracy, and Precision metrics, respectively. Individual data points represent subject-specific evaluations, while boxplots characterize the distribution across each metric.

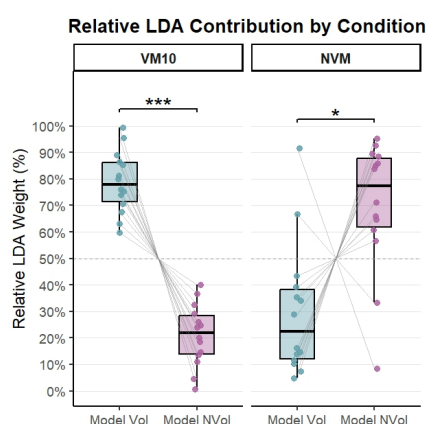


Figure 4: Relative LDA contribution by condition. The contribution of each model, **Vol** (dark cyan) and **NVol** (dark magenta), is shown for the two recordings, **VM10** and **NVM**. Triple asterisk (***) denote $p < 0.001$ and single (*) $p < 0.05$.

cance for AUC and ACC ($p = 0.0166$ and $p = 0.0245$ respectively). This indicates a greater difficulty in generalizing beyond its specific training condition.

The LDA weights shown in Figure 4 corroborate that each model exerts a greater influence when classifying its respective data, suggesting that the underlying dynamics of the two experimental conditions are distinct. In the **VM10** data, the LDA classifier relied predominantly on the **Vol** model, whereas in the **NVM** data, the contributions from both models were more balanced.

This asymmetry supports the hypothesis that a model trained on the dynamical patterns of volitional motor execution differs fundamentally from one trained on non-volitional movement, with the primary distinction residing in the involvement of efferent pathways. While both the **NVM** and **VM10** conditions involve afferent communication that elicits motor cortex responses, these results demonstrate that the presence of efferent communication significantly impacts classifier performance. The overall lower performance observed in the **NVM** condition relative to **VM10** may be attributed to the absence of these volitional motor commands and the associated sensorimotor feedback loop. Without a gen-

erated volitional drive, the stable cortical attractor that both models attempt to track may be disrupted; consequently, any resting-state epoch in which the subject incidentally reaches a dynamically stable cortical state could confound the model output, potentially destabilizing classification metrics independently of true motor intent. Given that efferent pathways are intrinsically linked to volitional control, these findings suggest that BCI paradigms must incorporate explicit feedback to effectively engage volitional mechanisms and establish robust dynamical regimes.

The LDA fusion model carries an inherent limitation in this context: by operating on the scalar outputs of each decoder, it cannot fully capture the nonlinear structure of the underlying feature distributions. Nevertheless, it provides a valuable interpretability tool, allowing the relative contribution of each decoder to be quantified directly from the discriminant coefficients.

Taken together, these findings indicate that the **Vol** model is better suited to capturing the dynamics of active motor execution, while the **NVol** model generalizes more robustly across conditions. Consequently, BCI paradigms calibrated on overt limb movement may exhibit performance degradation when transferred to scenarios lacking efferent pathways. The reduced performance observed in both models under the **NVM** condition supports the hypothesis that passive paradigms involve a disruption of the stable cortical attractor associated with overt movement. This effect may be further exacerbated by the absence of closed-loop sensorimotor feedback, a limitation that could be observed in motor imagery-based BCI contexts.

Additionally, the classification results demonstrate that the proposed method utilizing RQA-like features constitutes a reliable framework for the development of BCI systems. Its performance on unseen data, acquired several days after the initial calibration phase, aligns with current performance standards and suggests a low incidence of BCI “illiteracy” among participants. Mean AUC values were 0.772 ± 0.128 for the **Vol-VM10** pair and 0.697 ± 0.100 for the **NVol-NVM** pair. Figures 2a and 2b show representative outputs for each model, demon-

strating reliable control within a hypothetical BCI environment.

The relatively small sample size is taken as a limitation on this study, which may constrain statistical power and limit the generalization of the findings, particularly given the inherent inter-subject variability in cortical dynamics that could not be fully characterized.

CONCLUSION

The nonlinear features of the iBPTs, specifically the distance matrix derived from the phase-space embedding, are capable of encoding condition-specific dynamical structures sufficient to construct a functional BCI. Both the **Vol** and **NVol** models successfully captured the underlying iBPTs dynamics, yielding classification metrics consistent with current BCI standards. Nevertheless, it is important to emphasize that, given the inherently distinct dynamical regimes of volitional and non-volitional movement, each model was calibrated on fundamentally different dynamical parameters, and this divergence has a measurable effect on cross-condition prediction performance.

The **Vol** and **NVol** models captured distinct cortical dynamics, with the **Vol** model achieving optimal performance during volitional execution and the **NVol** model generalizing more consistently across both experimental conditions. This suggests that BCI calibration based on overt motor execution results in performance degradation when transferred to other states, whereas training on covert motor movement yields more consistent, albeit lower, results. Linear Discriminant Analysis (LDA) further revealed that the relative contribution of each decoder is condition-dependent, confirming that volitional and non-volitional motor states elicit separable dynamical regimes within the EEG signal.

The overall performance degradation observed under non-volitional conditions suggests that the absence of efferent motor drive and closed-loop sensorimotor feedback disrupts the stable cortical attractor that underlies reliable decoding. This has implications on BCI training paradigms, as it highlights the importance of model calibration and questions whether the decoders trained on motor execution transfer seamlessly to assisted or even imaginary-based conditions as the intrinsic dynamics of these signal changes.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Research Office and the Department of Engineering Studies for Innovation at Universidad Iberoamericana Ciudad de México for the financial support provided to attend this conference.

REFERENCES

[1] Pfurtscheller G. Central beta rhythm during sensorimotor activities in man. *Electroencephalogr Clin Neurophysiology*. 1981.

- [2] Pfurtscheller G, Brunner C, Schlögl A, Lopes da Silva F. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*. 2006;31(1):153–159.
- [3] Szurhaj W et al. Basic mechanisms of central rhythms reactivity to preparation and execution of a voluntary movement: A stereoelectroencephalographic study. *Clinical Neurophysiology*. 2003;114(1):107–119.
- [4] McFarland DJ, Miner LA, Vaughan TM, Wolpaw JR. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topography*. 2000;12(3):177–186.
- [5] Muthuraman. Source analysis of beta-synchronisation and cortico-muscular coherence after movement termination based on high resolution electroencephalography. *PLOS ONE*. 2012.
- [6] Stolk A et al. Electrographic dissociation of alpha and beta rhythmic activity in the human sensorimotor system. *eLife*. 2019;8:e48065.
- [7] Ulloa JL. The control of movements via motor gamma oscillations. *Frontiers in Human Neuroscience*. 2022;15.
- [8] Jayalakshmi Srinivasan K, Periasamy P, Gunasekaran S. Motor imagery and motor execution: A narrative review of electroencephalographic (EEG) signatures, methodological consistency, and translational applications. *Cureus*. 2025;17(9):e93011.
- [9] Yu Y et al. Enhanced neuroplasticity and gait recovery in stroke patients: A comparative analysis of active and passive robotic training modes. *BMC Neurology*. 2025;25(1):239.
- [10] Edelman BJ et al. Non-invasive brain-computer interfaces: State of the art and trends. *IEEE Reviews in Biomedical Engineering*. 2024;1–25.
- [11] Marwan N, Meinke A. Extended recurrence plot analysis and its application to ERP data. *International Journal of Bifurcation and Chaos*. 2004;14(02):761–771.
- [12] Pitsik E et al. Motor execution reduces EEG signals complexity: Recurrence quantification analysis study. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 2020;30(2).
- [13] Morales-Magallón F, Bojorges-Valdez E. Intended and non-volitional knee joint movements elicit distinct functional brain networks. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2025, 1–6.
- [14] Gramfort A et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013;7(267):1–13.
- [15] Delorme A, Sejnowski T, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*. 2007;34(4):1443–1449.
- [16] Takens F. Detecting strange attractors in turbulence. In: *Dynamical Systems and Turbulence*, Warwick 1980. Springer Berlin Heidelberg: Berlin, Heidelberg, 1981, 366–381.

Goal-Directed Quasi-Movements Based on Natural Hand Actions: EEG Correlates and Relevance for Rehabilitation BCI

E. P. Svirin¹, D. A. Berdyshev^{1,2}, S. L. Shishkin¹

¹ MEG-Center, Moscow State University of Psychology and Education

² Department of Research Automation, M.V. Lomonosov Moscow State University

E-mail: jogikint@gmail.com

ABSTRACT: Quasi-movements (QMs) are movements minimized to the point that they produce no visible movement or detectable muscle activation. Unlike kinesthetic motor imagery (MI), they are closer to actual movement in the sense of intention and may therefore provide a more natural basis for sensorimotor BCI control. However, previous QM research has been largely limited to thumb abduction. Here we investigated, for the first time, QMs based on more natural goal-directed hand actions, pressing and pointing. Healthy participants performed kinesthetic MI, non-goal-directed QMs, goal-directed QMs, and real movements while EEG and a custom force-sensor signal were recorded. Both QM conditions elicited stronger contralateral μ -rhythm desynchronization than MI, whereas ipsilateral μ -ERD did not differ reliably between conditions. Goal-directed instruction did not produce an additional detectable effect, and residual movement did not account for the EEG differences. These preliminary findings suggest that natural-action QMs are feasible and may represent a promising alternative to MI for non-invasive BCI applications.

INTRODUCTION

Kinesthetic motor imagery (MI) is one of the most widely used control strategies in non-invasive brain-computer interfaces (BCIs) [1]. However, MI is an internally directed task and may compete with the processing of external information, including visual feedback, as well as with oculomotor control and visual attention [2,3]. In addition, MI requires the user to perform a relatively abstract mental task rather than to prepare an actual action. This motivates the search for alternative motor-related tasks that would preserve informative EEG correlates while being closer to natural action.

A promising candidate is quasi-movement (QM), a phenomenon described by Nikulin et al., in which movement amplitude is reduced until no visible movement and no detectable muscle activation remain, while EEG patterns similar to those of actual movement are preserved [4,5]. QMs were previously proposed as a model of movement attempts, which had previously been discussed as a more natural alternative to MI for BCI, particularly in clinical and rehabilitation applications [6,7]. MI-based training is difficult to teach reliably within the limited time available in clinical settings [7]. Overt attempts are more natural for patients but carry their own risks in paresis and plegia: repeated effortful

contractions can aggravate spasticity and recruit pathological synergies. QMs, as movement attempts intentionally minimized to the point of no overt output, may avoid both problems: they preserve the natural feel of an action attempt while removing the peripheral output that drives the side effects of overt attempts. Recent work also indicates that residual muscle activation, when present, does not fully account for QM-related EEG effects [8], and that QMs are experienced as more similar to actual movement than MI, which may support the sense of control during BCI use [9]. In healthy participants, QMs additionally allow direct verification of the absence of overt output through force or EMG measurements, which is not feasible in patient studies of attempted movement. At present, however, QM research remains limited largely to a single movement, thumb abduction [4,8]. This leaves open the question of whether QMs can be extended to more natural, goal-directed actions that are closer to everyday control. This question is important because goal-directed movements have specific neural correlates and may be especially relevant for BCI applications [10,11].

Here we examined whether new goal-directed hand actions can be implemented as QMs and whether they retain the sensorimotor-rhythm modulation that makes QMs promising for BCI applications. Specifically, we developed, for the first time, task instructions for QM performance based on pressing and pointing actions, and compared the resulting QMs with kinesthetic MI of the same actions. We also assessed the effects of instruction and residual movement. We expected that these new QMs would be feasible in healthy participants and would produce sensorimotor-rhythm changes at least as strong as, and potentially stronger than, those observed during kinesthetic MI.

MATERIALS AND METHODS

Participants: 11 healthy volunteers (5 females, aged 21-32 years, median age 28 years) participated in the study after providing their written informed consent. Two of them had prior experience with MI and QM.

Signal acquisition: EEG (64 electrodes) was recorded at 1000 Hz using an NVX136 DC EEG amplifier (Medical Computer Systems, Moscow, Russia). The ground electrode was placed at Fpz. Electrode impedance was kept below 20 k Ω .

For mechanical registration of minimal residual movement, we developed a custom sensor based on a

DKFFF-6AA-120 strain gauge mounted on a 3D-printed compliant mechanical structure. Resistance changes were measured using a Wheatstone bridge circuit, and the midpoint voltage difference was amplified by a differential amplifier based on an LM358 operational amplifier. For comfortable hand placement, the sensor was installed in a housing that allowed the participant to rest the hand palm-down in a natural relaxed posture while the distal phalanx of the target finger was positioned on a circular plate (20 mm diameter) flush with the surrounding surface. The sensor was powered by an NVX136 amplifier AUX port, and the data was acquired using the analog input on the same port. Because the strain gauge registered deformation of the elastic sensor structure rather than force along a single separately calibrated axis, the sensor could detect very small changes in finger-applied force irrespective of direction. Surface EMG was not considered suitable for this study because pressing and pointing depend on multiple, partly deep muscle groups, so the minimal residual activation expected during quasi-movements would be difficult to detect and interpret reliably from surface recordings.

Software: Signal acquisition and real-time processing were performed using the Resonance platform [12]. The stimulus presentation sequence was delivered using a custom PyGame-based presentation script. Primary offline data processing was conducted in MNE-Python. Statistical analysis was performed in R using the lmerTest package, and post hoc comparisons were carried out with emmeans.

Experiment design: Participants sat in a chair in front of a computer screen. The right forearm rested comfortably on a support, and the right hand was placed palm-down on the force sensor so that the index finger contacted its sensitive plate. Stimuli were presented in blocks, each consisting of two motor-task sequences and one visual control sequence, with inter-sequence intervals varying randomly from 2 to 4 s.

In each motor sequence, an image of a right hand performing the corresponding action was presented for 3.2 s. One second after image onset, three brief auditory clicks were delivered with an inter-click interval of 600 ms. In response to each click, the participant performed one repetition of the task required in the current condition. In the visual control sequence, a figure with complex geometric elements was shown for 4 s, and the participant silently counted elements of a selected type at a comfortable pace, as described elsewhere [8]. Each condition comprised 20 blocks, yielding 40 motor sequences and 20 visual sequences; before each condition, participants completed 5–10 training blocks. For each of the two actions, four task types were recorded: overt movement (**OM**), kinesthetic imagery movements (**IM**), non-goal-directed quasi-movement (**nQM**), and goal-directed quasi-movement (**gQM**). OM condition was always performed first, IM and nQM were presented in random order, and gQM was always performed last for the given action. This was done because goal-directedness was introduced by instruction,

and earlier presentation of gQM could have induced premature association of the movement with its goal-directed meaning. Accordingly, before the gQM condition the experimenter avoided action-related keywords such as “press”, “button”, or “point”, describing the movement only in joint-motion terms. Immediately before gQM, the same movement was reintroduced in explicitly goal-directed terms.

Training: QM training was adapted from the procedure of Nikulin et al. and our previous work [4,8]. Participants were first shown the processed force-sensor signal on the screen and could verify that it changed with index-finger movement. To facilitate user comprehension, the Sum of Absolute Differences (SAD) between the current and lagged sensor force signals was displayed. This representation allows for the clear identification of force change regions and reduces the noise visually apparent in the graph [13]. Participants were then instructed to perform the target movement in a self-paced manner while gradually reducing its amplitude and the sensor output. After an individually determined threshold was reached, visual feedback was removed and the participant continued producing the movement in response to isolated auditory cues while further minimizing its amplitude if it remained distinguishable from sensor noise. Once stable QM performance was achieved, the experimenter asked control questions to verify that the participant had not switched to motor imagery or to another unintended movement. Finally, participants practiced responding to the triplets of auditory cues used in the main experiment. IM training was conducted immediately before each IM condition. Participants first performed the corresponding movement several times with full amplitude while attending to the sensations arising in the hand. They were then asked to describe these sensations, after which non-visual sensations such as tension, weight, stretching, or warmth were selected and used as the basis for kinesthetic imagery. Participants were instructed to imagine these sensations as localized in the hand, as if they were performing the actual movement. Training continued with auditory triplets until the participant reported being able to perform the imagery task confidently at the required pace.

Data processing: Offline processing was performed in MNE-Python. Bad EEG channels were detected automatically with PyPREP [14], using its noisy-channel detection procedures, and were then interpolated; the data were subsequently transformed into current source density (CSD). Residual movement was quantified from the force-sensor signal by robust peak amplitude (RPA), defined as the difference between the 95th and 5th percentiles within each epoch. We used this measure because it provides an estimate of overall movement-related force variation while being less affected than the conventional peak-to-peak range by isolated artifacts and extreme values. It was therefore well suited for detecting small and irregular residual movements during quasi-movements.

EEG analysis focused on the sensorimotor ROI

comprising FC5, FC3, C3, C1, CP3, CP1 and FC6, FC4, C4, C2, CP4, CP2. Single-trial Morlet time-frequency power was computed for each ROI channel and manually baseline-normalized in dB as $10\log_{10}(\text{power} / \text{PRE})$, where PRE was the mean baseline power in the selected pre-stimulus interval. μ -band ERD (8–13 Hz) was then derived from the band-limited ROI power time course using the selected post-stimulus interval relative to the first auditory cue. For each participant and condition, the electrode showing the maximal effect within the ROI was selected for analysis.

Statistical analysis: μ -band ERD was analysed with linear mixed-effects models fitted separately for contralateral μ ERD and ipsilateral μ ERD. The model included movement type, condition, epoch number, preceding task, prestimulus desynchronization level, and RPA, with subject as a random intercept. The interaction between RPA and task type was included to estimate the contribution of residual movement separately for different task types. Statistical analysis was performed in R using *lmerTest*, with post hoc comparisons conducted using *emmeans*.

RESULTS

Analysis of the force-sensor signal – Empirical cumulative distribution functions of RPA for different movements and conditions are shown in Fig. 1.

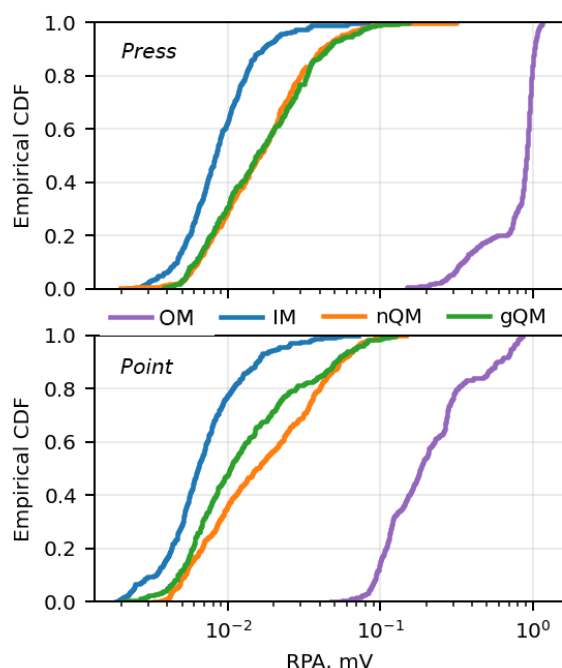


Figure 1. Cumulative distribution functions (CDFs) for the trial RPA values computed for different movements and conditions.

Table 1 shows the percentages of nQM and gQM epochs containing elevated residual movement, defined for each participant and movement separately as epochs whose RPA exceeded the 97.5th percentile of the RPA distribution observed in the corresponding IM condition. Participants varied highly in the percentage of trials

without elevated residual movement, and only four participants were able to refrain from producing elevated RPA in all the conditions. Still, no participants produced RPA stronger than minimal RPA observed during OM in the QM conditions.

Table 1. Percentages of epochs with elevated residual movement in QM conditions.

Participant	Press		Point	
	nQM	gQM	nQM	gQM
1	5	10	90	52.5
2	0	0	22.5	0
3	17.5	15	15	35
4	10	67.5	0	2.5
5	85	65	70	97.5
6	17.5	—	30	—
7	52.5	27.5	97.5	27.5
8	90	80	40	30
9	77.5	100	67.5	92.5
10	12.5	45	10	2.5
Mean	36.75	45.6	44.25	37.8

Repeated measures ANOVA with condition as a within-subject factor showed that movement magnitude, quantified by RPA, differed significantly across conditions for both *press* and *point* ($p < 0.001$ in both cases). For *press*, RPA was markedly higher in OM than in IM, nQM, and gQM (all $p < 0.001$). Among the non-overt conditions, nQM showed slightly higher RPA than IM ($p = 0.008$), whereas the difference between gQM and IM did not reach significance ($p = 0.064$). nQM and gQM did not differ from each other ($p = 0.941$).

For *point*, OM again showed higher RPA than IM, nQM, and gQM ($p = 0.028$, $p = 0.028$, and $p = 0.027$, respectively). By contrast, no significant differences were found among the non-overt conditions after correction for multiple comparisons: the contrast between IM and nQM was marginal ($p = 0.050$), whereas IM and gQM ($p = 0.163$) and nQM and gQM ($p = 0.599$) did not differ.

Overall, these results indicate that overt movement was associated with the largest mechanical signal, as expected, whereas the residual movement observed during quasi-movements was small and generally comparable to that observed during imagery, especially for the *point* movement.

Sensorimotor-rhythm modulation in the new tasks – The contralateral μ -band ERDs are shown in Fig. 2, and the corresponding grand average timecourses are shown in Fig 3. As far as the main goal of this preliminary study was to compare new QMs with corresponding IMs, the OM condition was excluded from this analysis. The analysis with linear mixed-effects models showed significant effects of movement type, task type, previous task, and prestimulus amplitude ($p = 0.001$, $p = 0.024$, $p < 0.001$, and $p < 0.001$, respectively). By contrast, neither RPA nor its interaction with task type reached significance ($p = 0.967$ and $p = 0.820$, respectively). Overall, contralateral μ -ERD was stronger for *press* than for *point* movements. With respect to task type, both quasi-movement conditions were associated with

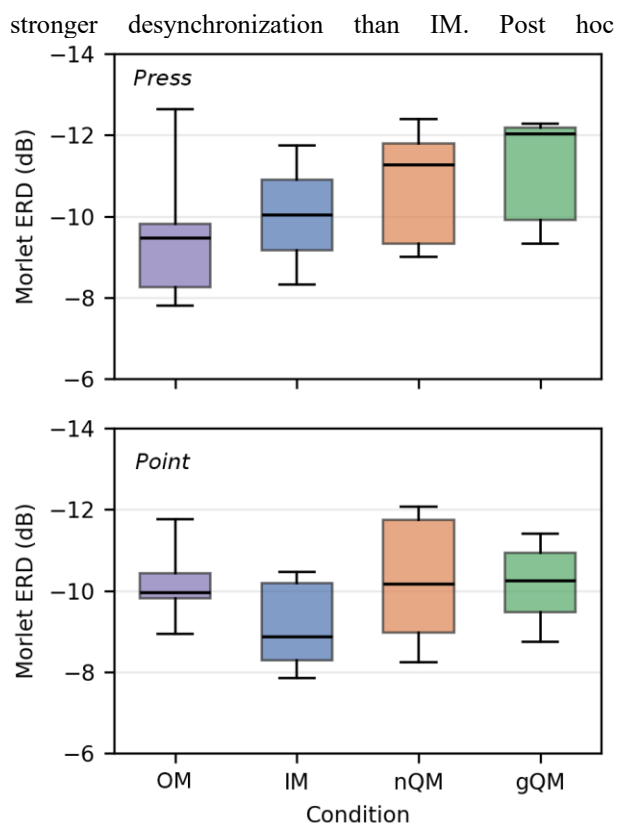


Figure 2. Desynchronization of contralateral sensorimotor rhythms in different conditions.

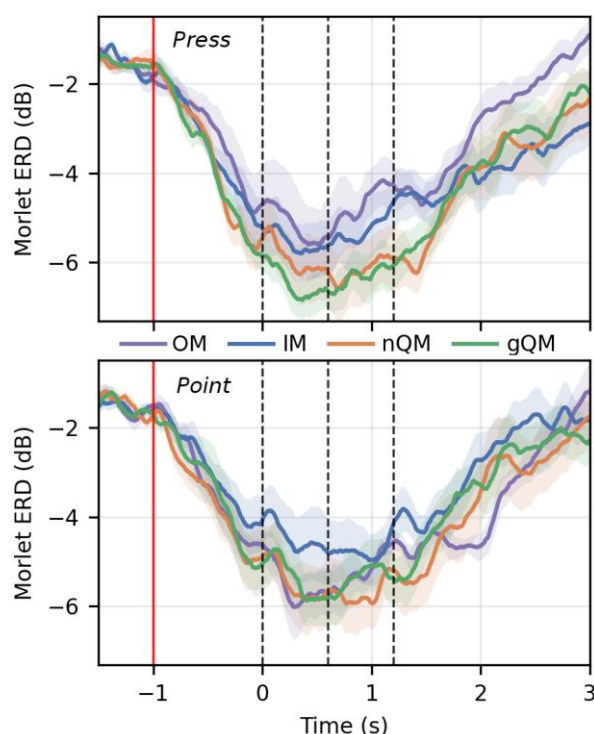


Figure 3. Group grand average time courses for ERD in sensorimotor rhythm band (8-14 Hz). Vertical red line corresponds to the moment of visual cue appearing on the screen. Vertical dashed lines correspond to the sound stimuli onset.

comparisons confirmed that μ -ERD was significantly stronger in nQM than in IM ($p=0.027$) and in gQM than in IM ($p=0.042$), whereas nQM and gQM did not differ ($p=0.792$).

Thus, both QM conditions elicited stronger contralateral μ -rhythm desynchronization than IM, while the goal-directed instruction did not produce an additional effect. The absence of significant RPA-related effects further indicates that residual movement was unlikely to account for these differences.

For ipsilateral μ -ERD, the analysis revealed significant effects of previous task and prestimulus amplitude ($p<0.001$ and $p=0.008$, respectively). By contrast, the effects of movement type and RPA were not significant ($p=0.236$ and $p=0.798$, respectively), and the main effect of task type did not reach significance ($p=0.086$). The RPA $\times$ task type interaction was also non-significant ($p=0.788$).

None of the pairwise comparisons between task types reached significance ($p=0.203$ for IM vs nQM, $p=0.955$ for IM vs gQM, and $p=0.135$ for nQM vs gQM). Thus, unlike contralateral μ -ERD, ipsilateral μ -ERD did not show reliable differences between imagery and quasi-movement conditions.

The absence of significant RPA-related effects further suggests that residual movement did not make a detectable contribution to ipsilateral μ -ERD.

DISCUSSION

These results are preliminary, given the small sample. Nevertheless, they suggest that quasi-movements can be extended beyond the “classical” thumb-abduction paradigm [4] to more natural hand actions such as pressing and pointing. The main result was that both nQM and gQM were associated with stronger contralateral μ -rhythm desynchronization than IM, whereas ipsilateral μ -ERD did not differ reliably between conditions. In addition, the contribution of residual movement, quantified by RPA, was not significant in either hemisphere. Overall, this pattern indicates that the advantage of the new quasi-movements over imagery was expressed primarily in contralateral sensorimotor activation and was unlikely to be explained simply by peripheral motor leakage.

In this respect, the present findings are broadly consistent with the original study of quasi-movements by Nikulin et al. [4], in which α -band ERD was strongest for overt movement, weaker for quasi-movement, and weakest for motor imagery, with the QM-IM difference being especially clear in the contralateral hemisphere. They are also in line with the later analysis of Vasilyev et al. [8], which showed that stronger contralateral ERD in QM than in imagery could not be explained by residual muscle activation alone. Thus, the present data extend the earlier QM literature by indicating that this advantage over imagery may also be obtained for more natural hand actions than thumb abduction.

At the same time, the absence of an additional effect of goal-directed instruction should be interpreted

cautiously. Existing evidence suggests that goal-directedness may be expressed less as larger mean ERD over central sensorimotor sites than as earlier onset, broader spatial distribution, and stronger parietal involvement, particularly in the upper alpha range [10,11,15–17]. However, the available direct comparisons are few, based on small samples, and partly confounded by movement complexity. Therefore, the lack of a clear gQM vs. nQM effect in the present study does not necessarily argue against goal-directedness-related neural differences. Rather, it may reflect the fact that the present analysis focused on mean amplitude within a central sensorimotor ROI, whereas the most specific effects of goal-directedness may lie in parietal recruitment, upper- μ dynamics, and temporal characteristics of ERD.

A further reason for caution is that the two actions used here, pressing and pointing, are themselves highly familiar and naturally associated with external objects and outcomes. Thus, even the nominally non-goal-directed variants may have retained part of their habitual action meaning. In addition, the gQM condition had to be presented last for each movement in order to avoid contaminating the neutral conditions, which was methodologically justified but may also have reduced the possibility of isolating a pure goal-directedness effect. Under these circumstances, the present data are better interpreted as showing that goal-directed instruction did not produce an additional detectable effect in the current design, rather than as evidence against the relevance of goal-directedness per se.

Although participants varied in the proportion of QM epochs containing elevated residual movement, the covert conditions were much closer to one another than to OM, especially for pointing. Most importantly, neither the main effect of RPA nor its interaction with task type was significant in the mixed-model analysis of μ -ERD. This argues against the simplest explanation that the stronger desynchronization in QMs merely reflected residual movement. Instead, the present findings support the view that QMs represent a distinct motor-related state that is closer to overt movement than imagery at the cortical level while remaining behaviorally non-overt. This interpretation is also compatible with the conclusion of Vasilyev et al. [8] that occasional peripheral activation is insufficient to account for the stronger cortical effects observed in QMs. The sensor was not separately calibrated for directional sensitivity, but the absence of a strong press-point dissociation in the covert conditions argues against this as a main source of the observed effects (Fig. 1).

A particularly interesting aspect of the present data is the pattern observed for the *press* movement. Although OM was not included in the main mixed-model comparison focused on IM and QMs, inspection of the data suggested that for at least some participants contralateral ERD during overt pressing tended to be smaller than in the QM conditions. At first sight, this appears inconsistent with the classical pattern reported by Nikulin et al. [4], where overt movement produced the strongest ERD. However,

the broader literature offers a plausible interpretation. Highly familiar or automated movements may be associated with reduced, rather than increased, sensorimotor ERD, especially in the α/μ range, in line with a neural-efficiency account [18–21]. Lower movement-related α desynchronization has been reported in highly trained participants, and practice may likewise reduce desynchronization or increase sensorimotor power [18–21]. On this view, overt pressing in the present experiment may have functioned as a highly familiar and automatized action, whereas pressing in the QM mode required a less habitual and more demanding control regime, since participants had to generate an effective motor command while simultaneously suppressing overt output. If so, the tendency toward smaller ERD in OM *press* would not necessarily contradict the QM framework but might reflect an interaction between quasi-movement-related activation and movement familiarity.

This interpretation should, however, remain cautious. The available literature on familiarity and sensorimotor ERD is heterogeneous with respect to task type, expertise, and movement phase [18–21]. Likewise, the literature on goal-directedness is limited by small samples and by incomplete separation of goal-related and complexity-related factors [15,16]. Therefore, the present data do not justify a general conclusion that familiar overt actions produce weaker ERD than quasi-movements. Rather, they suggest that the unexpected *press* pattern is compatible with findings on automatization and neural efficiency and deserves direct testing in a larger study.

A methodological limitation should also be noted. Band-power ERD does not separate cortical activation from active inhibition of motor output, which is particularly relevant for QMs, since participants generate a motor command while suppressing overt execution. The stronger contralateral μ -ERD for QMs relative to IM is therefore compatible with greater activation, stronger inhibitory control, or both. Direct characterization would require methods such as TMS-derived motor-evoked potentials, paired-pulse paradigms, or corticomuscular coherence, which have not yet been applied to QMs.

From the BCI perspective, the present findings are encouraging. The stronger contralateral μ -ERD observed in nQM and gQM relative to IM is consistent with the general idea that motor-related states closer to actual movement may provide more informative control signals than classical imagery [1,22]. The extension to pressing and pointing is relevant to rehabilitation in particular, since these actions are closer to functional grasp and reach targets than the thumb abduction used in previous QM studies, and goal-directed action representations have been proposed as a basis for more naturalistic BCI control [10,11]. Real-time QM-BCI control of hand exoskeleton has recently been demonstrated in both healthy participants and chronic stroke patients (Yashin et al., this conference), supporting the feasibility of the broader approach; the present results suggest that the task repertoire available for such applications can be

expanded beyond a single minimal movement. Future work should therefore include replication in a larger sample, more explicit manipulation of action goal-directedness, and testing of these new QMs in online BCI paradigms.

CONCLUSION

Our results suggest that pressing- and pointing-based quasi-movements are feasible and may provide more informative sensorimotor EEG correlates than imagery, as reflected in stronger contralateral μ -rhythm desynchronization. These preliminary findings support further investigation of natural-action quasi-movements as a promising alternative to motor imagery for BCI applications.

ACKNOWLEDGMENTS

This study was supported by the Russian Science Foundation under grant no. 24-75-00105.

The authors thank our colleague Anatoly Vasilyev for providing useful advices and feedback

REFERENCES

- [1] Pfurtscheller G, Neuper C. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*. 2001;89(7):1123-1134.
- [2] Korda Z, Walcher S, Korner C, Benedek M. Effects of internally directed cognition on smooth pursuit eye movements: A systematic examination of perceptual decoupling. *Atten Percept Psychophys*. 2023;85(4):1159-1178.
- [3] Walcher S, Korda Z, Korner C, Benedek M. The effects of type and workload of internal tasks on voluntary saccades in a target-distractor saccade task. *PLoS One*. 2023;18(8):e0290322.
- [4] Nikulin VV, Hohlefeld FU, Jacobs AM, Curio G. Quasi-movements: a novel motor-cognitive phenomenon. *Neuropsychologia*. 2008;46(2):727-742.
- [5] Hohlefeld FU, Nikulin VV, Curio G. Covert movements trigger repetition suppression of electroencephalography in sensorimotor cortex. *Neuroreport*. 2011;22(3):141-145.
- [6] Blokland Y, Spyrou L, Lerou J, Mourisse J, Jan Scheffer G, van Geffen GJ, et al.. Detection of attempted movement from the EEG during neuromuscular block: proof of principle study in awake volunteers. *Sci Rep*. 2015;5:12815.
- [7] Mansour S, Ang KK, Nair KPS, Phua KS, Arvaneh M. Efficacy of Brain-Computer Interface and the Impact of Its Design Characteristics on Poststroke Upper-limb Rehabilitation: A Systematic Review and Meta-analysis of Randomized Controlled Trials. *Clin EEG Neurosci*. 2022;53(1):79-90.
- [8] Vasilyev AN, Yashin AS, Shishkin SL. Quasi-Movements and “Quasi-Quasi-Movements”: Does Residual Muscle Activation Matter? *Life (Basel)*. 2023;13(2).
- [9] Yashin AS, Shishkin SL, Vasilyev AN. Is there a continuum of agentive awareness across physical and mental actions? The case of quasi-movements. *Conscious Cogn*. 2023;112:103531. doi:10.1016/j.concog.2023.103531
- [10] Pereira J, Ofner P, Muller-Putz GR. Goal-directed or aimless? EEG differences during the preparation of a reach-and-touch task. *Annu Int Conf IEEE Eng Med Biol Soc*. 2015;2015:1488-1491.
- [11] Pereira J, Ofner P, Schwarz A, Sburlea AI, Muller-Putz GR. EEG neural correlates of goal-directed movement intention. *Neuroimage*. 2017;149:129-140.
- [12] Nuzhdin, Y. Resonance - a BCI framework for working with multiple data sources, in *Proc. of the 8th Graz Brain Computer Interface Conference 2019*. Verlag der Technischen Universität Graz. Bridging Science and Application.
- [13] Wiktorski T, Królak A. Extended approach to sum of absolute differences method for improved identification of periods in biomedical time series. *MethodsX*. 2020;7:101094.
- [14] Appelhoff S, Hurst AJ, Lawrence A, Li A, Mantilla Ramos YJ, O'Reilly C, et al. PyPREP: A Python implementation of the preprocessing pipeline (PREP) for EEG data. Published online February 26, 2026.
- [15] Fumuro T, Matsuhashi M, Miyazaki T, Inouchi M, Hitomi T, Matsumoto R, et al. Alpha-band desynchronization in human parietal area during reach planning. *Clin Neurophysiol*. 2015;126(4):756-762.
- [16] Labyt E, Szurhaj W, Bourriez JL, Cassim F, Defebvre L, Destee A, et al. Changes in oscillatory cortical activity related to a visuomotor task in young and elderly healthy subjects. *Clin Neurophysiol*. 2003;114(6):1153-1166.
- [17] Pfurtscheller G, Neuper C, Krausz G. Functional dissociation of lower and upper frequency mu rhythms in relation to voluntary limb movement. *Clin Neurophysiol*. 2000;111(10):1873-1879.
- [18] Del Percio C, Babiloni C, Bertollo M, Marzano N, Iacoboni M, Infarinato F, et al. Visuo-attentional and sensorimotor alpha rhythms are related to visuo-motor performance in athletes. *Hum Brain Mapp*. 2009;30(11):3527-3540.
- [19] Del Percio C, Infarinato F, Iacoboni M, Marzano N, Soricelli A, Aschieri P, et al. Movement-related desynchronization of alpha rhythms is lower in athletes than non-athletes: a high-resolution EEG study. *Clin Neurophysiol*. 2010;121(4):482-491.
- [20] Kerick SE, Douglass LW, Hatfield BD. Cerebral cortical adaptations associated with visuomotor practice. *Med Sci Sports Exerc*. 2004;36(1):118-129.
- [21] Wright DJ, Holmes P, Di Russo F, Loporto M, Smith D. Reduced motor cortex activity during movement preparation following a period of motor skill practice. *PLoS One*. 2012;7(12):e51886.
- [22] Yashin AS, Vasilyev AN, Shevtsova YG, Shishkin SL. Can Quasi-Movements be Used as a Model of the BCI Based on Attempted Movements? In: 2024:2028-2033.

HD-EMG-Based BCI for Control and Assessment in Post-Stroke Hand Paresis: A Proof of Concept

Gian Maria Velardi^{1*}, Thien N. Le^{1*}, Andrea Cimolato¹, Bernhard Fasching², Wolfgang Serles²,
Stanisa Raspopovic^{1†}

¹Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria

²Department of Neurology, Medical University of Vienna, Vienna, Austria

* These authors contributed equally to this work

† Corresponding author. E-mail: stanisa.raspopovic@meduniwien.ac.at

ABSTRACT: Brain–computer interfaces (BCIs) offer a promising avenue for restoring motor function after stroke, but conventional BCIs are often limited by lengthy calibration or invasive implantation. High-density electromyography (HD-EMG) enables non-invasive decomposition of surface muscle signals into individual motor unit (MU) spike trains, providing a direct readout of neural drive to muscle, even when corticospinal transmission is impaired (e.g., after stroke). However, the blind source separation algorithms used for decomposition typically require approximately 15 seconds of sustained contraction, a demand that many stroke survivors cannot meet due to post-stroke fatigue and reduced motor endurance. Here we present a new HD-EMG processing pipeline that extracts movement intention through motor unit decomposition using repeated brief contractions rather than a single sustained effort. We evaluate this calibration strategy in both simulated signals (with ground-truth MU activity) and stroke-participant recordings. Our repeated-contraction approach maintains the same number validated motor units compared to sustained contraction (~6.75–6.95 MUs), indicating that our approach preserves decomposition sensitivity of state of art approaches. Furthermore, using two selectively recruited motor units, we demonstrate real-time classification capability of two distinct hand movements with high accuracy. Importantly, we identify a satellite potential in the motor unit action potential waveforms. This is a biomarker of corticospinal reorganization not previously observed with non-invasive recordings of stroke patients. These findings establish a clinically feasible calibration paradigm for HD-EMG-based BCIs in neurological populations and suggest that motor unit analysis may simultaneously serve both real-time control and longitudinal monitoring of recovery.

INTRODUCTION

Stroke is a leading cause of long-term disability, with approximately 12 million new cases annually worldwide [1]. Upper limb motor impairment is among the most common sequelae, affecting up to 80% of patients in the acute phase and persisting in close to 40% at the chronic stage [2]. Recovery of hand dexterity, grasping, and reaching is critical for independence in daily activities

such as feeding, dressing, and object manipulation, and is a significant predictor of health-related quality of life. Yet, only approximately half of survivors regain upper limb functional use within six months [3].

Brain–computer interfaces (BCIs) have been investigated as adjuncts to upper limb rehabilitation after stroke [4][5]. Muscles act as biological amplifiers of motor commands originating in the cortex, therefore Electromyographic (EMG) signals provide an indirect but accessible window into the neural control of movement, which can be used to interface with external computational systems [6]. However, when the corticospinal pathway is damaged, as in stroke, the descending drive to muscles is weakened, and conventional surface EMG may fail to capture meaningful activity [7].

High-density EMG (HD-EMG) addresses this limitation by enabling decomposition of the interference signal into individual motor unit action potentials and their discharge trains, providing direct access to the neural drive to the muscle. Critically, motor unit decomposition remains feasible even when corticospinal transmission is impaired. Studies in individuals with spinal cord injury [8][9] and Stroke [10] have shown that HD-EMG decomposition can extract motor unit activity despite reduced descending input, even from muscles that do not produce visible force. Among available decomposition methods, blind source separation (BSS) algorithms are the most widely validated [11]. These algorithms rely on the statistical properties of the signal accumulated over a sufficient observation window, typically requiring approximately 15 seconds of sustained contraction for reliable decomposition [11][13]. However, such trivial motor action for a healthy participant, could be problematic in stroke participants, as fatigue affects approximately 50% of the survivors [13]. To address this constraint, we propose a decomposition strategy that acquires the necessary signal information through multiple repetitions, accommodating the limited endurance of neurological patient. This modification may improve the clinical feasibility of HD-EMG-based systems for motor assessment and neurorehabilitation.

MATERIALS AND METHODS

Decomposition pipeline overview

We developed a motor unit decomposition pipeline designed to operate on short, repeated contractions rather than prolonged sustained efforts—a key practical requirement when recording from patients with severe motor impairment who cannot sustain isometric contractions for extended periods. The pipeline introduces two principal methodological contributions relative to the standard offline CBSS workflow [11]: (i) a segmentation and concatenation strategy that constructs a pseudo-continuous signal from repeated contraction plateaus, making the signal duration sufficient for reliable source separation without requiring a single long contraction; and (ii) recalibration of the offline separation filters for causal, peel-off-free deployment in a real-time streaming context.

The processing pipeline consists of the following

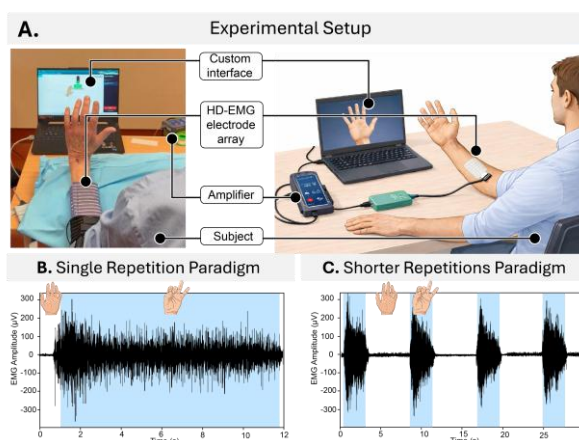


Figure 1. Experimental setup and calibration paradigm. (a) Custom recording interface providing animated visual cues for movement and rest periods. (c) Comparison of calibration conditions: a single sustained contraction (B) versus concatenated segments from multiple brief repetitions (C) regions shaded in blue correspond to the signal being decomposed.

subsequence. First, HD-EMG signals recorded during the repeated-movement condition are segmented into individual repetition windows corresponding to the plateau phase of each contraction. Second, the segmented windows are concatenated to form a single continuous signal of sufficient length for source separation. Third, CBSS [11] is applied to the concatenated signal to extract individual motor unit spike trains and action potential waveforms. Fourth, the decomposed motor units are evaluated on the basis of waveform morphology and firing-rate statistics to distinguish physiologically plausible units from decomposition artifacts [10]. Lastly, the motor unit filters derived during this calibration phase are recalibrated and applied to simulate real-time streaming conditions, using signals recorded during the same experimental session [14].

To validate the pipeline before applying it to patient data, the full workflow was first evaluated on simulated HD-EMG data with known ground-truth spike trains allowing objective quantification of decomposition accuracy. The

validated pipeline was then applied to recordings from a stroke patient to demonstrate feasibility in a clinical setting.

Signal processing

Raw HD-EMG signals were preprocessed offline as follows. The DC offset was removed from each channel, followed by band-pass filtering (4th-order Butterworth, 20–500 Hz) and notch filtering at 50, 100, and 150 Hz ($Q = 30$) to suppress power-line interference and its harmonics. Impulsive artifacts exceeding 8 standard deviations were removed by linear interpolation, and 0.5s of edge-contaminated samples was discarded at each signal boundary.

For the repeated-movement calibration condition, video footage was used to identify the sustained isometric phase of each contraction, excluding movement onset and return. These segments were extracted, concatenated across repetitions, and processed using the decomposition pipeline described below.

Motor unit decomposition

Motor unit decomposition was performed using a convolutive blind source separation (CBSS) framework [11] as implemented in MUniverse [15]. Prior to source separation, preprocessed signals were extended by a factor of $R = 16$ to capture delayed channel replicas and whitened using Zero-phase Component Analysis (ZCA) to enforce statistical decorrelation. FastICA was subsequently applied to extract the activity of the single MUs.

Spikes in this source activity were detected using a peak-finding algorithm with a 20ms refractory period, followed by k-means clustering ($k = 2$) to discriminate active peaks corresponding to spike firing from background peaks caused by noise. Motor units were retained only if they satisfied a silhouette score (SIL) > 0.85 between the two clusters, indicating good separation of the peaks from the baseline noise.

Units with sufficiently regular Inter Spike Interval (ISI) (Coefficient of Variance (CoV) < 0.4) were peeled off from the signal as they were found. Duplicate units sharing more than 50% of common discharge times within a ± 0.5 ms tolerance window were discarded; the duplicated unit with the higher SIL score was kept.

The output of the decomposition comprised, for each identified motor unit, the estimated motor unit action potential (MUAP) waveform and its corresponding binary spike train.

Validation of the extracted MUs is performed by visually inspecting the motor unit action potential (MUAP) waveforms, using the criteria mentioned in [10]. Only those motor units with clear, well-defined waveforms that are distinguishable above baseline noise, such as the in Figure 2 are considered physiologically plausible motor units and retained for subsequent analysis.

Simulated online processing

To simulate Real Time signal analysis and control, the

HD-sEMG signals recorded during the experimental session were divided in 50 ms chunks and streamed to the online decomposition algorithm via the Lab Streaming Layer (LSL) framework [16]. Each chunk was individually preprocessed in real time in the same manner as the offline preprocessing.

The offline separation filters, derived from decomposition of only the first four repetitions, were recalibrated via spike-triggered averaging on the raw HD-EMG to reintroduce temporal overlaps removed during the offline peel-off step, adapting the filters for causal, non-peel-off spike detection [17].

Because each movement type was recorded separately during calibration, a unified whitening matrix was computed by concatenating preprocessed HD-EMG segments across all movements. Spike and noise centroids, obtained offline, were then recalibrated against this unified matrix, eliminating the need for real-time peel-off or k-means clustering [17].

The incoming HD-EMG signal was processed with each chunk spatially extended and projected onto the pre-computed whitening matrix before applying the separation matrix necessary to obtain the source signal corresponding to the activities of the single motor units. The resulting source signal was normalized using offset and scale parameters estimated during the offline calibration and clipped to the unit interval [0, 1]. Spike events were identified by comparing the normalized source against the recalibrated centroids in Euclidean space; values closer to the spike centroid were classified as spikes. The instantaneous firing rate of each motor unit was estimated at every sample using a sliding rectangular window of 1 s duration. The pipeline updated every 50ms as the data chunks arrived.

We selected 2 motor units with the most distinctive firing rate for the tripod pinch and digit 4+5 flexion to use as control signal. Thresholding of the firing rate were used to determine the controller's output. We computed a confusion matrix between the 2 action and rest to evaluate the accuracy.

Validation on simulated HD-EMG data

In order to validate the proposed calibration approach against a controlled ground truth, the pipeline was first applied to simulated HD-EMG data. Simulated datasets were generated using Neuromotion [18], an integrated neuromusculoskeletal simulation framework that produces motor unit action potential waveforms and their corresponding spike trains from prescribed movement inputs.

We generated two datasets corresponding to a hand-closing movement: (1) a single sustained contraction of 10 s duration and (2) eight repetitions of 2 s held contractions. For each dataset, the simulation produced motor unit spike trains and action potential waveforms, which were convolved to generate HD-EMG signals representative of recordings obtained from a surface electrode grid.

Both datasets were processed through the full decomposition pipeline described above, including

segmentation and concatenation for the repeated-movement condition. Decomposition accuracy was evaluated using the Rate of Agreement (ROA) metric. The ROA is defined as:

$$ROA = (C / (C + O + I)) \times 100\%$$

where C is the number of common spikes identified in both the ground-truth and decomposed spike trains, O is the number of spikes present only in the ground-truth, and I is the number of spikes present only in the decomposition output. Decomposed MUs with ROA greater than 30% with any ground truth unit were retained as valid.

Patient recruitment

This study was approved by the Ethics Committee of the Medical University of Vienna (approval number: EK-2280/2024). One participant with stroke affecting upper limb motor function was recruited from the Department of Neurology at the Vienna General Hospital (AKH Wien). Inclusion criteria required a diagnosis of ischemic or hemorrhagic stroke, at least 2 weeks from the last event, ability to sit in up-right position, motor Fugl-Meyer-Upper Extremity Scale between 10 and 66, age between 18 and 80 years. The participant provided written informed consent in accordance with the Declaration of Helsinki prior to any study procedures. All recordings were carried out under the supervision of a licensed physiotherapist.

Experimental setup

High-density electromyographic (HD-EMG) signals were recorded using two monopolar electrode arrays (GR08MM1305, OT Bioelettronica) with an inter-electrode distance of 8 mm and 32 channels. One grid was placed longitudinally along the muscle belly of the flexor and the second on the extensors, on the forearm as required by the target movement. Signals were amplified and digitized at 2 kHz through a Grapevine Trek Amplifier (Ripple Neuromed).

The participant was seated comfortably in front of a screen with the forearm resting on a table. Four hand and wrist movements were tested: (1) wrist extension, (2) finger extension, (3) a three-digit pinch, and (4) flexion of the fourth and fifth digits. For each movement, two recording conditions were acquired. In the first condition, the participant performed a sustained contraction, limited to approximately 10s due to participant fatigue. In the second condition, eight repetitions of 3 s contractions were recorded, each initiated by cues from the recording interface. Between recordings the subject could rest to recover from fatigue.

A custom recording interface provided visual cues for movement and relaxation using an animated virtual hand. Cue timestamps were saved with the HD-EMG signals for alignment, and an external camera recorded the session to account for delays between cues and movement onset. Signals from both sustained and repeated conditions were then processed using the described decomposition pipeline.

RESULTS

The following results provide preliminary validation of the proposed approach. In simulation, repeated-contraction and sustained-contraction paradigms yielded comparable decomposition performance, supporting the validity of the proposed calibration strategy. We further applied this framework in a stroke participant, demonstrating that HD-EMG decomposition can enable two-degree-of-freedom external device control while also revealing candidate biomarkers relevant to recovery and rehabilitation.

Repetitions of a movement can be used in place of a single movement for calibration

In simulated signal, the decomposition algorithm yielded 43 MUs from the held contraction, of which 13 were validated against ground truth (69,81% Hallucination Rate) vs the repetition-based approach which yielded 40 MUs, of which 12 were validated against ground truth (70% Hallucination Rate).

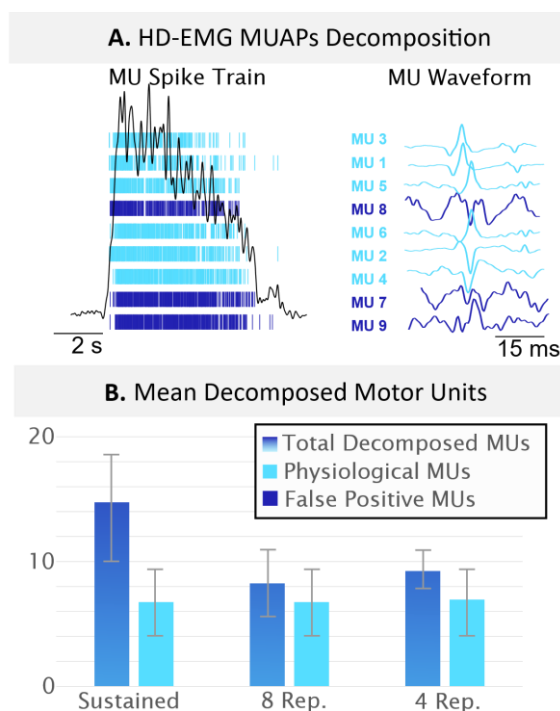


Figure 2. Repeated-contraction calibration preserves motor unit yield while suppressing spurious detections. A) Example of MUAPs Decomposition during 1 contraction of tripod pinch. Left: MU firing train corresponding to EMG envelop. Right: Decomposed waveform showing both physiological and false positive waveforms. B) total decomposed versus physiologically validated motor units across calibration conditions, averaged over four movements (N=4).

Although the yield and hallucination rate were similar in the simulations, a different pattern emerged from the data recorded from the subject (Fig. 2). Averaged across four movements, held contractions produced 14.5 ± 4.15

decomposed MUs of which only 6.75 ± 2.59 were validated (53.4% hallucination rate), whereas decomposition from repeated contractions yielded 8.25 ± 2.59 decomposed MUs with 6.75 ± 2.28 validated (18.2% hallucination rate).

Restricting the analysis to exactly four repetitions, in order to test the lower limit for effective decomposition, yielded similar results (9.25 ± 2.68 decomposed MUs, 6.95 ± 1.48 validated; 24.9% hallucination rate). Critically, the number of validated MUs remained stable across conditions (~ 6.75 – 6.95 MUs).

HD-EMG as BCI with 2 degrees of freedom

Motor unit filters and the whitening matrix were computed using four of the eight recorded repetitions.

From this first offline decomposition, two motor units with distinct activation profiles were identified. MU1 (Blue in Fig. 3) fired selectively during finger flexion (Movement 4), while MU2 (Orange in Fig. 3) was active during both tripod pinch (Movement 3) and finger flexion (Movement 4).

This differential recruitment enabled a straightforward classification rule: when only MU2 exceeded the firing rate threshold (10 Hz), the output was classified as tripod pinch; when both MU1 and MU2 were concurrently active, the output was classified as finger flexion.

As shown in Figure 3, the raw HD-sEMG signal exhibits clear modulation aligned with the cued movements, with shaded regions indicating the timing of each action. The firing rates of both motor units reliably crossed the detection threshold during their respective movements, allowing a simple transfer function to map motor unit activity to discrete commands for virtual or robotic hand control with perfect precision.

HD-EMG analysis can provide bio-markers to track recovery

In the offline analysis phase, we investigated whether motor units extracted from the stroke patient exhibited unique firing characteristics or waveform features that could reflect underlying corticospinal pathway alterations. During each recorded action, we identified the presence of a distinctive oscillation that was temporally locked to the MUAP waveforms

This satellite peak was observed at a median delay of 22.2ms (IQR: 20.5–22.6ms) across all recorded movements (Fig. 4). The proportion of motor units displaying the satellite potential varied by task, ranging from 33.3% in tripod pinch (5/15 units) to 58.3% in wrist extension (7/12 units), with finger extension and 4th+5th finger flexion at 44.4% (4/9) and 42.9% (3/7), respectively.

DISCUSSION

The findings of this study address three interrelated challenges in deploying HD-EMG-based BCIs for neurological populations: calibration feasibility, real-time control, and longitudinal monitoring of recovery.

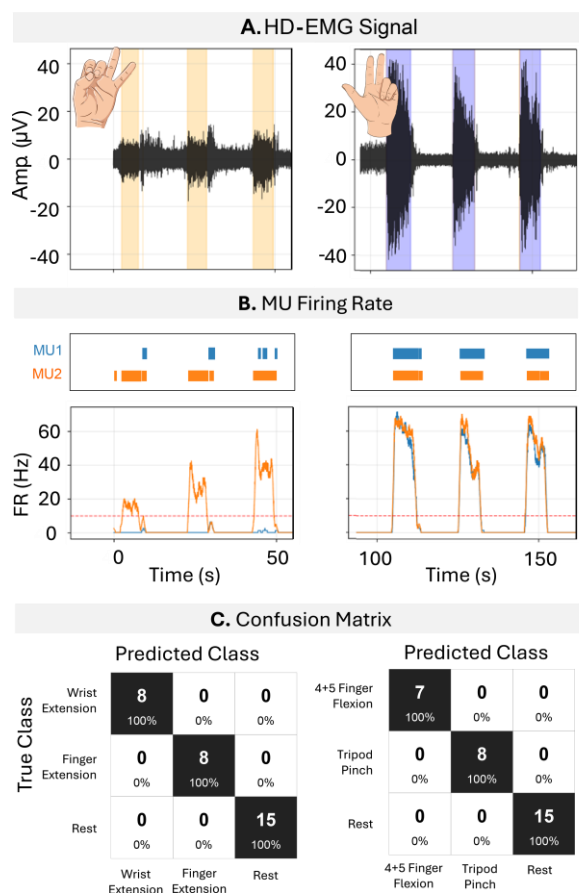


Figure 3. Real-time two-degree-of-freedom control driven by individual motor unit recruitment. (a) Raw HD-sEMG from a single channel during finger flexion and tripod pinch, with (b) instantaneous firing rates of MU1 (Blue) and MU2 (Orange) below. Selective recruitment — MU2 alone for tripod pinch, both units for finger flexion — enables discrete command encoding. (c) Confusion matrices confirming perfect classification accuracy using a simple firing-rate threshold rule.

The repeated-command calibration strategy addresses a fundamental barrier to the clinical deployment of HD-EMG decomposition in neurological populations. By concatenating signal segments from multiple brief contractions, the proposed approach preserves the statistical structure necessary for source separation while accommodating the reduced motor endurance characteristic of these populations.

Notably, the number of physiologically validated motor units remained stable across calibration conditions (~6.75–6.95 MUs), even as the hallucination rate dropped from 53.4%, with held contractions to 18.2% with repeated movements. This reduction in false-positive detections is particularly consequential for real-time applications, where spurious motor units would introduce erroneous control commands.

The demonstration of two-degree-of-freedom control using only two selectively recruited motor units highlights the potential of HD-EMG to move beyond amplitude-based classification schemes that have traditionally constrained myoelectric interfaces. By exploiting the differential recruitment patterns of individual motor units, a simple thresholding rule

achieved perfect classification accuracy during simulated online decoding.

Beyond real-time control, our findings suggest that motor unit analysis may yield clinically informative biomarkers. We observed satellite potentials occurring 44.2 % of the decomposed motor unit, a feature that earlier intramuscular EMG studies attributed to collateral reinnervation and newly formed synaptic connections following upper motor neuron lesions [21].

To our knowledge, the present findings represent the first demonstration of such satellite potentials detected from surface HD-EMG recordings, extending prior needle-based observations to a non-invasive recording modality. If validated in larger cohorts, this feature could complement existing discharge-based biomarkers, such

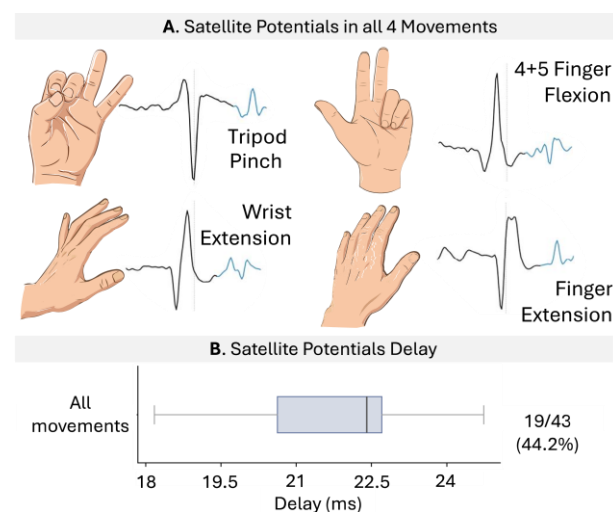


Figure 4. Surface-detected satellite potentials as candidate biomarkers of corticospinal reorganization. A secondary oscillation time-locked to the MUAP appears at a median delay of 22.2ms in 44.2% of decomposed motor units (N = 43).

as alterations in firing rate variability and motor unit synchronization [10][19][20] to provide a more comprehensive characterization of corticospinal recovery.

Several limitations should be acknowledged. The primary constraint is the small sample size, which precludes formal statistical inference and limits the generalizability of these findings.

The classification scheme, while effective, was designed for a controlled proof-of-concept setting; its robustness under conditions of fatigue, electrode shift, and variable contraction intensity remains to be established. Furthermore, the current two-degree-of-freedom paradigm, though clinically meaningful, falls short of the multi-joint control that would be required for functional neuroprosthetic applications, which will likely require decomposition of motor units across multiple muscles and the development of more sophisticated hardware pipelines [22]. Finally, the control was computed online but not in real time, owing to the computational cost of the decomposition step, which currently requires several minutes per movement. This processing time made it

impractical to keep the patient for the full duration, and we instead resorted to simulating the online control in the manner previously described.

CONCLUSION

This study demonstrates that a repeated-command calibration paradigm enables reliable HD-EMG decomposition in an acute stroke patient while substantially reducing spurious motor unit detections compared to sustained contractions. The stability of validated motor units across calibration conditions, combined with the successful real-time classification of two hand movements using only two selectively recruited units, supports the feasibility of motor unit-driven BCIs in neurological populations with limited motor endurance. Furthermore, the identification of satellite potentials in surface-recorded MUAP waveforms suggests that HD-EMG decomposition may serve a dual role: providing both real-time control signals and non-invasive biomarkers of corticospinal reorganization. While these findings are derived from a single participant and require validation in larger cohorts, they establish a foundation for clinically practical HD-EMG-based interfaces that accommodate the functional constraints of individuals with severe motor impairment.

ACKNOWLEDGEMENTS

Special thanks to our dear colleagues Anna Sparapani, and Hyunjeong Lee for their assistance with the experiment.

REFERENCES

- [1] GBD 2021 Stroke Risk Factor Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* 2024
- [2] Lawrence ES, Coshall C, Dundas R, Stewart J, Rudd AG, Howard R, Wolfe CDA. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke.* 2001
- [3] Kwakkel G, Kollen BJ, van der Grond J, Prevo AJH. Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke. *Stroke.* 2003
- [4] Chaudhary U, Birbaumer N, Ramos-Murguialday A. Brain–computer interfaces for communication and rehabilitation. *Nat Rev Neurol.* 2016
- [5] Biasucci A, Leeb R, Iturrate I, Perdakis S, et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nat Comm.* 2018
- [6] Philipp Raggam, Manuel Eder, Alexia-Theodora Popa, Peter Fugger, Moritz Grosse-Wentrup, An EMG-based brain-computer interface for communication-impaired patients, 9th BCI Conference Graz, 2024.
- [7] Farina D, Holobar A, Merletti R, Enoka RM. Decoding the neural drive to muscles from the surface electromyogram. *Clin Neurophysiol.* 2010
- [8] Ting JE, Del Vecchio A, Sarma D, Verma N, Colachis SC 4th, Annetta NV, Collinger JL, Farina D, Weber DJ. Sensing and decoding the neural drive to paralyzed muscles during attempted movements of a person with tetraplegia using a sleeve array. *J Neurophysiol.* 202
- [9] Oliveira DS, Ponfick M, Braun DI, Osswald M, et al. A direct spinal cord–computer interface enables the control of the paralysed hand in spinal cord injury. *Brain.* 2024
- [10] Tacca N, Levine JT, Heimann MK, Schlink BR, et al. Leveraging neural drive to assess hand motor function in individuals with chronic stroke. *J Neuroeng Rehabil.* 2026
- [11] Negro F, Muceli S, Castronovo AM, Holobar A, Farina D. Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation. *J Neural Eng.* 2016
- [12] Holobar A, Zazula D. Multichannel blind source separation using convolution kernel compensation. *IEEE Trans Signal Process.* 2007;55(9): 4487–4496
- [13] Alghamdi I, Ariti C, Williams A, Wood E, Hewitt J. Prevalence of fatigue after stroke: a systematic review and meta-analysis. *Eur Stroke J.* 2021;6(4): 319–332
- [14] Barsakcioglu DY, Bräcklein M, Holobar A, Farina D. Control of spinal motoneurons by feedback from a non-invasive real-time interface. *IEEE Trans Biomed Eng.* 2021;68(3): 926–935
- [15] Pranav Mamidanna, Thomas Klotz, Dimitrios Halatsis, Agnese Grison, et al. MUniverse: A Simulation and Benchmarking Suite for Motor Unit Decomposition, Open Review, 2025
- [16] Kothe C, Shirazi SY, Stenner T, Medine D, et al. The lab streaming layer for synchronized multimodal recording. *Imaging Neuroscience.* 2025;3: IMAG.a.136
- [17] Grison A, Gibbs C, Rawji V, Gouveia Vila L, et al. Multidimensional motoneuron control using intramuscular microelectrode arrays in tetraplegic spinal cord injury. *medRxiv [Preprint].* 2025.
- [18] Ma S, Mendez Guerra I, Caillet AH, Zhao J, et al. NeuroMotion: open-source platform with neuromechanical and deep network modules to generate surface EMG signals during voluntary movement. *PLoS Comput Biol.* 2024
- [19] Oliveira DS, Carbonaro M, Raiteri BJ, Botter A, Ponfick M, Del Vecchio A. The discharge characteristics of motor units innervating functionally paralyzed muscles. *J Neurophysiol.* 2025;
- [20] Shandiz E, Gaskell A, Ingram M, Trajano GS, et al. Applying high-density surface EMG to the study of neuromuscular disorders: a systematic review. *Clin Neurophysiol.* 2025
- [21] Benecke, R., Berthold, A. & Conrad, B. Denervation activity in the EMG of patients with upper motor neuron lesions: time course, local distribution and pathogenetic aspects. *J Neurol.* 1983
- [22] Cimolatto A, Raspopovic S. Closing the sensory feedback loop is necessary for effective neurorehabilitation. *PLoS Biol.* 2024

Single case study of Spatiotemporal Dynamics of Cortical Preparation and Motor Execution Across Lower-Limb Movements of Varying Complexity

S. Pathak¹, L. von der Linde¹, J.J. Hernández-Gloria^{1,2}, P. Ofner¹, N. Mrachacz-Kersting^{1,3}

1. Institute of Sport and Sport Science, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

2. Department of Microsystems Engineering-IMTEK, University of Freiburg, Freiburg, Germany

3. BrainLinks-BrainTools Center, IMBIT, Albert-Ludwigs University of Freiburg, Freiburg, Germany

E-mail: shreyansh.pathak@email.uni-freiburg.de

ABSTRACT: Most lower-limb brain-computer interface (BCI) studies focus on either isolated single-joint movements or complex movements like gait initiation and walking, leaving an unexplored gap between the transition from one movement type to another. As part of the REWIRING project, we aim to investigate the spatiotemporal relationship between cortical signals and muscle activation across various motor tasks of increasing biomechanical complexity and different volition types. To accomplish this, in an ongoing pilot study, electroencephalographic and electromyographic data were recorded and analysed from a single participant during three self-paced movements of varying complexity: dorsiflexion, plantar flexion, and gait initiation. A preliminary analysis of the spatially-filtered data extracted from Cz indicates identifiable movement related cortical potential (MRCP) patterns among all investigated tasks, with differences in peak-negativity and waveform morphology. Our findings demonstrate the feasibility of the proposed setup for investigating and understanding the differences of MRCPs across various lower-limb movements in future studies.

INTRODUCTION

Motor impairments following stroke or spinal cord injury (SCI) often result from disruption of descending efferent pathways, while cortical motor intent is preserved [1]. Non-invasive BCIs aim to decode movement intention using methods such as electroencephalography (EEG) to trigger external assistive devices [2–4]. When the decoded movement intention is synchronized with peripheral sensory feedback, it promotes plasticity, leading to reorganization of the central nervous system and functional recovery [5, 6].

One prominent neural marker for endogenous movement intention is the Movement-Related Cortical Potential (MRCP) [7–9]. MRCPs are slow negative shifts in EEG (typically 0.1-3 Hz), primarily originating from pre-motor cortical areas and primary motor cortex (M1), that accompany voluntary movement and reflect cortical processes of motor preparation, planning,

execution, and feedback processing [7, 10–12]. Some foundational studies have demonstrated the significance of precisely pairing the peak-negativity (PN) of MRCPs with peripheral afferent feedback in inducing associative cortical plasticity, making MRCPs an ideal trigger signal for BCIs in rehabilitation scenarios [5, 6]. Previous BCI studies that use the MRCP as the control signal have used cue-based movements [13]. However, natural body movements are predominantly self-paced. Studies have investigated the morphological disparities in MRCPs between self-paced and cue-based movements, noting that self-paced movements exhibit a distinct readiness potential (Bereitschaftspotential, BP), whereas cue-based movements are dominated by the Contingent Negative Variation (CNV) [7, 9, 14, 15]. Our upcoming study, for the REWIRING project [16], will include tasks such as seated plantar flexion (PF) and dorsiflexion (DF) of the dominant foot, gait initiation (GI), and continuous walking (WA) during both self-paced (SP) and cue-based (CB) conditions, measured across more than 20 able-bodied individuals. As an initial step toward this investigation, a pilot experiment was conducted to evaluate the technical feasibility of the experimental setup and to investigate the MRCP morphology across self-paced lower-limb movements of varying complexity. Here, we present a preliminary comparison of MRCP morphology obtained from a single participant from this ongoing pilot study.

MATERIALS AND METHODS

Participants: The current preliminary analysis involved one able-bodied subject aged 25 years, with written informed consent. The study protocol was conducted according to the Declaration of Helsinki. The protocol was structured as a 4x2 factorial design, consisting of eight randomized experimental blocks. Only three (DF, PF, and GI) of these eight blocks were analysed here. Each experimental block comprised 40 trials, yielding a total of 120 trials.

Data Acquisition: During each experimental block, EEG and electromyographic (EMG) data were recorded with two g.USBamps (g.tec Medical Engineering GmbH, Austria).

EEG data were acquired using a 28-channel active electrode system positioned according to the international 10–20 system. Particular emphasis was placed on the electrodes over the motor cortex and midline regions (Fz, FCz, Cz, CPz, Pz, C3, C4). Signals were recorded in a monopolar configuration referenced to the earlobe ipsilateral to the dominant leg, with a ground electrode placed at FPz.

For recording muscle activity, unipolar surface EMG electrodes were placed on the Tibialis Anterior (TA) and Soleus muscles of both the dominant and non-dominant legs. The reference and ground electrodes were placed on the electrically neutral bony prominence of the dominant knee. All signals were sampled at 1200Hz.

Additionally, an unimotor pedal system was used to measure the force exertion and foot positioning during dorsiflexion and plantar flexion tasks while sitting, as well as for gait initiation while standing. A force platform was used to measure ground reaction forces during the walking trials. Throughout the session, a goniometer and an accelerometer were placed on the dominant ankle foot to record ankle rotation angle and foot acceleration. This additional data was not included in this preliminary analysis.

		Factor B: Volition types	
		Self-paced	Cue-based
Factor A: Movement types	Planter flexion	Self-paced plantar flexion 40 trials Condition: SP-PF	Cue-based plantar flexion 40 trials Condition: CB-PF
	Dorse-flexion	Self-paced dorsiflexion 40 trials Condition: SP-DF	Cue-based dorsiflexion 40 trials Condition: CB-DF
	Gait initiation	Self-paced gait initiation 40 trials Condition: SP-GI	Cue-based gait initiation 40 trials Condition: SP-GI
	Walking	Self-paced walking 40 trials Condition: SP-WA	Cue-based walking 40 trials Condition: CB-WA

Figure 1: Experimental protocol: 4x2 factorial design.

Experimental Procedure: The first factor shown in Fig. 1, ranges over four levels of movement complexity, focusing on the dominant leg. During dorsiflexion and plantar flexion, the participant was instructed to perform a ballistic movement with the dominant leg while maintaining the rest of the body steady. Gait initiation was performed while standing on the pedal system and transitioning from a quiet stance to the first step of the dominant leg without lifting off. During walking trials, the participant was instructed to walk for 2-3 steps across a force platform and return to the start position. The second factor shown in Fig. 1, was related to the volition type. For self-paced trials, the participant was given a start cue before each run, upon which they could start the trials at their will, spacing each trial by roughly 8 to 10 seconds. A stop cue was given 5-10 seconds

after the last trial. For cue-based trials, the participant performed the movements by following a cursor on a screen. Each cursor screen was followed by a ‘REST’ and preceded by a ‘FOCUS’ screen, each of approximately 4-second duration [5]. All data was acquired using a previously developed MATLAB-based application. Here, the data from self-paced dorsiflexion, plantar flexion, and gait initiation were analysed for our preliminary analysis.

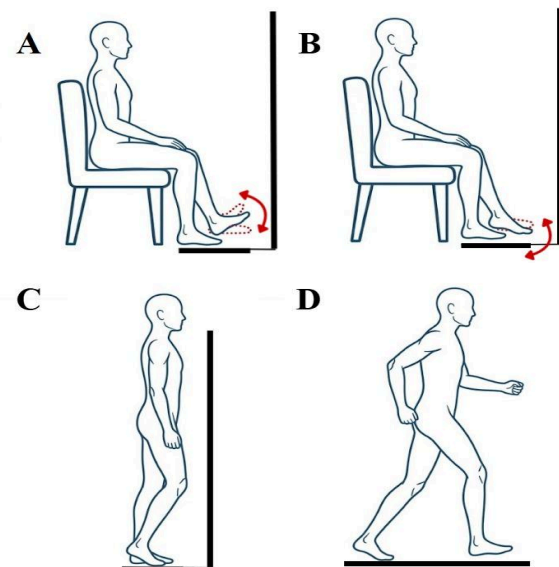


Figure 2: Movement Types; A: Dorsiflexion, B: Plantar Flexion, C: Gait Initiation, D: Walking. Only A, B, and C were evaluated in our preliminary analysis of our pilot study.

Signal Preprocessing: Continuous EEG data for DF, PF, and GI were recorded. The data was filtered using a 0.1 Hz high pass filter. Due to high noise level, channel P4 was removed after visual inspection. Subsequently, Independent Component Analysis (ICA) was performed to identify and remove artifacts, specifically eye movements and blinks. One component associated with eye blinks was removed and the components were projected back to EEG. A notch filter at 50 Hz was used to eliminate power-line noise. A zero-phase 2nd-order Butterworth band-pass filter (0.1 - 3 Hz) was applied to isolate the MRCPs. The data was then segmented into epochs of -4 to +6 seconds relative to movement onsets.

Epoching: To prevent temporal smearing of the grand-averaged MRCP, precise alignment of "Time Zero" is vital. For all discrete conditions (dorsiflexion, plantarflexion, gait initiation), $t=0$ was defined as the onset of the primary agonistic EMG burst (TA for dorsiflexion, Soleus for plantar flexion and gait initiation). The onset was detected using a

Teager-Kaiser Energy Operator (TKEO) thresholding method [17]. Epochs were extracted from $t = -4.0$ s to $t = +6.0$ s relative to EMG onset.

MRCP Alignment and Averaging: To enhance the spatial specificity of cortical activity over the motor cortex, a large Laplacian spatial filter centered at Cz was computed for each epoch. To reduce temporal variability across trials, the peak-negativities (PN) were aligned. For each epoch, the local minimum of the Laplacian Cz signal was identified within a window of -0.5 s to $+0.5$ s around the EMG onset. Trials were then realigned and cut to a duration of -3.0 s to $+5.0$ s based on PN before averaging. Each epoch was then individually baseline corrected. The mean signal value during the first second of the epoch (-3 to -2 s relative to PN) was calculated and subtracted from the entire trial to remove slow drifts and inter-trial offsets.

For each movement condition, the aligned trials were averaged to obtain the mean MRCP waveform. The standard error of the mean (SEM) was computed across trials to quantify variability. The resulting mean MRCPs for dorsiflexion, plantar flexion, and gait initiation were then superimposed to allow visual comparison of waveform morphology across the three tasks.

MRCP Feature Extraction: To further characterize task-dependent differences in MRCP morphology, several temporal and amplitude-based features were extracted from the averaged Laplacian Cz waveforms after peak-negativity (PN) alignment and baseline correction. The early Bereitschaftspotential (BP) slope (NS1) was calculated within the interval from -2.0 s to -0.5 s relative to PN, while the late BP slope (NS2) was calculated from -0.5 s to 0 s. A post-movement recovery slope was additionally computed from 0 s to $+1.0$ s. For each interval, a first-order linear fit was applied to each trial, and the resulting slopes were averaged across trials for each movement condition. The peak-to-peak (P2P) amplitude was calculated as the difference between the most negative deflection and the maximum positive deflection observed within the interval of ± 2 s around the PN. All quantitative features are reported as mean $\pm$ standard error of the mean (SEM) across trials.

RESULTS

Fig. 3 shows the averaged MRCPs at the Cz electrode for the three self-paced tasks: dorsiflexion (DF), plantar flexion (PF), and gait initiation (GI), time-locked to peak negativity (PN) ($t = 0$). All conditions exhibit the characteristic MRCP morphology consisting of a gradual negative potential preceding movement onset, followed by a return toward baseline after movement initiation.

The MRCP features differed across the three movement conditions. The Peak-to-peak amplitudes were comparable for DF ($11.68 \mu\text{V}$) and PF ($11.10 \mu\text{V}$), whereas GI exhibited a smaller overall peak-to-peak excursion ($7.85 \mu\text{V}$).

Differences were also observed in the temporal evolution of the readiness potential. GI exhibited the steepest early BP slope (NS1: $-5.69 \pm 1.68 \mu\text{V/s}$), whereas DF demonstrated the strongest late BP slope (NS2: $-16.51 \pm 2.33 \mu\text{V/s}$). PF showed comparatively smaller negative slopes during both preparatory phases. In the post-movement interval (NS3), all conditions exhibited positive recovery slopes, with DF showing the largest recovery rate.

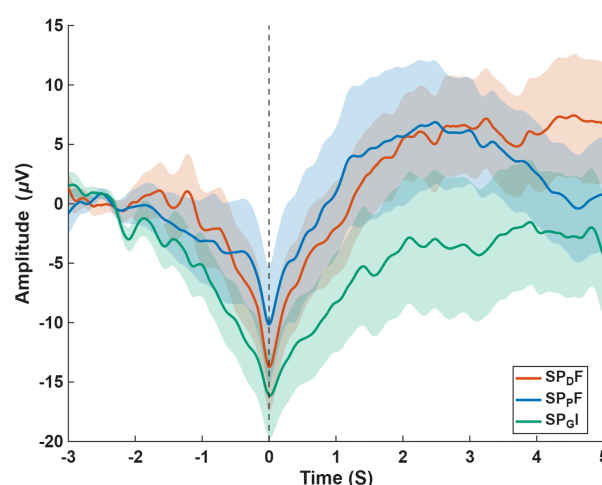


Figure 3: Mean MRCPs at electrode Cz for self-paced (SP) DF (SPDF), PF (SPPF), and GI (SPGI). Shaded regions represent $\pm$ SEM across trials. The dashed line indicates peak negativity (PN) ($t=0$).

DISCUSSION

The framework presented in this study addresses a critical bottleneck in the translation of BCI-based neuro-rehabilitation: the transition from simple, isolated motor intentions to the control of complex locomotor actions. By characterizing the MRCPs across a spectrum of movement complexities, we intend to establish a comprehensive “morphological library” of motor intention.

The present work represents an exploratory pilot analysis based on data from a single participant ($N = 1$). Consequently, statistical comparisons between movement conditions on a group-level were not feasible. In addition, only three self-paced movement conditions were evaluated, limiting the generalizability of the findings to broader motor tasks or clinical rehabilitation populations.

Nevertheless, the quantified MRCP features revealed distinct waveform characteristics across dorsiflexion,

plantar flexion, and gait initiation. In particular, gait initiation exhibited the largest motor potential amplitude and the steepest early BP slope, whereas dorsiflexion demonstrated the strongest late BP slope immediately preceding movement onset. These observations suggest the possibility of task-dependent differences in cortical preparatory dynamics associated with varying movement complexity.

Importantly, the present study does not establish that these features can reliably discriminate movement classes in a classifier framework. Rather, the results serve as an initial indication that MRCP-derived features such as early and late BP slopes, motor potential amplitude, and latency may contain discriminative information worth investigating in larger cohort studies. If validated across multiple participants and sessions, such features could potentially support the development of more adaptive BCI systems capable of differentiating between simple ankle movements and more complex locomotor actions such as gait initiation [2].

Although preliminary, the results presented here provide an initial validation of the proposed experimental framework and motivate future studies involving larger participant cohorts, additional movement conditions, and clinical populations relevant to neuro-rehabilitation applications.

CONCLUSION

This study outlines the feasibility of a comprehensive, multimodal paradigm for characterizing the spatiotemporal and cortico-muscular dynamics of lower-limb movements. Building upon this pilot, we aim to map the diverse neural signatures required for asynchronous, intent-driven, closed-loop BCI rehabilitation systems that can account for variations in the interactions involved in movements of varying complexity.

ACKNOWLEDGEMENTS

This work was supported by ERA-NET NEURON - The Network of European Funding for Neuroscience Research.

REFERENCES

[1] Ang KK, Guan C, Chua KSG, Ang BT, Kuah C, Wang C, et al. Clinical study of neurorehabilitation in stroke using EEG-based motor imagery brain-computer interface with robotic feedback. In: Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC). Buenos Aires, Argentina, 2010, pp.

5549–5552.
[2] Savić AM, Aliakbaryhosseinabadi S, Blicher JU, Farina D, Mrachacz-Kersting N, Došen S. Online control of an assistive active glove by slow cortical signals in patients with amyotrophic lateral sclerosis. *J. Neural Eng.* 2021;18(4):046085.
[3] Hernández-Gloria JJ, Jaramillo-Gonzalez A, Savić AM, Mrachacz-Kersting N. Toward brain-computer interface speller with movement-related cortical potentials as control signals. *Front. Hum. Neurosci.* 2025;19:1539081.
[4] Wolpaw JR, McFarland DJ. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proc. Natl. Acad. Sci. U.S.A.* 2004;101(51):17849–17854.
[5] Mrachacz-Kersting N, Jiang N, Stevenson AJ, Niazi IK, Kostic V, Pavlovic A, et al. Efficient neuroplasticity induction in chronic stroke patients by an associative brain-computer interface. *J. Neurophysiol.* 2016;115(3):1410–1421.
[6] Mrachacz-Kersting N, Kristensen SR, Niazi IK, Farina D. Precise temporal association between cortical potentials evoked by motor imagination and afference induces cortical plasticity. *J. Physiol.* 2012;590(7):1669–1682.
[7] Shakeel A, Navid MS, Anwar MN, Mazhar S, Jochumsen M, Niazi IK. A review of techniques for detection of movement intention using movement-related cortical potentials. *Comput. Math. Methods Med.* 2015;2015:346217.
[8] Niazi IK, Jiang N, Tiberghien O, Nielsen JF, Dremstrup K, Farina D. Detection of movement intention from single-trial movement-related cortical potentials. *J. Neural Eng.* 2011;8(6):066009.
[9] Jiang N, Gizzi S, Mrachacz-Kersting N, Dremstrup K, Farina D. A brain-computer interface for single-trial detection of gait initiation from movement related cortical potentials. *Clin. Neurophysiol.* 2015;126(1):154–159.
[10] Shibasaki H, Hallett M. What is the Bereitschaftspotential? *Clin. Neurophysiol.* 2006;117(11):2341–2356.
[11] Cui RQ, Huter D, Lang W, Deecke L. Neuroimage of voluntary movement: topography of the Bereitschaftspotential, a 64-channel DC current source density study. *Neuroimage.* 1999;9(1):124–134.
[12] Cunningham R, Ianssek R, Bradshaw JL, Phillips JG. Movement-related potentials associated with movement preparation and motor imagery. *Exp. Brain Res.* 1996;111(3):429–436.
[13] Aliakbaryhosseinabadi S, et al. Detection of movement intention from movement-related cortical potentials with different paradigms. In: Jensen W, Andersen O, Akay M (Ed.). *Replace, Repair, Restore, Relieve – Bridging Clinical and Engineering Solutions in Neurorehabilitation.* Biosyst. Biorobot., vol. 7. Springer, Cham, 2014.
[14] Jankelowitz SK, Colebatch JG. Movement-related potentials associated with self-paced, cued and

- imagined arm movements. *Exp. Brain Res.* 2002;147(1):98–107.
- [15] Savić A, et al. Movement related cortical potentials and sensory motor rhythms during self initiated and cued movements. In: Jensen W, Andersen O, Akay M (Ed.). *Replace, Repair, Restore, Relieve – Bridging Clinical and Engineering Solutions in Neurorehabilitation*. Biosyst. Biorobot., vol. 7. Springer, Cham, 2014.
- [16] Neuron ERANET. JTC2024 – Brain-Body: REWIRING. Unlocking and re-establishing effective mechanisms of the wiring between brain and lower limb to improve recovery of walking after stroke. Available from: <https://www.neuron-eranet.eu/projects/REWIRING/>
- [17] Yang D, Zhang H, Gu Y, Liu H. Accurate EMG onset detection in pathological, weak and noisy myoelectric signals. *Biomed. Signal Process. Control.* 2017;33:306–315.

Combining Mindfulness and Implicit Neurofeedback to Upregulate Frontal Alpha Asymmetry

E. Percolla¹, S. Drapart¹, A. Kübler¹, M. Pfeiffer¹

¹ Intervention Psychology, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

E-mail: elisa.percolla@uni-wuerzburg.de

ABSTRACT: This contribution presents an experimental design that integrates an implicit auditory neurofeedback protocol in a mindfulness intervention to pursue an upregulation of frontal alpha asymmetry (FAA). The following sections detail the design of such an experiment, together with the development and implementation of the materials in use. Lastly, preliminary data are provided to support its feasibility and discuss future improvements and applications.

INTRODUCTION

Neurofeedback (NF) originates with the neural signal of interest and the neuroimaging method used for its acquisition, as in standard observational studies but adds two more steps: the online processing of the neural signal received and its real-time return to the source in form of comprehensible sensory stimuli. Through the iteration of this finely tuned closed-loop communication, participants can learn to modulate their neural activity in a target direction by adjusting their internal states or behavior accordingly [1].

Explicit and Implicit NF modalities are distinguished by the level of awareness experimentally afforded to participants regarding the feedback signal. In the explicit modality, participants achieve experimental goals through intentional effort while in contrast, the implicit modality withholds explicit feedback and let the underlying learning process operate independently of conscious awareness [2]. Explicit NF protocols currently predominate the literature, a trend largely driven by their proven efficacy for specific neural targets e.g., the sensorimotor rhythm (SMR) [3]. We argue that implicit NF offers a functional framework for those brain-behaviour correlates that are less responsive to traditional explicit protocols as is the case for Frontal Alpha Asymmetry (FAA).

FAA is the power difference in the alpha band (8–13 Hz) between frontal cortices. Since alpha power is inversely related to cortical activation, relatively lower left-alpha is a measure of greater left-frontal activation, while lower right-alpha represents greater right-frontal activation. A FAA upregulation results from a right-alpha increase, a left-alpha decrease or both, the opposite trend defines a FAA downregulation [4]. Lower left-alpha is associated with approach motivation and positive affect, except for anger, while lower right-alpha is associated to

withdrawal and negative emotional states [5].

FAA NF and the Capability Model: Many baseline recordings and classical approach-avoidance and/or NF tasks did not yield significant FAA modulation, particularly when the objective was FAA upregulation [6,7]. These findings can be explained by the capability model, which states that active emotional entanglement is an essential requisite of any FAA investigation [8]. We hypothesize that the use of traditional explicit NF for the upregulation of FAA may not be adequate. Neutral stimuli, such as visual scales or simple tones, lack the emotional salience required to effectively elicit FAA; this could explain the absence of modulation often reported in the FAA literature. Even when a FAA modulation is elicited via explicit NF, it may be intrinsically biased toward downregulation. For instance, if the experimental goal is FAA upregulation and the participants cannot succeed, repeated failed trials might bias them toward frustration or disengagement. In such cases, the resulting negative emotional state acts as a functional barrier, keeping FAA consistently downregulated regardless of the intended training.

Mindfulness-based Implicit NF: The strategy behind the experimental design here is to provide solely positive implicit feedback (PIF) when FAA is upregulated. The implicit modality should help to reduce evaluation-related FAA downregulation [9]. However, presenting PIF to the participants in a way such that they perceive it as pleasant but do not detect it as feedback is a challenge. We aimed at addressing it by masking the PIF within a mindfulness meditation.

Mindfulness is the process of regulating attention to bring a quality of non-elaborative awareness to current experience and a quality of relating to one's experience within an orientation of curiosity, experiential openness, and acceptance [10]. Given that practicing mindfulness is proven to upregulate FAA [11,12], mindfulness can serve as a robust implicit strategy for the proposed NF protocol. In this contribution, we present a mindfulness-based auditory NF protocol, outlining solutions to potential pitfalls and pilot data confirming its feasibility.

METHODS AND MATERIALS

Participants: Three expert meditators: two young instructors and one senior advanced practitioner

participated in the study (biological women ≥ 18 years old, right-handed, without a history of neurological disorders age:). All participants were recruited through digital advertising and were experienced practitioners with 3 to 10 years of experience in meditation and regular mindfulness practice over the last year. All participants were invited to the EEG lab at the Department of Clinical Psychology and Psychotherapy, Section Intervention Psychology at the Julius-Maximilian Universität (JMU) in February 2026 and received monetary compensation (13.00 € per hour) for their participation. All procedures were carried out in accordance with the ethical standards of the Association of German Professional Psychologists (Deutsche Gesellschaft für Psychologie, DGPs) and the Declaration of Helsinki [13].

Procedure: Participants were initially informed that the study aimed at testing a novel, hybrid mindfulness intervention combining a reduced voice-over with instrumental components. This framing was used to maintain the implicit nature of the NF protocol and ensure the authenticity of the evaluation. Participants completed a single session of the custom-made mindfulness intervention. In detail, they listened to pre-recorded mindfulness guidelines, performed a resting state recording and four meditation-practice blocks. The meditation-practice blocks will henceforth be referred to as 'blocks'. At the end of the experiment, they filled in a questionnaire about state mindfulness. Following the session participants were debriefed.

EEG Setup: EEG data were recorded using 64 ActiCap active electrodes and two BrainAmp DC amplifiers (Brain Products GmbH, Gilching, Germany). Electrodes were positioned according to the international 10–20 system, with the reference at FCz and the ground at Fpz. The impedance check was performed using ActiCap Control Software (v. 1.2.5.5); all initial impedance levels were below 5 k Ω before data collection started. Data were streamed via the BrainAmp LSL Connector (v. 1.18) using liblsl (v. 1.16). Recording was performed at 500 Hz using LabRecorder (v. 1.16), and signal quality was monitored online through BrainVision LSL Viewer (v. 1.1.1).

FAA Resting State: Various methods exist for calculating FAA [14], this contribution adopts the formula from Coan et al. [15], which involves subtracting log-transformed alpha power at the electrode site F3 (left frontal hemisphere) from that at electrode site F4 (right-frontal hemisphere). We recorded individual baselines following the guidelines proposed by Hagemann et al. [16] to calibrate the system for the derivation of artifact thresholds and to obtain a baseline FAA distribution for z-normalization and further comparison between spontaneous thinking and meditation. The resting state lasted 8 minutes and consisted of 8 1-min blocks of spontaneous thinking with eyes opened (EO) or closed (EC) in the following fixed sequence: EO - EC - EC - EO - EC- EO - EO - EC. Participants have been

instructed to either keep their gaze on the fixation cross (EO) or close their eyes (EC) and to avoid movement or muscle contractions. In the baseline instructions, it was specified that this was not yet a meditation-practice block and that the participants were invited to let their thoughts run spontaneously as in everyday life.

NF Processing was achieved by a Python framework [17], adapted for this study.

Online Processing Before the NF: Since the participants were asked to meditate with eyes opened, for calibration and online processing, just the EO blocks of the resting state were used. The signal was re-referenced online to common average reference (CAR), detrended (DC offset removal) and down sampled to 250 Hz. Artifact thresholds were determined as the smallest absolute value at which the z-score exceeded 2.58, constrained to 80–180 μ V. Data were then bandpass-filtered (1–40 Hz) and used to fit Riemannian Artifact Subspace Reconstruction (rASR) for online artifact correction. Data were then transformed using current source density (CSD). Power spectral density (PSD) was computed using Welch's method. Mean and standard deviation of band power between 8–13 Hz at F3 and F4 were stored independently and then used to calculate FAA baseline distribution.

Online Processing During the NF: FAA was computed from a 2 s sliding window updated every 100 ms. The stream was referenced to a common average reference (CAR), detrended, bandpass filtered, corrected using the pre-trained rASR, transformed to CSD before PSD was computed with Welch's method. Power in alpha band at F3 and F4 electrode sites was extracted, log-transformed and used to calculate FAA. FAA was normalized to baseline z-scores which were mapped to the auditory feedback stimulus with a direct logic such that higher z-scores (FAA upregulation) corresponded to an increase in the volume of the positive feedback stimulus.

Offline Processing after the NF: Continuous data were referenced to a Common Average Reference (CAR), processed with a 1–100 Hz bandpass filter (8th-order Butterworth IIR) and then segmented into 5s epochs. Independent Component Analysis (ICA) (Infomax; max 500 iterations) was used for artifact rejection. Components accounting for 99% of the variance were retained and classified using ICLabel; any non-brain components identified with $\geq 70\%$ certainty were removed. After discarding remaining epochs with amplitudes exceeding 150 μ V, we applied the same processing used in our online pipeline: data were re-referenced to CSD and bandpass-filtered to a 1–40 Hz range. PSD at electrodes F3 and F4 was then estimated using Welch's method (Hamming window, 50% overlap), and alpha power was extracted and log transformed to calculate FAA. This offline processing pipeline was developed in Python 3.14.3.

Auditory Implicit NF: Designing the PIF for the blocks posed a complex balancing act. The stimulus had to be noticeable and recognizable to enable implicit learning yet remain aesthetically consistent with the background audio. Smooth transitions were essential as well; any sudden 'on/off' switching of the PIF would have created a discordant sound, breaking the participant's immersion. At the same time, the challenge behind smooth transitions was ensuring these seamless fades did not compromise the precise timing required for real-time NF. To address these design challenges, we implemented several key solutions. First, we opted for instrumental background audio over a guided meditation to ensure the voice-over would not overshadow the PIF. To guarantee harmonic compatibility between the background audio and the PIF, instead of using two different audio tracks, we chose one and decomposed it into two distinct but complementary layers: a static, non-evolving audio for the background and a more salient, "short-peaking" one for the PIF. We adjusted the melody from myNoise [18] relying also on pleasantness ratings coming from the platform users in the forum.

Although previous FAA studies utilized fixed thresholds (e.g., ± 0.85 SD) [7,15], we mapped z-scores to a continuous volume scale (0–100%). However, initial testing revealed that the PIF was still discontinuous near the threshold of human hearing due to rapid FAA fluctuations. To prevent abrupt audio cut-offs, we introduced a buffered decay zone within the lower 10% of the volume range. In this zone, volume changes were restricted to 1% change per sample, ensuring that even if the FAA value dropped suddenly, the audio would take at least 10 samples (1s) to fade out completely. While this introduces a minor trade-off in temporal precision, it was necessary to maintain a seamless meditative flow. Above this 10% threshold, the volume tracked the timing and magnitude of the FAA modulation in real-time. The auditory pipeline was implemented using PyQt5 with supplementary audio processing performed in Audacity (v3.7.7).

Mindfulness Guidelines: The mindfulness guidelines consisted of a 7-minute script, which has been recorded with Audacity 3.7.7 in a soundproof booth at the lab.

Adjustments to the Lab Setting: While meditators are traditionally encouraged to remain non-reactive to external distractions—such as the buzzing of an insect or distant traffic, these examples were adapted to the lab environment. Participants were invited to maintain this same patience and non-judgmental awareness toward sensations specific to the setup, such as the EEG cap on their head, the contact with the conductive gel, or the cables connecting them to the equipment. To help participants feel at ease in an unfamiliar laboratory setting, the script also employed a specific communication style characterized by a welcoming tone of voice, simple language, the use of a direct second person singular and classic mindfulness metaphors, such

as 'watching your thoughts like clouds passing through the sky'. While the instructions were delivered through a conversational narrative—designed to feel like a personal sharing of practical tips—the script followed a rigorous, cyclical structure intended to systematically address the foundational principles of mindfulness, consistently returning to the core concept of present-moment awareness [10,19].

Participants' Auditory Priming: The mindfulness guidelines were overlaid with the auditory stimulus reserved for the PIF in the subsequent meditation sessions. This approach was designed to prime the participants, ensuring that the first instance of FAA-driven PIF would effectively facilitate the implicit recognition of the signal as a marker of present-moment awareness. Audio implementation was performed with Audacity (v3.7.7).

Meditation-Practice Blocks: The meditation was not guided to prioritize the auditory saliency of the PIF. However, the background stimulus had to be mildly pleasant and immersive nonetheless to ensure the participant succeeded being presented with the PIF (FAA successful upregulation). A strategy to keep the participants engaged in a meditative state involved the insertion of short verbal anchors. These mindfulness reminders followed the thematic progression of the initial mindfulness guidelines and were presented every 50 seconds. A more detailed description can be found in Tab. 1. The participants performed four 5min identical blocks.

Table 1. Mindfulness Intervention – Core Elements

Mindfulness Principles in the Guidelines	Reminders during the Meditation with English Translation
Present-focused Awareness	1 Dieser Moment gehört nur dir. <i>This moment belongs entirely to you.</i>
	2 Lausche deinen Gedanken, ohne sie zu beurteilen. <i>Listen to your thoughts without judging them.</i>
Non-Judgment	3 Sei geduldig, sei freundlich zu dir selbst. <i>Be patient, be kind with yourself.</i>
	4 Entspanne deinen Körper, befreie deinen Geist. <i>Relax your body, free your mind.</i>
Patience and Self-Compassion	5 Kehre zurück zum gegenwärtigen Augenblick. <i>Return to the present moment.</i>

Subjective Measure of State Mindfulness: Instruments for measuring state mindfulness are limited, as most existing scales focus on trait mindfulness and/or lack validation in German [20]. The State Mindfulness Scale (SMS) [21] is one of the few state mindfulness questionnaires translated and validated in German (SMS-PA) [22]. The

SMS-PA consists of 12 items divided into two subscales: Mindfulness of the Mind (items 1–6) and Mindfulness of the Body (items 7–12). Given that the physical subscale only contains items irrelevant to the current sedentary intervention (e.g. movement, temperature changes, or physical fatigue), participants completed only the 'Mindfulness of the Mind' subscale. Original items and their validated translations are presented in Tab. 2.

Table 2: Items from the State Mindfulness of the Mind Subscale (SMS-PA-M)

No	Original English Items (Cox et al., 2016)	Validated German Items (Wallner et al., 2024)
M1	I was aware of different emotions that arose in me.	Ich war mir verschiedener Emotionen bewusst, die in mir entstanden sind.
M2	I noticed pleasant and unpleasant emotions.	Ich habe angenehme und unangenehme Emotionen wahrgenommen.
M3	I noticed pleasant and unpleasant thoughts.	Ich habe angenehme und unangenehme Gedanken wahrgenommen.
M4	I noticed emotions come and go.	Ich habe wahrgenommen, wie Emotionen kamen und gingen.
M5	I noticed thoughts come and go.	Ich habe wahrgenommen, wie Gedanken kamen und gingen.
M6	It was interesting to see the patterns of my thinking.	Es war interessant, meine Denkmuster zu beobachten.

RESULTS

As we present here a feasibility study, we report results only descriptively.

Figure 1: Fig. 1 (A) illustrates the standardized FAA for each participant during each block using box- and strip plots. Each data point represents a 5s epoch, color-coded to distinguish between the resting state (pink; spontaneous thinking) and meditation-practice blocks (turquoise; meditation). Horizontal black lines in the boxplot indicate the median FAA score per block. The y-axis indicates the block type, where 0 represents the resting state and 1–4 represent the subsequent meditation-practice blocks. All recordings were processed with the offline pipeline described above and visualized with R (v. 4.5.2). The last two blocks for Participant 1 were excluded due to technical problems in the lab. Fig. 1(B) displays individual participant responses to the six items of the State Mindfulness of the Mind Subscale reported in Tab. 1. Ratings range from 0 ('not at all') to 4 ('very much'). The bars are coloured turquoise to maintain visual consistency with the meditation-practice blocks in Figure 1(A), as the evaluation focused explicitly on the meditation-practice

period rather than the resting state.

FAA and SMS-PA-M Descriptives: our results indicate that FAA levels were modulated by the experimental condition: in detail, while a downregulation was observed across all subjects for the first meditation-practice block, subsequent blocks were marked by a shift toward upregulation. One exception was Participant 3, whose data reflected higher variance across all blocks. Participant 2 reported the highest state mindfulness ratings and exhibited the most pronounced shift toward FAA upregulation, in fact, after the first meditation-practice block, this participant's values remained consistently equal or above baseline. Interestingly, the data revealed a collective decrease in ratings for items 2 and 3.

Post-experimental Interviews: confirmed that participants remained unaware of the NF. None reported unusual observations or held correct prior assumptions. Notably, they could not differentiate the PIF from the baseline audio until they listened to them separately. Regarding study duration, participants noted that while four blocks were manageable for experts, a 15-minute intervention might be more appropriate for non-expert populations to avoid fatigue. One participant suggested maintaining the 5-min block duration, noting that physical immobility—rather than the meditative state itself—becomes challenging over longer periods. The auditory materials were consistently described as 'pleasant', 'calming' or 'relaxing,' with no modifications suggested for the mindfulness guidelines or reminders. The first two participants reported concerns regarding their performance, noting that they required additional time to reach a meditative state within the experimental environment. All participants, especially 1 and 3, found the eye-open requirement highly counterintuitive. Participant 2 noted that although the experience with eyes open was unusual, she was able to acclimate quickly, instead, participant 1 reported that upon entering a deep meditative state, she began to close her eyes almost unintentionally, only to be momentarily distracted by the need to consciously re-open them to comply with the protocol.

DISCUSSION

Achieving a meditative state is rarely instantaneous, even outside a laboratory; here, experimental constraints likely exacerbated this latency and may have elicited performance anxiety. This could explain the initial FAA downregulation observed across the sample. Furthermore, the effort required to maintain open eyes, especially for those with different meditation habits, may have increased cognitive workload and disrupted the meditative flow. For instance, Participant 1 reported the need to actively counteract the urge to close her eyes when a deeper meditation state unfolded. In contrast, Participant 2 struggled less with this requirement, potentially explaining her higher mindfulness ratings and

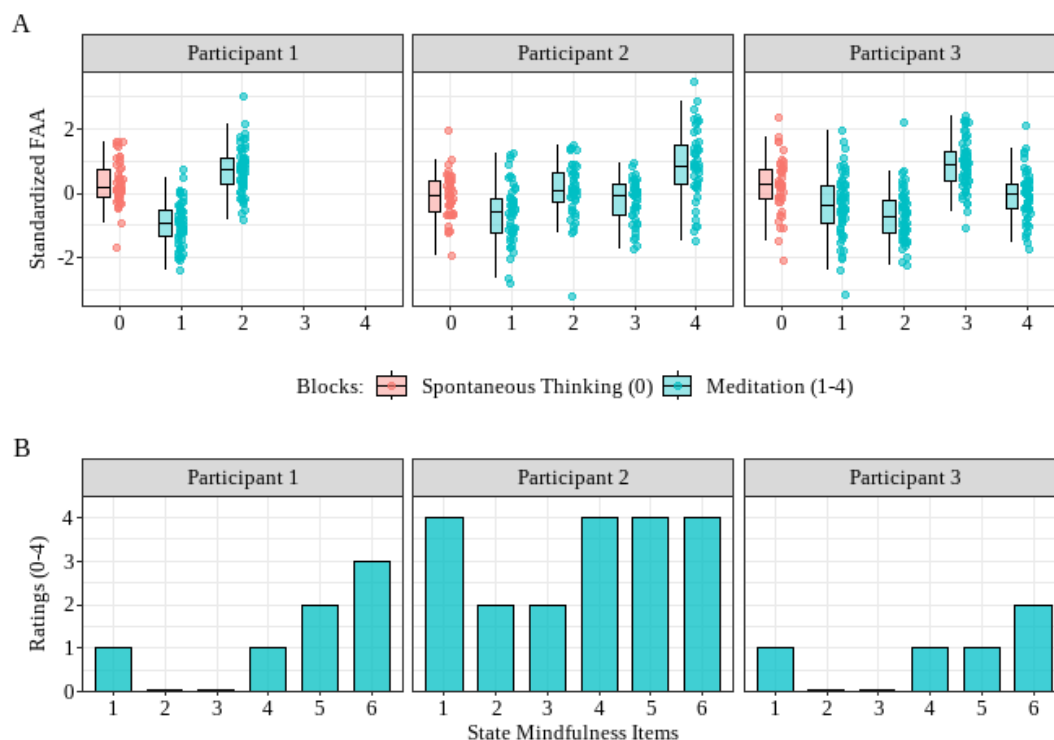


Figure 1: Combined visualization of (A) standardized Frontal Alpha Asymmetry (FAA) across experimental blocks and (B) behavioral ratings for State Mindfulness items. Data are presented for three individual participants.

clearer FAA upregulation. Together, these factors may when a deeper meditation state unfolded. Together, these factors may account for the lack of consistent upregulation across blocks. The general lower ratings for items 2 and 3 of the SMS-PA-M warrant further investigation.

Eyes Open/Closed in FAA Literature: Approach-avoidance FAA research is more focused on eyes-open designs [5,23] and the same applies to NF research both aimed at investigating FAA [6,7,24] or meditation [12,25,26], likely because the NF protocols adopted were visual. To date, eyes-closed NF-FAA research has been limited to a single publication investigating its role as a baseline predictor for learners/non-learners [27]. Our eyes-open design contrasts with the broader meditation literature, where studies targeting FAA exclusively employed eyes-closed protocols [11,28,29]. Given that our protocol does not strictly require active gazing, exploring an eyes-closed direction is a logical next step.

Choice of Meditation: Mindfulness remains one of the most successful meditation techniques for targeting FAA, alongside breathwork meditation [29]. We opted against breathing-based protocols to protect the signal-to-noise ratio, as irregular or deep breathing patterns can introduce artifacts that compromise EEG signal quality and NF accuracy. Coherently, our mindfulness guidelines focus on present-moment awareness with minimal bodily reference beyond general relaxation.

CONCLUSION

As a proof-of-concept, this contribution confirms that positive auditory feedback can be integrated into an instrumental mindfulness meditation without compromising the qualitative depth of the practice. To establish whether sustained FAA modulation is achievable, future research should move beyond the illustrative single sessions by programming a minimum of 5-6 sessions in accordance with previous NF-FAA literature. Moreover, rigorous control groups are necessary to disentangle the effects of mindfulness from the NF, alongside more comprehensive psychometric testing and experimental tasks enabling us to measure pre-to-post-performance changes.

ACKNOWLEDGMENTS

The authors thank Dr. Botrel for the NF framework, Philipp Ziebell and Dr. Rodrigues (Julius-Maximilians-Universität Würzburg) for their expertise on FAA models and Prof. Dr. Kleih-Dahms (Otto Friedrich Universität Bamberg) for her insights on neurofeedback. This work was supported by the German Research Foundation [grant number RTG 2660 – Approach Avoidance] assigned to Andrea Kübler.

AI Disclosure: Google Gemini (2.0 Flash) refined the text clarity and aesthetics of Figure 1. The authors validated the output and remain fully responsible.

REFERENCES

- [1] Sitaram R, Ros T, Stoeckel L, Haller S, Scharnowski F, Lewis-Peacock J, et al. Closed-loop brain training: the science of neurofeedback. *Nat Rev Neurosci* 2017;18(2):86–100.
- [2] Muñoz-Moldes S, Cleeremans A. Delineating implicit and explicit processes in neurofeedback learning. *Neuroscience & Biobehavioral Reviews* 2020;118:681–8.
- [3] Reiner M, Gruzelier J, Bamidis PD, Auer T. The Science of Neurofeedback: Learnability and Effects. *Neuroscience* 2018;378:1–10.
- [4] Coan JA, Allen JJB. Frontal EEG asymmetry and the behavioral activation and inhibition systems. *Psychophysiology* 2003;40(1):106–14.
- [5] Rodrigues J, Müller M, Mühlberger A, Hewig J. Mind the movement: Frontal asymmetry stands for behavioral motivation, bilateral frontal activation for behavior. *Psychophysiology* 2018;55(1):e12908.
- [6] Peeters F, Ronner J, Bodar L, Van Os J, Lousberg R. Validation of a neurofeedback paradigm: Manipulating frontal EEG alpha-activity and its impact on mood. *International Journal of Psychophysiology* 2014;93(1):116–20.
- [7] Mennella R, Patron E, Palomba D. Frontal alpha asymmetry neurofeedback for the reduction of negative affect and anxiety. *Behaviour Research and Therapy* 2017;92:32–40.
- [8] Coan JA, Allen JJB, McKnight PE. A capability model of individual differences in frontal EEG asymmetry. *Biological Psychology* 2006;72(2):198.
- [9] Yonah R. In Neurofeedback Training, Harder is Not Necessarily Better: The Power of Positive Feedback in Facilitating Brainwave Self-Regulation. *NR* 2023;10(1):31–41.
- [10] Bishop SR, Lau M, Shapiro S, Carlson L, Anderson ND, Carmody J, et al. Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice* 2004;11(3):230–41.
- [11] Krishna D, Singh D, Manjunath NK. Determining the depth of meditation through frontal alpha asymmetry. *Front Hum Neurosci* 2025;19:1576642.
- [12] Navarro Gil M, Escolano Marco C, Montero-Marín J, Minguez Zafra J, Shonin E, García Campayo J. Efficacy of Neurofeedback on the Increase of Mindfulness-Related Capacities in Healthy Individuals: a Controlled Trial. *Mindfulness* 2018;9(1):303–11.
- [13] WMA - The World Medical Association-WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants n.d. <https://www.wma.net/policies-post/wma-declaration-of-helsinki/> (accessed March 16, 2026).
- [14] Vincent KM, Xie W, Nelson CA. Using different methods for calculating frontal alpha asymmetry to study its development from infancy to 3 years of age in a large longitudinal sample. *Developmental Psychobiology* 2021;63(6):e22163.
- [15] Allen JJB, Harmon-Jones E, Cavender JH. Manipulation of frontal EEG asymmetry through biofeedback alters self-reported emotional responses and facial EMG. *Psychophysiology* 2001;38(4):685–93.
- [16] Hagemann D. Individual differences in anterior EEG asymmetry: methodological problems and solutions. *Biological Psychology* 2004;67(1–2):157–82.
- [17] Botrel L, Kreilinger A, Müller M, Pfeiffer M, Scheu V, Vowinkel N, et al. The influence of time and visualization on neurofeedback-guided parietal alpha downregulation and sense of presence in virtual reality. *Front Neurosci* 2025;19:1476264.
- [18] Pigeon DIS. Background Noises • Ambient Sounds • Relaxing Music | myNoise ® n.d. <https://mynoise.net/index.php> (accessed March 16, 2026).
- [19] Kabat-Zinn J. An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results. *General Hospital Psychiatry* 1982;4(1):33–47.
- [20] Blanke ES, Brose A. Mindfulness in Daily Life: a Multidimensional Approach. *Mindfulness* 2017;8(3):737–50.
- [21] Cox AE, Ullrich-French S, French BF. Validity Evidence for the State Mindfulness Scale for Physical Activity. *Measurement in Physical Education and Exercise Science* 2016;20(1):38–49.
- [22] Wallner J, Kind L, Donath C, Kornhuber J, Luttenberger K. Validation of the German version of the state mindfulness scale for physical activity in a clinical sample. *J Patient Rep Outcomes* 2024;8(1):140.
- [23] Paul K, Beauducel A, Hennig J, Hewig J, Hildebrandt A, Kührt C, et al. Frontal alpha asymmetry as a marker of approach motivation? Insights from a cooperative forking path analysis. *Journal of Personality and Social Psychology* 2025;128(1):196–210.
- [24] Quaedflieg CWEM, Smulders FTY, Meyer T, Peeters F, Merckelbach H, Smeets T. The validity of individual frontal alpha asymmetry EEG neurofeedback. *Soc Cogn Affect Neurosci* 2016;11(1):33–43.
- [25] Brandmeyer T, Delorme A. Closed-Loop Frontal Midline θ Neurofeedback: A Novel Approach for Training Focused-Attention Meditation. *Front Hum Neurosci* 2020;14:246.
- [26] Treves IN, Greene KD, Bajwa Z, Wool E, Kim N, Bauer CCC, et al. Mindfulness-based Neurofeedback: A Systematic Review of EEG and fMRI studies 2024.
- [27] Nan W, Wan F, Tang Q, Wong CM, Wang B, Rosa A. Eyes-Closed Resting EEG Predicts the Learning of Alpha Down-Regulation in Neurofeedback Training. *Front Psychol* 2018;9:1607.
- [28] Moyer CA, Donnelly MPW, Anderson JC, Valek KC, Huckaby SJ, Wiederholt DA, et al. Frontal Electroencephalographic Asymmetry Associated With Positive Emotion Is Produced by Very Brief Meditation Training. *Psychol Sci* 2011;22(10):1277–9.
- [29] Barnhofer T, Chittka T, Nightingale H, Visser C, Crane C. State Effects of Two Forms of Meditation on Prefrontal EEG Asymmetry in Previously Depressed Individuals. *Mindfulness* 2010;1(1):21–7.

Beyond Accuracy: User Experience Matters in the Comparison of Two EEG Headsets for BCI

S. Petrinotti¹, A. Bigioni^{1,3}, V. Caracci^{1,2,3}, S. Boesso¹, J. Toppi^{1,2}, F. Cincotti^{1,2}, D. Mattia¹, A. Riccio¹

¹ Neuroelectrical Imaging and BCI Laboratory, Fondazione Santa Lucia IRCCS, Rome, Italy

²Dep. of Computer, Control and Management Engineering, Sapienza University of Rome, Italy

³Fondazione Neurone Onlus, Rome, Italy

E-mail: s.petrinotti@hsantalucia.it

ABSTRACT: Brain-Computer Interfaces (BCIs) are a promising Assistive Technology (AT) for individuals with severe motor impairments, enabling communication and interaction through neural activity without requiring muscular control. However, their integration into everyday assistive services remains limited, partly due to usability issues related to EEG acquisition devices. In particular, the comfort and acceptability of EEG headsets are key factors for long-term adoption.

This study compared users' subjective experience with two EEG headsets during a P300-based BCI spelling task: a gel-based cap (EasyCap), commonly used in laboratory and clinical settings, and a mobile semi-dry system (X.on) designed for out-of-lab applications. Ten healthy volunteers completed two sessions, performing the same task with each headset. After each session, participants filled out a questionnaire assessing satisfaction with headset, along with Visual Analogue Scales (VAS) evaluating perceived physical discomfort. Results showed higher overall satisfaction with the semi-dry headset, suggesting greater suitability for everyday BCI communication.

INTRODUCTION

Damage or dysfunction affecting the nervous system can severely impair motor function, limiting communication and interaction with the environment and reducing quality of life [1, 2].

Assistive Technologies (AT) have evolved to support communication and interaction for individuals with motor impairments through alternative input devices (e.g., eye-trackers, switches, alternative keyboards) and adaptive software solutions [1 - 4].

Despite advances in AT, individuals with severe motor impairments still face barriers to access. Brain-Computer Interfaces (BCIs) offer an alternative by translating neural activity into control signals without

requiring muscular output [2, 4, 5].

Among the different types of BCI systems, a subset is designed for communication and relies on the P300 event-related potential [1, 2, 6]. These systems typically operate according to a visual oddball paradigm, in which the user focuses attention on a target item displayed on a screen without performing any physical movement. The target appears as a rare stimulus among frequent non-target stimuli and elicits a P300 component in the EEG signal. By detecting this response across successive stimulus presentations, the system can infer the attended item, allowing the user's selection to be identified and used to enable communication or control external devices through brain activity [7].

Despite the growing body of research on BCIs, the integration of these devices into everyday assistive technology services is still limited. One of the main challenges is that most systems remain confined to laboratory settings, requiring specialized expertise for setup and operation. To fully realize their potential, BCIs need to be optimized for daily use, becoming more portable, easier to use, and suitable for home and everyday living environments [1, 2].

To achieve this goal, a User-Centered Design (UCD) approach is essential, ensuring that development is guided not only by technical performance but also by user experience factors such as comfort, usability, and satisfaction, which are crucial for system acceptance [8]. One of the key components that can influence the usability and acceptability of a BCI system is the EEG acquisition device worn by the user. The comfort, weight, fit, and overall experience of wearing the headset can significantly affect both performance and motivation to use it. A variety of EEG headsets are currently available, differing in their montage, number and type of electrodes, and acquisition technology. Based on the method of contact with the scalp, it is possible to distinguish between *gel-based*, *dry* and *semi-dry* headsets [9 - 14]. The first ones, also known as *wet* systems, represent the standard in neuroscience

and clinical research due to the high quality and stability of the signal they provide, but require a relatively long setup time and the need to wash hair, which influence their use in everyday contexts [9 - 11]. Dry headsets, in contrast, do not show this constraint, but they can cause physical discomfort due to the pressure of pin electrodes on the scalp [12, 13]. An intermediate category consisting of semi-dry systems, based on water or saline solution, represents a compromise between feasibility and signal reliability: they are quick to apply and do not leave any residues in the hair [10, 11, 14].

Furthermore, in order to incorporate the user's point of view, it is important to highlight that comfort and physical well-being are crucial factors for the prolonged use of EEG headsets: dry systems with pin electrodes often cause significant discomfort or pain over time, leading to headaches and user fatigue [9, 11, 12, 13]. User preferences are also strongly influenced by aesthetics and design. Gel-based EEG caps are often not appreciated, while modern and discreet designs are preferred [11, 13]. Headsets with electrodes positioned on the mastoids or around the ears are especially appreciated for being less noticeable [15]. Moreover, users tend to prefer headsets that are easy to use and quick to set up. Wet systems, with their slow preparation and the need to wash hair after each session, are less appealing for everyday or home use. In contrast, dry and semi-dry systems are preferred because they can be put on autonomously, without technical assistance, making them more practical and user-friendly for daily use [9, 13, 15].

Nonetheless, in this context, it is important to also consider the accuracy of BCI performance, that is to say the success rate of correctly classified user intentions during BCI operations. Most of the literature on EEG-based BCIs focuses on gel-based systems, which have been extensively studied in terms of signal quality and classification accuracy. However, some studies directly compare wet, dry, and semi-dry headsets, highlighting some differences in performance: dry EEG systems seems to exhibit lower accuracy compared to gel-based systems, with gel electrodes consistently providing more reliable classification performance [11, 15]. Furthermore, when evaluating signal quality, gel-based systems appear to maintain their superiority, exhibiting better signal-to-noise ratios [16]. Thus, the focus should be on achieving an optimal trade-off between BCI performance and user preference.

This study compares two types of EEG headsets, a gel-based cap widely used in clinical and research settings, and a mobile semi-dry system intended for use outside the laboratory, to control a P300-based BCI for communication. Both classification accuracy and user satisfaction are evaluated. The ultimate aim is to identify which system provides the best balance between BCI performance and user satisfaction, and is therefore better suited for integration into AT

solutions.

MATERIALS AND METHODS

Participants

Ten healthy volunteers participated in the study (mean age $\pm$ sd = 28.8 ± 4.3 ; 3 males and 7 females). According to the inclusion criteria, all participants were over 18 years of age and reported no history of neurological or psychiatric disorders. All research was conducted in accordance with the principles embodied in the Declaration of Helsinki. The study protocol was approved by the Independent Ethics Committee of Fondazione Santa Lucia in Rome.

Each participant completed two experimental sessions on the same day, separated by an interval of approximately 30 minutes, resulting in a total experiment duration of about 2 hours and half. The order of the sessions was counterbalanced across participants.

EEG acquisition

EEG signals were recorded to compare the performance of two different headsets during a P300-based BCI spelling task. In one session, participants wore an *EasyCap EEG headset* (Brain Products) with active gel electrodes, connected to a LiveAmp amplifier. A 12-channel montage was used (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, PO7, PO8, Oz), with reference and ground placed on the right earlobe. Although EEG activity was recorded from all 12 channels, only signals from 8 channels (Fz, Cz, P3, Pz, P4, PO7, PO8, Oz) were used for the online task. These channels correspond to the well-established *8-channel montage* widely employed in the literature for P300-based BCIs [17]. While signals from all 12 channels were recorded, only these 8 channels were transmitted to the BCI2000 software, and the classifier was trained using features derived exclusively from these channels [17]. In the other session, participants performed the same task using the *X.on EEG headset* (Brain Products), equipped with passive semi-dry electrodes in a 7-channel montage (F3, F4, C3, Cz, C4, P3, P4), with reference and ground on the right earlobe. The amplifier was integrated into the headset.

To ensure comparable conditions between the two montage types, EEG cap placement was performed by the same operator for all participants and across all sessions. The operator had no prior experience with either montage procedure, thereby ensuring an equivalent level of expertise for both setups. As no technicians with extensive experience in both cap types were available, an operator with no prior experience with either system was selected. Furthermore, the same setup procedure was adopted for both montages, consisting of skin preparation with alcohol cleansing in order to reduce impedance, cap positioning, and application of conductive gel or saline solution to facilitate electrical conduction between the scalp and the electrodes.

EEG headset design

From an aesthetic point of view, the two headsets are different. The EasyCap is made of fabric and has a fixed size, equipped with a chin strap for secure fastening. In addition to the fabric headset, an earclip is provided for ground and reference. The connections are made via cables to a central connection unit, which must subsequently be connected to the amplifier for EEG signal recording. Conversely, the X.on features a rigid structure composed of multiple arms, includes an adaptable frontal strap, and is adjustable for head circumferences between 54 and 60 cm, with an earclip also present for ground and reference. It integrates sponge electrodes and a built-in amplifier, so that no cables are needed between the headset and the external amplifier. Both the headsets are shown in Fig. 1.

P300 spelling task

As mentioned above, the spelling task was implemented using the BCI2000 platform [18]. Participants were seated approximately 60 cm from a computer monitor displaying a 5 x 6 matrix of 30 alphabetic and control characters. Rows and columns of the matrix were intensified in random order. Participants were instructed to focus their attention on a target letter indicated on the screen and to silently count the number of times it was highlighted. Each letter selection consisted of 8 row-column stimulation sequences, corresponding to 16 total intensifications. Stimulus duration was 120 ms and the inter-stimulus interval was 120 ms, resulting in a stimulus onset asynchrony (SOA) of 240 ms.

Experimental procedure

Each session consisted of nine runs. First, participants performed four runs of calibration, during which they were shown five-letter words (BACIO, POSTO, METRO, ZEBRA) verbally indicated before each run, with no feedback provided. EEG data from these runs were used to train the BCI classifier via Stepwise Linear Discriminant Analysis (SWLDA), which selects the most relevant features of the EEG signal and builds a linear model to discriminate between target and non-target responses, i.e., to correctly identify the letter the participant is intending to select [17].

Participants then completed five copy-mode runs which followed a structure similar the calibration runs and included online feedback, which was provided through the trained classifier parameters, enabling real-time classification. The words used in this phase were DANZA, GUSTO, MENTE, VIOLA and FIORE.

All the words employed were commonly used words selected from an Italian dictionary and chosen to cover most of the orthographic matrix.



Figure 1: Image of EasyCap (on the left), with the earclip (a) and the LiveAmp amplifier (b) and X.on (on the right), with the earclip (c), sponges (e) and their holders (d).

User perspective assessment

At the end of each session, the participants' subjective experience of the EEG headset used for the task was assessed.

A short questionnaire inspired by the extended Quebec User Evaluation of Satisfaction with assistive Technology ("extended QUEST 2.0") used in Zickler et al., [19] was administered to evaluate satisfaction regarding headset comfort, dimensions, weight, and aesthetic design. Each item was rated on a 5-point Likert scale ranging from 1 ("not satisfied at all") to 5 ("very satisfied").

To further investigate physical discomfort, three Visual Analogue Scales (VAS) adapted from Berger et al. [20] were administered to measure perceived pressure, perceived discomfort and perceived pain, each rated on a scale from 0 (none) to 10 (maximum). All subjective measures were administered by a trained psychologist. To allow the participant to provide an evaluation of the headset's aesthetic design, a mirror was placed in front of them during the headset setup and they were instructed to pay attention to their own image while wearing the headset. The mirror was removed during the task to avoid distractions and was then reintroduced at the end of the session for the administration of questionnaire and VAS.

Statistical analysis

For each participant and each headset, the following subjective indices were extracted from the questionnaire and VAS: *comfort satisfaction*, *dimensions satisfaction*, *weight satisfaction*, *aesthetic design satisfaction*, *overall satisfaction* (calculated as the mean of the four previous scores), *perceived pressure*, *perceived discomfort*, *perceived pain*, and *overall perceived discomfort* (calculated as the mean of the three previous scores).

In addition to these subjective measures, *accuracy* was extracted, calculated as the ratio of correctly selected letters to the total number of selections in the copy-mode phase. Additional performance metrics were also

evaluated during the experiment and will be reported in a separate paper related to the same experimental protocol (Caracci et al., “Evaluating the impact of electrode type and montage on classification performance in assistive P300-based BCIs”, submitted). The present work specifically focuses on the participants’ subjective evaluation.

Differences between the two headsets were then assessed for each index using a two-tailed Wilcoxon signed-rank test. The significance level was set at $\alpha = 0.05$. Then, multiple comparisons were controlled using the Benjamini–Hochberg false discovery rate (FDR) correction. All statistical analyses were conducted using the software *Statistica*.

RESULTS

The primary outcome of interest was the accuracy achieved with the two headsets. It differed significantly between the two systems, with better performance for the EasyCap (p-value = 0.008; results of all the Wilcoxon tests were reported in Tab. 1). The median accuracy was 98 % for the gel-based headset and about 85 % for the semi-dry system, as shown in Fig. 2.

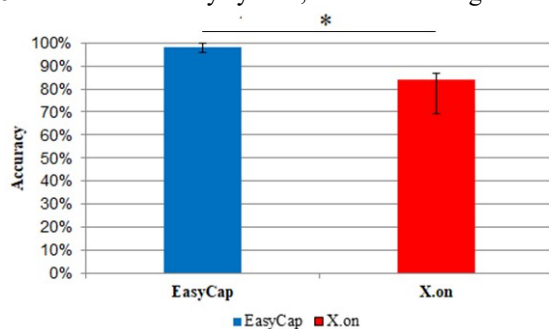


Figure 2: The figure shows the median values of the accuracy distributions for performances obtained with EasyCap (in blue) and with X.on (in red). Bars represents the median values, while error bars indicate the interquartile range (IQR). * denotes significant differences between the conditions ($p < 0.05$).

The results of the subjective assessment obtained through the questionnaire inspired by the extended QUEST 2.0 revealed a statistically significant difference in the overall satisfaction between the two conditions. As shown in Fig. 3a, participants reported significantly higher overall satisfaction (p-value = 0.005) for the X.on (median 4.5, IQR 4.5 - 4.75) compared to the EasyCap (median 3.38, IQR 3 - 3.5). A more detailed analysis showed that comfort satisfaction, weight satisfaction, and aesthetic design satisfaction were also significantly higher for the X.on (comfort: median 5, IQR 4 - 5 vs. median 3, IQR 3 - 3, p-value = 0.008; weight: median 5, IQR 5 - 5 vs. median 4, IQR 4 - 4.75, p-value = 0.028; aesthetic design: median 4.5, IQR 4 - 5 vs. median 2, IQR 2 - 2, p-value = 0.005). The test revealed no significant differences for dimensions satisfaction between the two conditions (median 4, IQR 4 - 5 vs. median 4, IQR 3.25 - 4.75, p-value = 0.286).

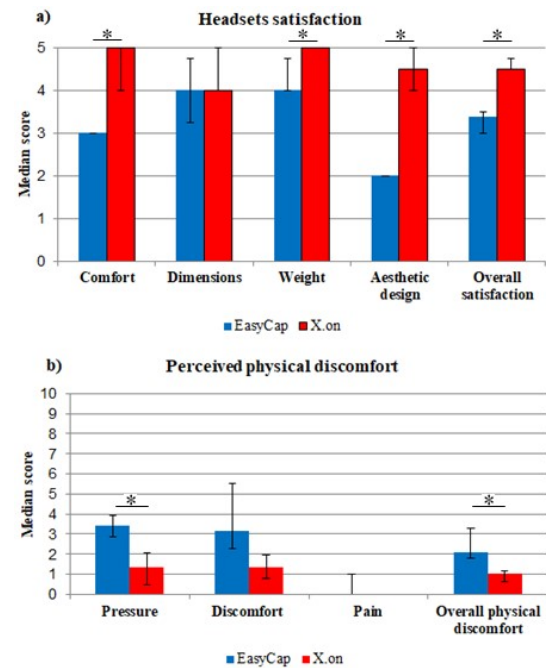


Figure 3: a) Median scores for the items assessing participants’ satisfaction with the two headsets in terms of comfort, dimensions, weight, aesthetic design and overall satisfaction (calculated as the mean of the four item scores). Each item was rated on a scale from 1 to 5. b) Median scores for the VAS assessing perceived pressure, perceived discomfort, and perceived pain (0 - 10 scale). An overall physical discomfort measure (calculated as the mean of the three VAS measures) was also reported. In both a) and b), blue and red bars represent the median scores for the EasyCap and X.on, respectively, with error bars showing the interquartile range. * denotes significant differences between the conditions ($p < 0.05$).

Complementing these quantitative ratings, participants’ qualitative comments on the questionnaire items provided a qualitative assessment of their satisfaction with the two headsets.

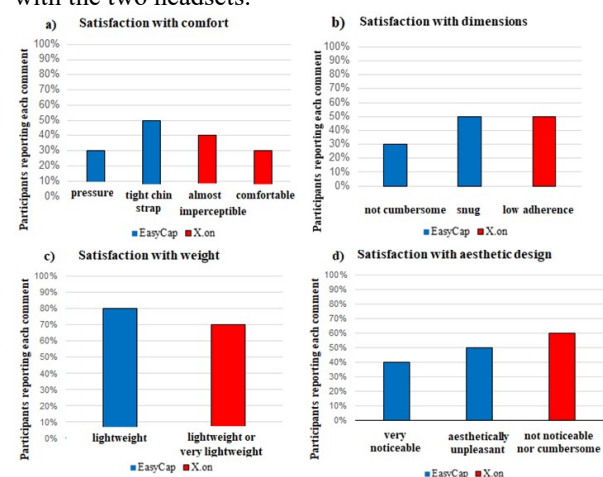


Figure 4: The figure shows participants’ most frequent comments to questionnaire items and corresponding percentages for each headset (blue for EasyCap and red for X.on). Each of the four panels presents a different aspect: a) comfort, b) dimensions, c) weight and d) aesthetic design. The bars indicate the proportion of participants reporting each comment.

As illustrated in Fig. 4, qualitative feedback revealed a generally more positive perception of the *X.on*, particularly in terms of comfort and aesthetic design. Participants frequently described it as comfortable and less noticeable, whereas the *EasyCap* was more often associated with tightness, perceived pressure, and unfavorable aesthetic impressions.

With regard to the results of the VAS, which rated perceived physical discomfort, they are shown in Fig. 3b. The overall discomfort score was significantly different between the two headsets, with higher values reported for EasyCap compared to X.on (median 2.08, IQR 1.78 - 3.31 vs. median 1.03, IQR 0.75 - 1.15, p-value = 0.005).

When examining the individual components, statistically significant differences were also observed for perceived pressure and perceived discomfort, both reported as higher for the EasyCap (pressure: median 3.40, IQR 2.85 - 3.95 vs. median 1.35, IQR 0.45 - 2.08, p-value = 0.005; discomfort: median 3.15, IQR 2.30 - 5.53 vs. median = 1.35, IQR 0.78 - 1.98, p-value = 0.041). However, the effect observed for discomfort did not remain statistically significant after correction for multiple comparisons. No significant differences were found for perceived pain (median 0, IQR 0 - 1 vs. median 0, IQR 0 - 0, p-value = 0.178).

Table 1. Wilcoxon signed-rank test results (Z statistics and p-values, Benjamini-Hochberg corrected).

Variable	Z	P-value	Significant (after Benjamini-Hochberg)
Accuracy	2.7	0.008	significant
Comfort	2.7	0.008	significant
Dimensions	1.1	0.286	not significant
Weight	2.2	0.028	significant
Aesthetic design	2.8	0.005	significant
Overall satisfaction	2.8	0.005	significant
Pressure	2.8	0.005	significant
Discomfort	2.0	0.041	not significant
Pain	1.3	0.178	not significant
Overall physical discomfort	2.8	0.005	significant

DISCUSSION

The present study aimed to determine whether there were differences in users' subjective experience between two EEG headset - a gel-based cap commonly employed in clinical and research environments, and a novel mobile semi-dry headset, designed for out-of-lab applications - during the same task, and to identify the specific factors underlying any observed differences.

Results from the accuracy analyses indicated that the EasyCap was more effective than the X.on. Specifically, the median accuracy was 98 % for EasyCap and about 85 % for X.on. While the latter performed significantly lower than the former, its mean accuracy remained well above 70 %, a threshold previously identified as sufficient for effective communication [21]. This finding suggests that, from a purely effectiveness-driven perspective, both systems could support functional BCI use, even though one clearly outperforms the other.

However, when considering the participants' subjective experience, a different picture emerged. Overall satisfaction, comfort, weight and aesthetic design were consistently rated higher for the X.on. Participants described it as lightweight, almost imperceptible, and visually unobtrusive, whereas EasyCap was frequently associated with excessive tightness, perceived pressure, and less favorable aesthetic impressions. Perceived physical discomfort was also significantly higher for EasyCap, particularly regarding pressure and general discomfort. These qualitative and quantitative insights highlight that user experience does not always align with measures of device effectiveness: a headset can be highly accurate yet cause discomfort or dissatisfaction, potentially limiting its practical usability. This points to the crucial role of user feedback in the optimization of BCI systems, to ensure that devices are not only effective but also acceptable and comfortable for daily use. To promote device acceptance, encourage frequent use in daily life, and prevent withdrawal, it is essential to consider both system effectiveness and user satisfaction during implementation. Adopting a BCI device that may sacrifice a modest amount of accuracy -still remaining above the communication threshold - to provide superior comfort and overall satisfaction, may ultimately result in higher usage rates compared to a highly effective device that is uncomfortable or poorly received by users.

The present study confirms previous literature indicating a user preference for EEG headsets that do not use conductive gel. The study by Hinrichs and colleagues highlights a clear user preference for EEG systems that do not use conductive gel compared to gel-based systems: 74% of participants preferred the gel-free system, whereas only 10% favored the gel-based one [9]. Even during longer recording sessions (up to 60 minutes), the gel-free system was rated as more comfortable [9]. The main reasons for this preference included the absence of conductive gel, which eliminates the need to wash the hair after use, the absence of a chin strap, which was perceived as uncomfortable in the gel-based system, and greater freedom of movement compared to wired setups. In addition, the gel-free system was positively evaluated in terms of usability, with most participants reporting that they would be able to wear the headset independently without technical assistance [9].

In any case, what appears to most strongly drive user preferences is the visual appearance of the device. Gel-based EEG caps were often rated negatively, being perceived as unattractive and overly conspicuous [11, 13]. Discreet designs are generally preferred, as they are better able to accommodate individual preferences and avoid drawing excessive attention in public settings [11, 13, 15].

All of these factors should be carefully considered when implementing a BCI device to ensure both effective performance and user acceptance.

CONCLUSION

The present study demonstrates that a semi-dry headset, designed for out-of-lab use, is significantly preferred over a gel-based research cap in terms of comfort, weight, aesthetics, and reduced physical discomfort. These results suggest that consideration of user-centered factors, including ergonomic and perceptual aspects (while maintaining a focus on effectiveness), is essential for promoting the practical adoption and sustained use of BCIs for individuals with severe motor disabilities.

ACKNOWLEDGMENT

This work was partially supported by the Italian Ministry of Health (PNRR POC-2023- 12377627) and by the ERANET Neuron JTC2024 - Brain-Body (REWIRING).

REFERENCES

- [1] Pitt KM, Brumberg JS, Pitt AR. Considering augmentative and alternative communication research for brain-computer interface practice. Assistive technology outcomes and benefits. 2019;13(1):1-20
- [2] Riccio A, Pichiorri F, Schettini F, Toppi J, Risetti M, Formisano R, et al. Interfacing brain with computer to improve communication and rehabilitation after brain damage. In: Mattia D, Astolfi L, Cincotti F (Eds.). Progress in Brain Research. Elsevier, Amsterdam 2016, pp. 357-387
- [3] Andrich R, Mathiassen NE, Hoogerwerf EJ, Gelderblom GJ. Service delivery systems for assistive technology in europe: An aaate/eastin position paper. Technology and Disability. 2013;25(3):127-146.
- [4] Galiotta V, Riccio A, Pichiorri F, Mattia D, Astolfi L, Cincotti F, et al. P300-based brain-computer interface for communication in assistive technology centres: Influence of users' profile on BCI access. Journal of Neural Engineering. 2025;22(3):036044.
- [5] Wolpaw JR, Millán JdR, Ramsey NF. Brain-computer interfaces: Definitions and principles. In: Birbaumer N, Cohen LG (Eds.). Handbook of Clinical Neurology. Elsevier, Amsterdam 2020, pp. 15-23.
- [6] Pitt KM, Brumberg JS, Burnison JD, Mehta J, Kidwai J. Behind the scenes of noninvasive brain-computer interfaces: A review of electroencephalography signals, how they are recorded, and why they matter. *Perspect ASHA Spec Interest Groups*. 2019;4(6):1622-1636.
- [7] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and clinical Neurophysiology. 1988;70(6):510-523
- [8] ISO 9241-210:2019. Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. International Organization for Standardization; 2019.
- [9] Hinrichs H, Scholz M, Baum AK, Kam JWY, Knight RT, Heinze HJ. Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG system for clinical applications. Scientific Reports. 2020;10(1):5218.
- [10] Wang F, Li G, Chen J, Duan Y, Zhang D. Novel semi-dry electrodes for brain-computer interface applications. J. Neural Eng. 2016;13(4):046021
- [11] Nijboer F, van Laar B, Gerritsen S, Nijholt A, Poel M. Usability of three electroencephalogram headsets for brain-computer interfaces: A within-subject comparison. Interact Comput. 2015;27:500-511.
- [12] Oliveira AS, Schlink BR, Hairston WD, König P, Ferris DP. Proposing metrics for benchmarking novel EEG technologies towards real-world measurements. Frontiers in Human Neuroscience. 2016;10:188
- [13] Radüntz T, Meffert B. User experience of 7 mobile electroencephalography devices: Comparative study. JMIR Mhealth Uhealth. 2019;7(9):e14474.
- [14] Pinegger A, Hiebel H, Wriessnegger SC, Müller-Putz GR. Composing only by thought: Novel application of the P300 brain-computer interface. PLoS One. 2017;12(9):e0181584.
- [15] Cardoso ASS, Andreassen Struijk LNS, Kaeseler RL, Jochumsen M. Comparing the usability of alternative EEG devices to traditional electrode caps for SSVEP-BCI controlled assistive robots, in Proc. IEEE Int. Conf. Rehabil. Robot. (ICORR), Rotterdam, Netherlands, 2022, 1-6
- [16] Radüntz T. Signal Quality Evaluation of Emerging EEG Devices. Front. Physiol. 2018;9:98.
- [17] Krusienski DJ et al. A comparison of classification techniques for the p300 speller. Journal of neural engineering. 2006;3(4):299-305
- [18] Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR. Bci2000: A general-purpose brain-computer interface (bci) system. IEEE Transactions on biomedical engineering. 2004;51(6):1034-1043.
- [19] Zickler C et al. A brain-computer interface as input channel for a standard assistive technology software. Clinical EEG and neuroscience. 2011;42(4):236-244.
- [20] Berger LM, Wood G, Kober SE. User experience in neurofeedback applications using AR as feedback modality. Computers. 2024;13(5):110.
- [21] Kübler A, Neumann N, Wilhelm B, Hinterberger T, Birbaumer N. Predictability of brain-computer communication. Journal of Psychophysiology. 2004;18(2/3):121-129

First real-time study of a BCI based on quasi-movements: Comparison to the MI BCI within a new hybrid eye–brain–computer interface paradigm

Y.G. Shevtsova^{1,2}, A.S. Yashin¹, E.P. Svirin¹, A.N. Vasilyev^{1,2}, S.L. Shishkin¹

¹ Moscow State University of Psychology and Education, Moscow, Russia

² M.V. Lomonosov Moscow State University, Moscow, Russia

E-mail: shevtsova.jg@gmail.com

ABSTRACT: We report the first real-time study of a brain–computer interface (BCI) based on quasi-movements (QM) and compare it with standard motor imagery (MI) within a novel hybrid eye–brain–computer interface (EBCI). Eighteen healthy participants used gaze to select on-screen objects and then activated EEG-based control to interact with a game requiring precise initiation and termination of sensorimotor-task execution. Online classification performance was comparable across conditions: mean balanced accuracy was 74.25% for QM and 74.63% for MI, with 13 of 18 participants exceeding 70% in each condition. No significant differences were found between QM and MI in classifier performance, game interaction measures, or subjective ratings. Nevertheless, marked interindividual variability was observed, with some participants clearly favouring one condition over the other. These findings demonstrate that QM can support stable online BCI control in a demanding hybrid gaze–EEG paradigm and may serve as a practical alternative to MI, potentially broadening accessibility of active sensorimotor BCIs.

INTRODUCTION

Kinesthetic motor imagery (MI), i.e., mental simulation of kinesthetic sensations associated with certain movements, is an especially popular method to issue a command via non-invasive brain-computer interfaces (BCIs). This mental task is associated with changes in EEG sensorimotor rhythms roughly similar to what is observed under real movements. However, MI BCIs are slow and often misinterpret the user's intentions. Some studies demonstrated improved BCI classification performance [1] and clinical outcome of BCI-based rehabilitation [2] in case of using an alternative to MI, attempted movements (AM). AM involve the effort to produce actual physical movement and elicit changes in the EEG with a pattern similar to changes during MI execution but not result in such movement due to paralysis.

Unfortunately, studying AM in healthy subjects is complicated. Limb constraint (e.g., [3]) is associated with strongly increased sensory input, which may significantly alter the effects under study. Blokland et al. in their study of AM [1] used neuromuscular block to create temporal paralysis, an approach which cannot be widely used. Experimentation in paralysed and amputee

participants is limited by ethical issues.

Nikulin et al. [4] discovered a new motor phenomenon, quasi-movements (QM), and proposed to use it as a model for AM in healthy participants. QM are overt movements minimized to the extent that no muscle activation is detectable [4]. Similarly to attempted movements, QM elicit stronger event-related desynchronization (ERD) of sensorimotor rhythms compared to MI [4,5]. Previous studies, however, did not explore real-time control using QM.

We suggested previously that, because AM and QM are closer to real movements than to MI in the sense of intention [7], they could potentially be more compatible than MI with tasks requiring attention to outside world, such as receiving visual feedback from a BCI and, especially, sending commands to a computer using eye gaze [6,8]. So far, tight hybridization of noninvasive MI BCIs with gaze-based control (e.g., when the user has to do MI during intentional gaze fixation) has not been very successful [9,10]. Designs where gaze direction has been used to select an object for subsequent interaction via MI BCI seem more effective, although still prone to frequent errors [11, 12].

In the current study we, for the first time, explored real-time QM BCI control. Participants had to activate the BCI immediately after selecting an object on a screen using their gaze and stop performing QM or MI tasks at given moments as precisely in time as possible, based on visual feedback from the BCI. The need to use gaze intentionally for control and then to closely monitor visual feedback from the BCI had to increase external attentional focus. We expected that in such a hybrid eye–brain–computer interface (EBCI) paradigm QM would perform better than MI due to their hypothetically better compatibility with a task requiring external focus of attention (gaze-based control), and, as a result, better EBCI performance, as well as subjectively more successful and comfortable control, would be observed under the use of QM compared to MI.

Preliminary results from this study, with a focus on gaze data, have been reported at the ACM ETRA conference [13].

MATERIALS AND METHODS

Participants: 18 healthy volunteers (13 females and 5

males, aged 19–41 years; median age 24) participated in the study. Written informed consent was obtained from all of them.

All participants had prior experience with motor imagery (MI), and nine had experience with quasi-movements (QM) as they had taken part in one of our earlier studies. Experienced participants were intentionally recruited to ensure reliable contralateral event-related desynchronization (ERD) during sensorimotor tasks essential for BCI control.

Signal acquisition: EEG (64 electrodes) and EMG were recorded at 1000 Hz using an NVX136 DC EEG amplifier (Medical Computer Systems, Moscow, Russia). The ground electrode was placed at Fpz. Electrode impedance was kept below 20 k Ω . EMG electrodes were positioned on the participant's right palm of a participant: one over the belly of the *abductor pollicis brevis* muscle and the other near the proximal base of the phalanx.

Gaze data were acquired through the EyeLink 1000 Plus eye tracker (SR Research, Canada) in remote mode (without a chin rest) at 500 Hz from the dominant eye, determined using the Miles and Porta tests. Eye tracker calibration was performed every 10 minutes (before each second game), ensuring validation error not exceed 1.0°.

Software: Signal acquisition and real-time processing were performed using the Resonance platform. The EBCI game was implemented in Unity 3D game engine, gaze data were processed online in Python, offline classification was conducted in MATLAB (MathWorks, USA), and surveys were administered using PsychoPy.

Offline Data Processing: The EEG classification pipeline followed [14]. Individual μ -rhythm frequency bands were identified by computing Common Spatial Pattern (CSP) filters in overlapping sub-bands (7–9 to 14–16 Hz) and manually selecting the band showing contralateral and/or ipsilateral ERD. EEG signals were then spatially filtered using two CSP components (typically those with the smallest eigenvalues). Spectral features were extracted using a short-time Fourier transform (1-s window, 90% overlap), and Fisher's Linear Discriminant with 5% covariance regularization was applied to obtain optimal frequency-bin combinations for class separation. Feature vectors consisted of power values from the selected spatial components within the individual frequency band. A Naïve Bayes classifier was trained on these features and applied online.

Online Data Processing: For real-time control, features were extracted from the ongoing EEG signal, and the trained classifier was applied every 100 ms to estimate the probability of sensorimotor task execution over the preceding 1-s window. To reduce the impact of classifier errors, these outputs were aggregated into one decision per second: a positive decision (task detected) was made if more than half of the ten consecutive classifier outputs exceeded 0.5.

EBCI algorithm: Gaze was used to select objects on

the screen, triggering the BCI-controlled phase. Classifier decisions then generated on-screen feedback. At the end of each trial, BCI control was blocked until a new object was selected. Thus, the gaze and BCI components operated sequentially and complementarily.

Experiment design: Two conditions corresponding to sensorimotor tasks (MI or QM) were presented in random order. Each condition comprised three stages: training, calibration, and EBCI interaction.

During training, participants were instructed how to perform MI or QM. EEG data were then recorded while participants performed the sensorimotor and counting tasks during the calibration session to obtain two classes of data for training an individual classifier. The resulting model enabled BCI control in the game AstroSync, which was played three times by each participant during the EBCI interaction session. After the calibration session, after the first and third games in each condition, and at the end of both conditions, participants completed questionnaires assessing their user experience.

We further describe training, calibration, and EBCI interaction stages in reverse order, as it helps to clarify the relations between them.

EBCI interaction session: Participants played AstroSync, a game designed in this study for testing combined gaze- and BCI-based control. The game field consisted of 10 constellations, each comprising a central star surrounded by multiple mini-stars (Fig. 1a). Maintaining gaze fixation on a constellation for 0.5 s caused it to be outlined by a circular frame; after an additional 1.5 s of continuous fixation, the central star turned yellow (Fig. 1b), signaling that it is time to start interaction via the BCI.

During this period, one of eight concentric zones, invisible to the player, was activated every second from the center towards the periphery. Mini-stars within the active zone were illuminated in green and awarded points when the EEG classifier detected execution of the sensorimotor task; otherwise, they were illuminated in red and points were deducted (Fig. 1c).

If the sensorimotor task execution was detected before the center star turned yellow, the player was prompted to relax: the constellation disappeared and a distractor counting task was presented (Fig. 1d). In such cases, the participants were required to enumerate mini-stars within the triangles and squares. Failure to terminate the sensorimotor task after completion of the circular sequence resulted in point deductions per outer zones (up to four zones).

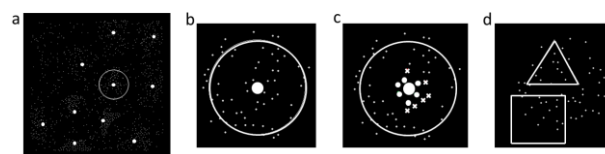


Figure 1: AstroSync gameplay features. (a) Star selection, (b) interaction with a constellation, (c) blocked interaction.

Calibration session: Each calibration trial followed the

sequence of phases in the EBCI interaction session's trials; however, as no classifier was available at this stage, all interactions were predefined as successful. No interaction was blocked, all mini-stars were illuminated in green, zone by zone, and the outer zones were not activated. The session comprised 40 trials presented in random order: 20 sensorimotor-task and 20 counting-task trials.

Training: Prior to the calibration session participants were trained approximately 15 minutes per condition. First, they were familiarized with the target action, thumb abduction. This movement was chosen following previous studies [4-8], because the muscle activity related to it could be readily monitored using EMG to verify the absence of muscle activation in the QM condition. For this purpose, the EMG signal was monitored online using Teager-Kaiser Energy Operator (TKEO), which highlighted bursts in the signal.

QM training was adapted from the standard protocol described in the original QM study [4] and used in our previous work. Participants watched their transformed EMG at a computer screen and were instructed to gradually reduce the amplitude of right thumb abduction until the visualized signal approached noise level. They were then guided to transit to QM. Unlike in previous studies, we explicitly defined QM for the participants as motor attempts that produce no visible muscle contractions. Also specific to this study, participants were instructed to focus on generating an "impulse" or motor "command" to the thumb, rather than on muscle sensations or somatosensory feedback.

Although all participants were familiar with the MI task, they received brief instructions on kinesthetic MI. First, they performed full-amplitude thumb abductions while attending to the associated sensations, such as the feeling of movement, muscle tension, and a sense of heaviness or warmth in the hand. They were then instructed to mentally rehearse these movements while keeping the hand completely relaxed.

Performance evaluation: Online EEG classifier performance was evaluated using balanced accuracy to account for class imbalance. In addition to overall performance, we examined how accurately the classifier detected task execution. The true positive rate (TPR) reflects how often the classifier correctly identified periods when participants were performing the sensorimotor task, whereas the true negative rate (TNR) reflects how often it correctly identified periods when the task was not being performed.

BCI interaction performance was quantified using several measures. The primary metric was the percentage of successfully completed constellations, with success defined as achieving >50% correct classifier predictions per constellation. Because the classifier is sensitive to the timing of task execution, additional metrics assessed participants' ability to start and stop tasks appropriately: the proportion of constellations activated without blocking, and a measure of how long participants remained in the active state after completing a task.

Surveys: After completing both conditions, participants rated how effective and comfortable each was, using scales anchored by QM on the left and MI on the right. They also reported how often they were able to interact with the game without consciously attending to the sensorimotor task, separately for QM and MI, using scales anchored by "never" on the left and "very frequently" on the right. Participants also rated, after each game in the QM condition, how much they attended to preventing quasi-movements from turning into actual movements, using a scale anchored by "constantly monitored" on the left and "did not monitor at all" on the right.

Statistics: To analyze classifier performance and game metrics, we used mixed-effects models with Geisser-Greenhouse correction, assessing the main effects of condition and game, as well as their interaction. Questionnaire data were analyzed using t-tests for the automation scales and one-way ANOVA for amplitude-tracking ratings.

RESULTS

EMG monitoring: Mean RMS EMG levels remained low and comparable between the QM and MI conditions during both calibration ($1.91 \pm 0.8 \mu\text{V}$ for MI; $2.04 \pm 0.78 \mu\text{V}$ for QM) and game sessions ($2.49 \pm 1.77 \mu\text{V}$ for MI; $2.37 \pm 1.27 \mu\text{V}$ for QM), indicating no systematic increase in muscle activity in the QM condition compared to MI.

Classifier performance evaluation: Fig. 2 summarizes online EEG classifier performance across conditions and games. Mean balanced accuracy was $74.63 \pm 6.52\%$ for the MI condition and $74.25 \pm 6.73\%$ for the QM condition. Notably, 13 out of 18 participants achieved balanced accuracy above 70% in each condition, indicating a confident level of interface control.

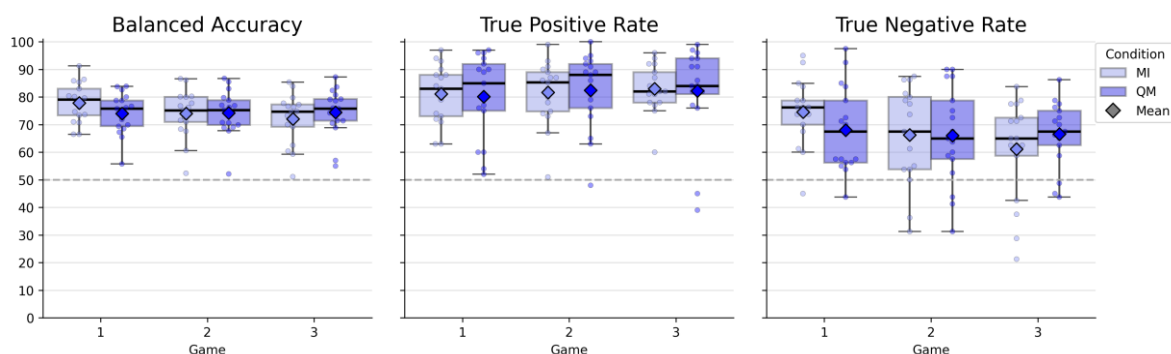


Figure 2: Performance of the online EEG classifier. Balanced accuracy, true positive rate (TPR), and true negative rate (TNR) in the MI and QM conditions across games.

Analysis of TPR and TNR indicated that both task execution and non-execution periods were correctly classified in more than half of the trials. Mean TPR was $81.94 \pm 8.26\%$ for MI and $81.61 \pm 12.25\%$ for QM, while mean TNR was $67.32 \pm 11.21\%$ for MI and $66.89 \pm 11.72\%$ for QM. However, in the MI condition, two participants had TNR below 50%, and in the QM condition, one participant had TNR below 50%, suggesting that the classifier occasionally failed to correctly identify periods without task execution for some users.

A mixed-effects model with the Geisser–Greenhouse correction revealed no significant main effects of condition ($F(1, 32) = 0.026, p = .871$) or game ($F(1.78, 56.57) = 2.136$) on balanced accuracy. The interaction between condition and game approached significance ($F(2, 64) = 2.713, p = .074$). For TPR, there were no significant effects of condition ($F(1, 32) = 0.008, p = .928$), game ($F(1.601, 51.24) = 0.440, p = .603$), or their interaction ($F(2, 64) = 0.109, p = .897$). For TNR, no significant main effect of condition was observed ($F(1, 32) = 0.011, p = .916$), nor was there a significant

condition $\times$ game interaction ($F(2, 64) = 1.965, p = .148$). However, there was a marginal effect of game ($F(1.956, 62.58) = 3.123, p = .052$), suggesting a possible influence of game progression. Exploratory post-hoc comparisons indicated a difference between Game 1 and Game 3 in the IM condition.

Interaction evaluation: Game interaction measures are summarized in Fig. 3. For the percentage of successfully completed constellations, no significant effects were found for condition ($F(1, 32) = 0.017, p = .898$), game ($F(1.804, 57.52) = 0.246, p = .760$), or their interaction ($F(2, 64) = 1.469, p = .238$). Similarly, the proportion of constellations activated without blocking showed no significant effects of condition ($F(1, 32) = 0.047, p = .829$), game ($F(1.910, 61.13) = 0.732, p = .483$), or interaction ($F(2, 64) = 0.036, p = .965$). For the mean number of activated outer zones, no significant effects were observed for condition ($F(1, 32) = 0.073, p = .788$) or the interaction between condition and game ($F(2, 64) = 2.145, p = .125$). There was a marginal effect of game ($F(1.817, 58.13) = 2.707, p = .080$), suggesting a possible trend across games.

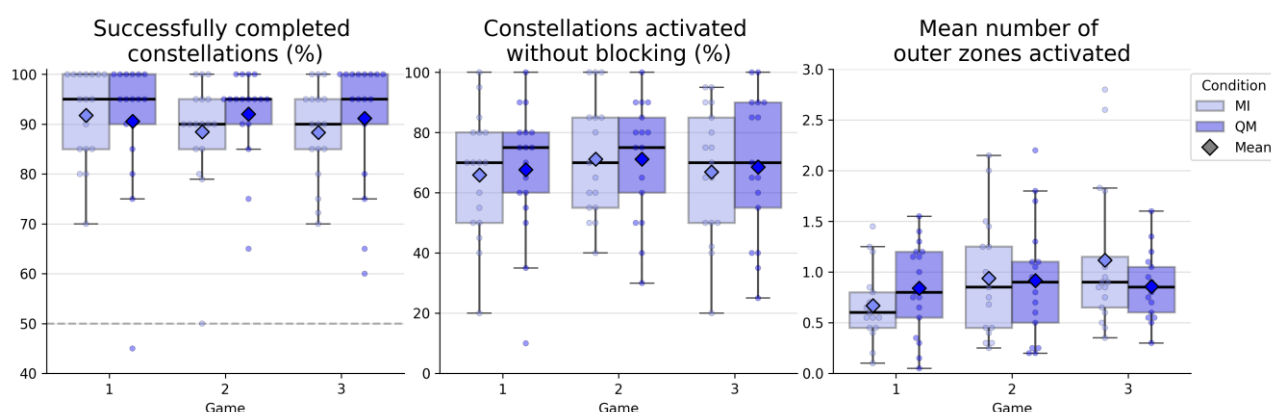


Figure 3: Game interaction performance in the MI and QM conditions across games.

Surveys: Ratings of perceived control effectiveness and comfort showed a bimodal distribution, with participants clearly favoring either QM or MI; none rated the conditions near the midpoint (Fig. 4). This pattern emphasizes strong individual differences in subjective preference, consistent with variability

observed in objective performance.

According to a one-way ANOVA, participants reported paying progressively less attention to the amplitude of QM as the games progressed (Fig. 3c), $F(2, 20) = 4.755, p = .0496$. Post-hoc comparisons indicated a significant difference between the “before” and “after” conditions

($p = .001$).

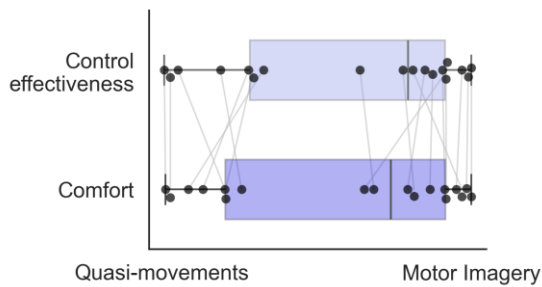


Figure 4: Final questionnaire results comparing the two conditions (MI and QM). Lines link responses from the same participant across conditions.

Ratings of control automatization revealed that participants were rarely able to interact with the constellation without focusing on the sensorimotor task in the QM condition ($t(17) = -2.598$, $p = .019$), whereas no significant deviation was observed in the MI condition ($t(17) = -1.584$, $p = .132$).

DISCUSSION

This study provides the first demonstration of online control using quasi-movements (QM) in an active brain-computer interface (BCI). Importantly, this control was implemented in a hybrid eye-brain-computer interface (EBCI) paradigm, in which participants had to combine gaze-based interaction with sensorimotor BCI control while continuously monitoring visual feedback. Despite this increased task complexity and the need to maintain attention to external stimuli, participants achieved stable BCI control.

Our results indicate that QM can support performance comparable to that of the standard motor imagery (MI) paradigm. Balanced accuracy remained consistently above chance, and many participants achieved reliable levels of BCI control in both conditions. Thus, QM appears to be a viable alternative control strategy for active BCI systems. Contrary to our initial hypothesis, however, the more externally oriented QM task did not outperform MI in either objective performance or subjective ratings. No significant differences between conditions were observed on any of the evaluated metrics. At the same time, substantial interindividual variability was evident: some participants performed markedly better in one condition than in the other. This pattern suggests that QM may be particularly useful for a subset of users, especially those who struggle with conventional MI-based control, a challenge often discussed in terms of BCI illiteracy.

Notably, this comparable performance was observed despite substantially lower prior familiarity with QM. All participants had previous experience with MI, whereas only about half had prior exposure to QM, and the revised instructions required even experienced participants to partially relearn the task. This finding suggests that QM can be learned sufficiently for online control even with limited prior training in healthy

participants, which supports its potential applicability in broader user populations.

QM has previously been proposed to offer practical advantages, including simpler training requirements [4] and a closer resemblance to natural movement [7]. However, in the present study only 7 of 18 participants reported preferring QM, while many described it as relatively difficult. One possible reason is the lack of somatosensory feedback from actual muscle activity, which may have reduced the subjective sense of control during QM. Visual feedback alone may not have fully compensated for the absence of proprioceptive cues, particularly given its limited congruence with the task and the approximately one-second delay imposed by technical constraints. This, in turn, underscores the importance of feedback design and temporal responsiveness for effective interaction with QM-based interfaces.

Taken together, these findings suggest that QM is a promising alternative control paradigm for sensorimotor BCI systems. By offering a control strategy distinct from traditional MI while achieving comparable performance, QM may help broaden the range of users who can effectively operate brain-computer interfaces.

CONCLUSION

This study demonstrates that quasi-movements (QM) can support stable online control in a brain-computer interface, achieving performance comparable to the standard motor imagery (MI) paradigm. The absence of systematic differences between conditions, together with strong interindividual variability, suggests that QM may serve as a useful alternative control strategy for some users, particularly those who struggle with MI. Overall, incorporating QM alongside traditional approaches may help broaden the range of available sensorimotor BCI control strategies.

ACKNOWLEDGMENTS

This study was supported by the Russian Science Foundation under grant no. 22-19-00528.

REFERENCES

- [1] Blokland Y, Spyrou L, Lerou J, Mourisse J, Scheffer GJ, Van Geffen GJ, Farquhar J, Bruhn J. Detection of attempted movement from the EEG during neuromuscular block: proof of principle study in awake volunteers. *Sci. Rep.* 2015;5: 12815
- [2] Mansour S, Ang KK, Nair KPS, Phua KS, Arvaneh M. Efficacy of Brain-Computer Interface and the impact of its design characteristics on poststroke upper-limb rehabilitation: a systematic review and meta-analysis of randomized controlled trials. *Clin. EEG Neurosci.* 2022;53(1): 79-90
- [3] Mondini V, Sburlea A-I, Müller-Putz GR. Towards unlocking motor control in spinal cord injured by applying an online EEG-based framework to decode

motor intention, trajectory and error processing. *Sci. Rep.* 2024;14(1): 4714

[4] Nikulin VV, Hohlefeld FU, Jacobs AM, Curio G. Quasi-movements: a novel motor-cognitive phenomenon. *Neuropsychologia* 2008;46(2): 727-742

[5] Vasilyev AN, Yashin AS, Shishkin SL. Quasi-movements and “quasi-quasi-movements”: does residual muscle activation matter? *Life* 2023;13(2): 303

[6] Yashin AS, Vasilyev AN, Shevtsova YG, Shishkin SL. Can quasi-movements be used as a model of the BCI based on attempted movements?, in *Proc. 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Sarawak, Malaysia, October 6-10, 2024, pp. 2028-2033

[7] Yashin AS. A challenge for bringing a BCI closer to motor control: the “interface uncanny valley” hypothesis, in *Proc. 2023 IEEE Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB)*, Novosibirsk, Russian Federation, September 28-30, 2023, pp. 242-247

[8] Shishkin SL, Yashin AS, Shevtsova YG, Vasilyev AN. Should attempted movements replace motor imagery in BCI? The issue of compatibility with gaze use, in *Proc. 9th Graz Brain-Computer Interface Conference 2024*, Verlag der Technischen Universität Graz, pp. 455-460

[9] Zander TO, Gaertner M, Kothe C, Vilimek R. Combining eye gaze input with a Brain-Computer Interface for touchless human-computer interaction. *Int. J. Hum.-Comput. Interact.* 2011;27(1): 38-51

[10] Hou BJ, Bekgaard P, MacKenzie S, Hansen JPP, Puthusserypady S. GIMIS: gaze input with motor imagery selection, in *Proc. Eye Tracking Research and Applications Symposium (ETRA)*, 2020, Article 18, pp. 1-10

[11] Frisoli A, Loconsole C, Leonardis D, Banno F, Barsotti M, Chisari C, Bergamasco M. A new gaze-BCI-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* 2012;42(6): 1169-1179

[12] Soekadar SR, Birbaumer N, Slutzky MW, Cohen LG. Brain-machine interfaces in neurorehabilitation of stroke. *Neurobiol. Dis.* 2015;83:172–179

[13] Yashin AS, Shevtsova YG, Svirin EP, Vasilyev AN, Shishkin SL. Combining intuitive gaze-based control with EEG-based detection of motor imagery and quasi-movements, in *Proc. 2025 ACM Symposium on Eye Tracking Research & Applications (ETRA)*, Article 38, pp. 1-3

[14] Vasilyev AN, Nuzhdin YO, Kaplan AY. Does real-time feedback affect sensorimotor EEG patterns in routine motor imagery practice? *Brain Sci.* 2021;11(9):1234

Contrast-dependent decoding performance in classic and grating c-VEP stimuli

Sena Er¹, Donja Schipper¹, Yuzhen Qin¹, Jordy Thielen¹, Michael Tangermann¹

¹Machine Learning & Neural Computing department, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

E-mail: sena.er@donders.ru.nl

ABSTRACT: Designing effective visual brain-computer interfaces (BCIs) requires careful selection of stimulus parameters that shape the strength and reliability of evoked neural responses. Stimulus contrast is one such parameter in code-modulated visual evoked potential (c-VEP) BCIs, yet its effects on decoding performance and reliability remain incompletely characterized. Here we investigate how stimulus contrast influences decoding performance in a c-VEP speller using two stimulus families: classic luminance flicker and grating-based textured stimulation. Four participants were tested across six contrast levels (10–100 %). Trial-level performance was quantified using a Fisher-transformed margin metric and modeled with Gaussian processes. In this exploratory dataset, performance generally increased with contrast, with larger decoding margins for the classic stimulus than for the grating stimulus. The magnitude and shape of these effects varied across participants, while estimated trial-to-trial noise changed only modestly across contrast levels. These findings suggest that trial-level Gaussian-process modeling can support exploratory characterization of stimulus-parameter effects in c-VEP BCIs.

INTRODUCTION

Designing an effective brain-computer interface (BCI) requires decisions not only about decoding methods, but also about how neural responses are elicited. In code-modulated visual evoked potential (c-VEP) BCIs, the design of stimuli shapes the evoked brain response and therefore the resulting decoding performance [1].

One important stimulus parameter is contrast. Across psychophysics and visual physiology, contrast is known to shape the strength of early visual responses, with neural activity in visual cortex increasing systematically with stimulus contrast over much of the tested range [2]. Human electrophysiology shows a similar pattern: visual evoked responses measured with EEG/VEP paradigms vary reliably with contrast, indicating that stronger contrast can produce stronger and more discriminable visually evoked activity [3–5]. In c-VEP BCIs, this suggests that increasing contrast could improve decoding performance by strengthening the stimulus-evoked response, although the magnitude of this benefit may vary across

participants and may need to be balanced against practical considerations such as visual comfort and usability [6–8], which may be higher for lower contrast levels.

However, contrast does not act independently of the stimulus design. In addition to conventional spatially uniform luminance flicker, recent work has considered textured stimulation patterns that introduce spatial structure into the stimulus [7]. In the present study, we compare a *classic* condition, consisting of spatially uniform luminance modulation around a mid-gray baseline, with a *grating* condition, in which the active state contains a textured pattern composed of multiple Gabor-like patches with different orientations. In the grating condition, modulation is therefore defined by the appearance of oriented spatial texture around a mid-gray baseline rather than a purely global luminance change. Consequently, it could be plausible that contrast effects may differ between the two stimulus families.

Most previous evaluations of c-VEP stimulation settings have focused on aggregate outcomes such as decoding accuracy or subjective ratings [1, 6–8]. While such measures are important, they do not fully capture how reliably a given stimulus setting produces usable responses across trials. For practical BCI design, a good parameter setting should not only yield high average performance, but should also produce strong trial-level performance with limited uncertainty and trial-to-trial variability [9, 10]. Therefore, it is important to characterize the contrast space not only in terms of performance but also in terms of trial-to-trial variability and uncertainty.

The present study examines how stimulus contrast affects c-VEP decoding in both classic and grating conditions using a trial-level performance metric derived from the class-wise trial scores. Rather than considering only average decoding accuracy, we model how this metric changes across contrast levels and quantify the associated uncertainty and trial-to-trial variability. To this end, we use Gaussian process regression to characterize subject- and condition-specific contrast-response functions. Specifically, we test whether trial-level performance generally increases with contrast, whether this effect differs between classic and grating stimuli, and whether variability remains approximately stable across contrast levels or differs across subjects and conditions. We expect higher contrast to generally yield stronger trial-level performance, although the shape of

the contrast-response function may vary with stimulus type and individual differences.

METHODS

Dataset and experimental design: Five healthy participants took part in the study (age 21–56 years, mean 34 years; 4 female). All had normal or corrected-to-normal vision and no history of neurological disorders. The study was approved by the Ethics Committee of the Social Sciences of the Radboud University, and written informed consent was obtained from all participants. One participant was excluded because of insufficient EEG data quality, leaving four participants for analysis.

The dataset comprised a visual speller experiment with two stimulation conditions: a *classic* condition and a *grating* condition. In the classic condition, stimulation consisted of spatially uniform luminance modulation around a mid-gray baseline. In the grating condition, the active state ($bit = 1$) consisted of a textured pattern composed of multiple Gabor-like patches with different orientations, producing a spatially structured stimulus rather than a uniform luminance field. In both conditions, the stimulus modulation was centered around the mean luminance of the display (see Figure 1).

Visual stimulation was presented on a 120 Hz display. Stimuli were presented in a c-VEP speller with 32 symbols arranged in a 4×8 grid. Each symbol in the speller was assigned a unique code derived from a 63-bit m -sequence, with circular shifts of 2 bits used to generate 32 distinct target codes. The codes were presented at a 60 Hz presentation rate for a cycle duration of 1.05 s.

The stimulus contrast parameter was defined relative to the mean luminance of the display. A contrast of 100% corresponds to a modulation of $\pm 50\%$ around the mean luminance, whereas 40% contrast corresponds to $\pm 20\%$ of modulation in both directions. Thus, across all conditions, the temporal modulation of each target was centered around the same mean luminance, such that the average brightness remained constant and only the modulation amplitude varied.

Six contrast levels, $c \in C = \{10, 20, 30, 40, 60, 100\}$, were tested in both classic and grating conditions, yielding 12 stimulus conditions in total. For each condition, 32 trials of 4.2 s duration (four m -sequence cycles) were collected, one for each target. The order of trials and conditions was fully randomized. Before stimulation, the target was cued to the participant by highlighting its background in green for 800 ms. The 384 trials collected per participant in a single session were divided into eight runs, between which self-paced breaks could be taken.

EEG acquisition and preprocessing:

EEG was recorded with a 64-channel BioSemi ActiveTwo system. Preprocessing was implemented in Python using MNE-Python (version 1.11.0) [11]. Continuous EEG was band-pass filtered from 6 to 21 Hz using MNE’s default zero-phase FIR filter with a Hamming window and `firwin` design; the filter length was au-

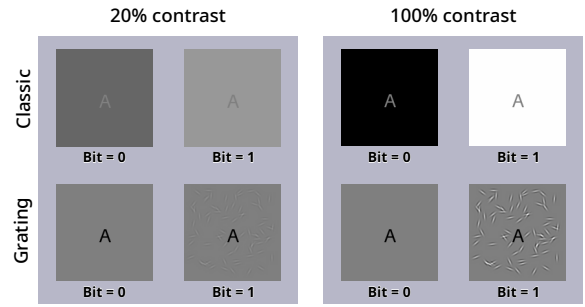


Figure 1: **Example stimuli used in the experiment.** Two contrast levels (20% and 100%) are shown. In the classic condition the stimulus consists of spatially uniform luminance flicker. In the grating condition the flicker is modulated by Gabor grating patterns.

tomatically determined from the transition bandwidths. Epochs were extracted from -0.5 to 4.7 s relative to trial onset, resampled to 120 Hz, and cropped to the analyzed 0 to 4.2 s stimulation interval. No baseline correction was applied.

Decoding: Decoding was performed offline, separately for each subject, stimulus condition, and contrast level, using a two-stage pipeline implemented in Python. Within each subject–condition–contrast combination, a 4-fold cross-validation procedure was used to estimate decoding performance.

Within each trial, overlapping encoding windows were extracted using an encoding length of 300 ms and a stride of $1/60$ s, that is, synchronized to the bit presentation rate. These epochs were first transformed using canonical correlation analysis (CCA) with one component and covariance regularization parameters γ_x and γ_y ($\gamma_x = \gamma_y = 0.1$) [12, 13].

The transformed epochs were subsequently vectorized and classified using a linear discriminant analysis (LDA) classifier with the *eigen* solver and Ledoit–Wolf shrinkage to regularize the covariance estimate, as implemented in scikit-learn (version 1.8.0) [14].

Epoch-level class probabilities were recombined across the trial and correlated with the tiled stimulus codebook to obtain class-wise trial scores ρ_{ij} , where ρ_{ij} denotes the score assigned to the target candidate symbol j in trial i [13]. The predicted symbol was $\hat{y}_i = \arg \max_j \rho_{ij}$.

Trial-level performance metrics:

Trial-level performance metrics were derived from the class-wise trial scores to quantify how strongly each trial supported the correct class. Let ρ_{i,y_i} denote the score of the true class for trial i . Two Fisher-transformed performance metrics were analyzed:

$$\begin{aligned} z_{\text{true},i} &= \text{atanh}(\rho_{i,y_i}), \\ z_{\text{margin},i} &= \text{atanh}(\rho_{i,y_i}) - \text{atanh}\left(\max_{j \neq y_i} \rho_{ij}\right). \end{aligned} \quad (1)$$

Both metrics are expressed on the Fisher-z scale, which stabilizes the variance of correlations and makes the values more suitable for regression and between-trial comparisons [15].

The two metrics capture complementary aspects of trial-level performance. z_{true} reflects the absolute strength of the match between the EEG response and the true stimulus code. In contrast, z_{margin} measures the separation between the true class and the strongest competing class, that is, the runner-up symbol, and therefore more directly reflects decision-relevant trial-level performance. As classification depends on this relative separation, z_{margin} was used as the primary target variable in the Gaussian process analysis.

Before GP modeling, we applied a shared trial-level outlier filter based on z_{margin} . Within each subject $\times$ condition $\times$ contrast group, trials with MAD-based robust z-scores exceeding 3.5 in absolute value were removed, where the robust z-score was computed relative to the group median and scaled by the median absolute deviation (MAD). The same exclusion mask was applied to all modeled targets.

Contrast-response modeling: The goal of the statistical analysis was to characterize how the trial-level performance metric varies with stimulus contrast. Contrast was modeled as $x_i = \log_{10}(c_i)$ and standardized as $\tilde{x}_i = (x_i - \mu_x)/\sigma_x$, consistent with psychophysics and visual-neuroscience work showing that contrast-response relationships are often more regular when expressed as a function of contrast than on the raw linear scale [2, 16]. For a given subject-condition pair, let

$$D = \{(\tilde{x}_i, y_i)\}_{i=1}^N$$

denote the retained trial-wise observations after outlier removal, where N is the number of retained trials and y_i is either $z_{\text{true},i}$ or $z_{\text{margin},i}$. The observed trial-level metric was modeled as

$$y_i = f(\tilde{x}_i) + \varepsilon_i, \quad (2)$$

where $f(\tilde{x})$ denotes the latent mean contrast-response function and ε_i is a zero-mean residual term capturing trial-to-trial deviations around that mean. Separate Gaussian process models were fit for each subject and stimulus condition.

Gaussian process contrast-response model: A Gaussian process prior was placed on the latent contrast-response function,

$$f(\tilde{x}) \sim \mathcal{GP}(0, k(\tilde{x}, \tilde{x}')), \quad (3)$$

where $k(\tilde{x}, \tilde{x}')$ is a squared-exponential covariance kernel,

$$k(\tilde{x}, \tilde{x}') = \eta^2 \exp\left(-\frac{(\tilde{x} - \tilde{x}')^2}{2\ell^2}\right), \quad (4)$$

with signal amplitude η and lengthscale ℓ . The kernel hyperparameters were assigned weakly regularizing positive priors,

$$\eta \sim \text{HalfNormal}(2), \quad \ell \sim \text{Gamma}(2, 1). \quad (5)$$

Posterior summaries were evaluated on a dense grid of 200 contrast values spanning the observed range. Figures

report the posterior mean together with 94 % highest-density intervals.

Two observation models were considered. First, a homoskedastic model:

$$y_i | f(\tilde{x}_i) \sim \mathcal{N}(f(\tilde{x}_i), \sigma^2), \quad \sigma \sim \text{HalfNormal}(1). \quad (6)$$

Second, a heteroskedastic model, in which the observation noise was allowed to differ across the discrete tested contrast levels. Let $c_i \in \mathcal{C}$ denote the raw contrast level associated with trial i . Then

$$y_i | f(\tilde{x}_i) \sim \mathcal{N}(f(\tilde{x}_i), \sigma_{c_i}^2), \quad \sigma_c \sim \text{HalfNormal}(1), \quad c \in \mathcal{C}. \quad (7)$$

Thus, the mean contrast-response relationship was modeled as a smooth function of standardized log-contrast $\tilde{x}_i$, whereas the observation noise was allowed to vary across the six discrete contrast conditions. Both models were fit for diagnostic purposes, but the main contrast-response curves reported in the results section are based on the homoskedastic model.

Posterior inference was performed in PyMC (version 5.28.0) using the No-U-Turn sampler with the NumPyro backend (version 0.20.0) [17]. Sampling used 1000 tuning iterations and 1000 posterior draws with a target acceptance probability of 0.95.

RESULTS

Trial-level performance as a function of stimulus contrast:

Figure 2 shows Gaussian process estimates of the trial-level performance metric z_{margin} as a function of stimulus contrast for each subject and stimulus condition. Across subjects, trial-level performance increased systematically with stimulus contrast. At low contrast levels (10–20 %), z_{margin} values were typically close to zero, indicating weak separation between the target and competing symbols. As contrast increased, the expected decoding margin increased monotonically, indicating more reliable decoding.

Analyses using z_{true} yielded the same qualitative conclusions, with fitted mean curves nearly identical to those for z_{margin} ($r \approx 0.995$ – 0.999). As z_{margin} directly reflects the separation between the true class and its strongest competitor, it was used as the primary metric in the figures and quantitative summaries, while analyses based on z_{true} yielded the same qualitative conclusions.

Homoskedastic Gaussian process fits revealed substantial variability between subjects and a clear interaction with the stimulus condition. In the classic condition, three of four subjects already showed positive trial-level performance at 10 % contrast. Across subjects, the mean posterior estimate increased from 0.08 at 10 % contrast to 0.26 at 100 % contrast. Subject 4 showed the strongest classic response, rising from 0.20 to 0.48 across the tested

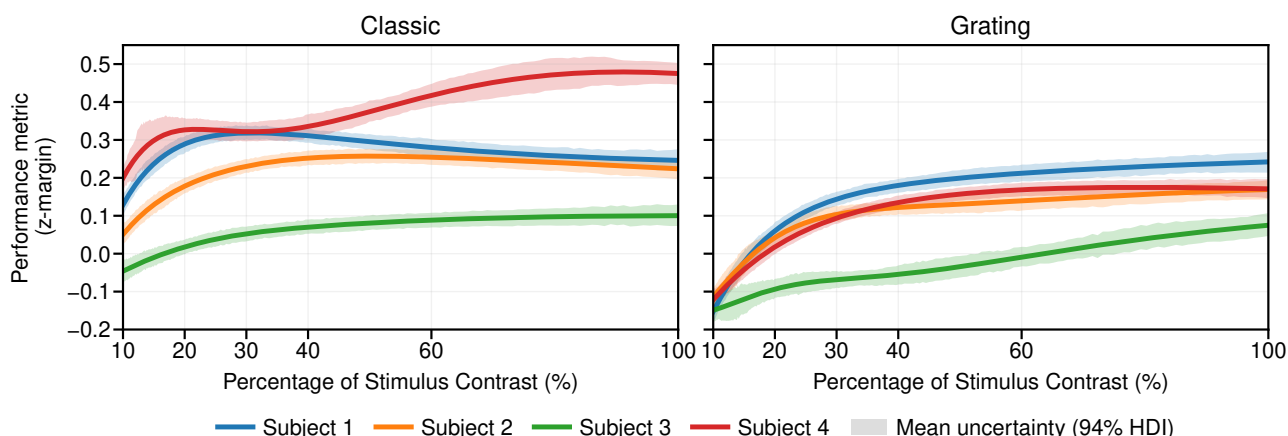


Figure 2: **Subject-wise trial-level performance curves as a function of contrast.** Gaussian process estimates of the trial-level performance metric z_{margin} are shown for each subject in the classic stimulus condition (left) and the grating stimulus condition (right). Higher values indicate stronger separation between the true target and competing targets. Solid lines show the posterior mean of the homoskedastic Gaussian process estimate of the expected trial-level performance as a function of contrast, and shaded regions indicate the 94 % highest-density interval.

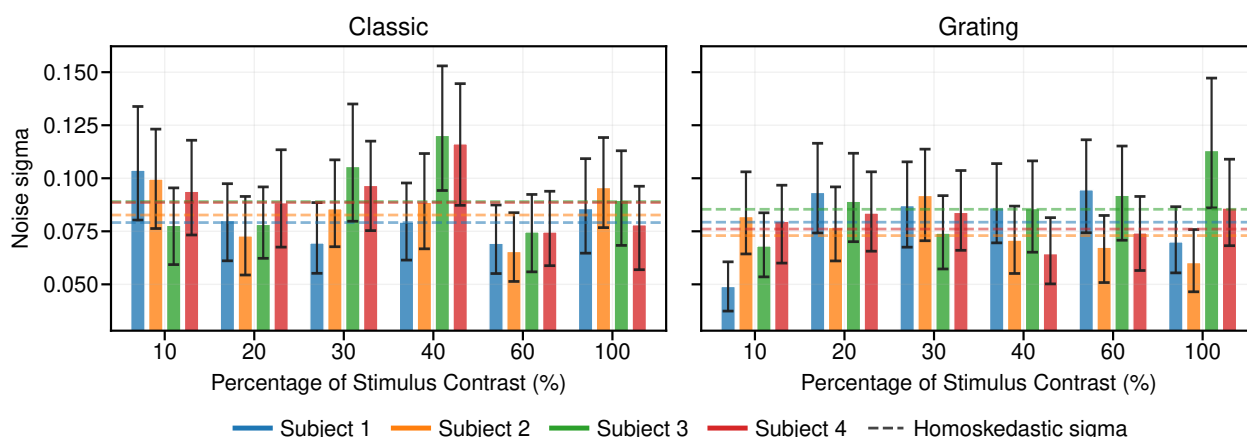


Figure 3: **Estimated observation noise by contrast and subject.** Estimated observation noise is shown across contrast levels for the classic (left) and the grating condition (right). Bars indicate posterior mean noise estimates with 94 % highest-density intervals.

contrast range, whereas Subject 3 remained weakest and reached only 0.10 at 100 % contrast.

The grating condition produced lower trial-level performance overall than the classic condition. All grating curves started near or below zero at 10 % contrast and increased more gradually with contrast, reaching a lower across-subject mean of 0.16 at 100 % contrast. The magnitude of the condition effect differed across subjects. Subject 1 showed nearly identical performance across conditions at high contrast (the difference between classic and grating at 100 % was only 0.004), whereas Subject 4 retained a strong classic advantage throughout the range (difference at 100 % contrast: 0.304).

Contrast-dependent observation noise:

Allowing the observation noise to vary by contrast revealed a modest level of heteroskedasticity in z_{margin} (Figure 3), although the pattern was subject- and condition-specific rather than globally monotonic.

At the individual level, noise in the classic condition was typically increased either at low contrast (Subjects 1–2: $\sigma \approx 0.10$ at 10 % contrast) or around mid contrast (Subjects 3–4: $\sigma \approx 0.12$ at 40 % contrast).

Noise levels in the grating condition were generally lower at 10 % contrast (mean across subjects = 0.070) and tended to increase toward intermediate or high contrasts. The clearest high-contrast increase was observed for Subject 3 in the grating condition, where σ reached 0.113 at 100 % contrast.

However, within-subject changes in noise across contrast were relatively small overall (ranges approximately 0.02–0.05), and many 94 % HDIs overlapped. This indicates that contrast-dependent noise effects were weaker and less systematic than the subject- and condition-dependent differences observed in the mean contrast-response curves in Figure 2.

DISCUSSION

We examined how stimulus contrast shapes trial-level performance in c-VEP stimulation, comparing a classic luminance-flicker condition with a spatially structured grating condition. Across subjects and conditions, trial-level performance generally increased with contrast, supporting the expectation that stronger stimulus contrast elicits more separable target-related responses. However, the results also showed that this effect is not captured well by a single global contrast rule: the shape and magnitude of the contrast–response relationship varied substantially across subjects and differed between the two stimulus families.

In this dataset, increasing contrast appeared to primarily improve the *mean* trial-level performance, while contrast-dependent changes in estimated trial-to-trial noise were comparatively modest. In other words, higher contrast mainly improved the separation between the true target and competing targets rather than strongly altering the variability of this trial-level performance metric across trials. This is consistent with the interpretation that contrast strengthens the evoked response itself, in line with known contrast effects in visual cortex and electrophysiology, and that the practical decoding benefit arises mainly from a stronger target-related signal rather than from a substantial reduction in uncertainty. In the present data, this effect was visible already at the level of subject-wise Gaussian process curves, where the increase in z_{margin} with contrast was clearer and more systematic than any corresponding pattern in the noise estimates.

A second observation is that the classic condition generally yielded stronger trial-level performance than the grating condition, with spatially uniform luminance modulation producing a larger decision-relevant margin than the textured grating stimulus. This contrasts with the findings of Dehaes and colleagues, who reported that textured flickers based on Gabor and Ricker patterns achieved higher classification accuracy than plain flicker stimuli, while also improving perceived visual comfort [7]. The discrepancy likely reflects differences in stimulus sequences, the user interface, the number of symbols, the spatial structure, and the decoding pipelines across studies, suggesting that the benefit of textured stimulation is not universal but depends on implementation details. Spatial structure is known to interact with contrast processing in the visual system, meaning that nominally identical contrast levels can drive different neural populations depending on stimulus texture and spatial frequency [18].

Importantly, the difference between the two stimulus families was not uniform across subjects. In some subjects, the grating condition approached the classic condition at higher contrast levels, whereas in others the classic condition retained a clear advantage across the entire range. This inter-individual variability is practically important, as it suggests that both the optimal contrast level and the preferred stimulus family may vary across

users. The present sample is too small to identify reliable predictors of these differences. Age could contribute, given known age-related changes in visual processing, but this cannot be tested with $N = 4$. Other factors, including visual acuity, attention, fatigue, cortical response strength, or preference for spatially structured stimulation, may also play a role. Larger studies should examine whether demographic, perceptual, or neurophysiological measures predict which users benefit from specific stimulus settings.

The margin-based metric z_{margin} provides a trial-level view of how strongly the neural response supports the correct decision within the decoder. Although it is not itself an online BCI performance measure, it is directly related to classification: larger positive margins indicate more robust correct selections, whereas margins close to zero indicate decisions that are more vulnerable to noise or changes in stopping time. This decoder-level perspective complements physiological stability measures, such as correlations or covariance between repeated EEG epochs, which capture response reproducibility more directly and have been shown to predict individual differences in c-VEP performance [10]. Future work could therefore combine signal-level stability measures with trial-level decision metrics when evaluating stimulus design.

Several limitations should be noted. First, the number of analyzed subjects was small ($N = 4$), and the study should therefore be interpreted as preliminary and exploratory. The goal was not to establish population-level effects, but to test whether trial-level modeling can characterize how stimulation parameters relate to both mean decoding performance and trial-to-trial reliability. Contrast level and stimulus family therefore served as example parameters for a broader goal: individualized stimulus-parameter selection in c-VEP BCIs. The observed between-subject differences should accordingly be viewed as possible individual trends, not as stable or generalizable group-level effects.

Second, the number of trials per stimulus setting was limited to 32. This limits the precision of absolute performance estimates, but the present analysis aimed to compare parameters rather than obtain a definitive performance value for each setting. Sosulski and Tangermann showed that reduced ERP-BCI datasets can still preserve information about which stimulation parameters are preferable, even when absolute estimates are noisy [19]. Thus, 32 trials per contrast level were considered sufficient for an exploratory within-subject comparison of contrast-dependent parameter effects, but not for precise population-level estimates.

Third, the analysis was performed offline and within a single decoding pipeline. Although the CCA–LDA approach was chosen as a standard and regularized c-VEP decoding pipeline, different preprocessing choices or classifiers could interact differently with stimulus structure, particularly for the grating condition. In addition, contrast was defined physically relative to display lu-

minance. Perceived contrast may differ across participants and could therefore contribute to the observed inter-individual variability. Finally, practical stimulus selection should also consider comfort, fatigue, and visual strain, which may favor stimuli that do not maximize decoding performance alone.

In summary, the present findings suggest that contrast can influence trial-level performance in c-VEP, primarily through its effect on the strength of the decision-relevant margin between the target and competing classes. In this dataset, the classic condition produced stronger trial-level performance overall, whereas the grating condition showed weaker but qualitatively similar contrast dependence. Modeling trial-wise performance with Gaussian processes provided an interpretable way to characterize both the mean contrast–response relationship and its uncertainty, and may be useful for future studies seeking to optimize visual stimulation parameters.

CONCLUSION

Overall, stimulus contrast influenced trial-level c-VEP decoding performance in this preliminary dataset, mainly by increasing the decision-relevant decoding margin. The effect was stronger for classic luminance flicker than for grating stimulation, while estimated trial-to-trial noise changed only modestly. These findings illustrate how trial-level evidence metrics and Gaussian process modeling can characterize stimulus-parameter effects beyond aggregate accuracy.

ACKNOWLEDGMENTS

This publication is part of the project Dutch Brain Interface Initiative (DBI2) with project number 024.005.022 of the research programme Gravitation which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

- [1] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): A literature review. *Journal of Neural Engineering*. 2021;18(6):061002.
- [2] Geisler WS, Albrecht DG. Cortical neurons: Isolation of contrast gain control. *Vision Research*. 1992;32(7):1409–1410.
- [3] Kubová Z, Kuba M, Spekreijse H, Blake-more C. Contrast dependence of motion-onset and pattern-reversal evoked potentials. *Vision Research*. 1995;35(2):197–205.
- [4] Ash RT, Nix KC, Norcia AM. Stability of steady-state visual evoked potential contrast response functions. *Psychophysiology*. 2024;61(1):e14412.
- [5] Tsoneva T, Garcia-Molina G, Desain P. Electrophysiological model of human temporal contrast sensitivity based on SSVEP. *Frontiers in Neuroscience*. 2023;17:1180829.
- [6] Xu H, Hsu SH, Nakanishi M, Lin Y, Jung TP, Cauwenberghs G. Stimulus design for visual evoked potential based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2023;31:2545–2551.
- [7] Dehais F, Castillos KC, Ladouce S, Clisson P. Leveraging textured flickers: A leap toward practical, visually comfortable, and high-performance dry EEG code-VEP BCI. *Journal of Neural Engineering*. 2024;21(6).
- [8] Castillos KC, Ladouce S, Darmet L, Dehais F. Burst c-VEP based BCI: Optimizing stimulus design for enhanced classification with minimal calibration data and improved user experience. *NeuroImage*. 2023;284:120446.
- [9] Ahmadi S, Desain P, Thielen J. A bayesian dynamic stopping method for evoked response brain-computer interfacing. *Frontiers in Human Neuroscience*. 2024;18:1437965.
- [10] Velut SRH, Thielen J, Chevallier S, Corsi MC, Dehais F. Neurophysiological screening of individual variability for robust decoding in c-VEP-based BCI. *Imaging Neuroscience*. 2026.
- [11] Gramfort A et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013;7(267):1–13.
- [12] Thielen J, van den Broek P, Farquhar J, Desain P. Broad-Band visually evoked potentials: Re(con)volution in brain-computer interfacing. *PLOS One*. 2015;10(7):e0133797.
- [13] Thielen J, Marsman P, Farquhar J, Desain P. From full calibration to zero training for a code-modulated visual evoked potentials for brain–computer interface. *Journal of Neural Engineering*. 2021;18(5):056007.
- [14] Pedregosa F et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
- [15] Zimmerman DW, Zumbo BD, Williams RH. Bias in estimation and hypothesis testing of correlation. *Psicológica*. 2003;24(1):133–158.
- [16] Boynton GM, Demb SA, Glover DJ, Heeger DJ. Neuronal basis of contrast discrimination. *Vision Research*. 1999;39(2):257–269.
- [17] Abril-Pla O et al. Pymc: A modern and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*. 2023;9(e1516).
- [18] Sclar G, Maunsell JHR, Lennie P. Coding of image contrast in central visual pathways of the macaque monkey. *Vision Research*. 1990;30(1):1–10.
- [19] Sosulski J, Tangermann M. Extremely reduced data sets indicate optimal stimulation parameters for classification in brain-computer interfaces. In: 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE, 2019, 2256–2260.

DECODING NEUROETHICS: QUALITY AS THE CODE FOR NEURO-AI GOVERNANCE

E. Kalugina¹

¹HSE University, Perm

E-mail: kalugina.law@gmail.com

ABSTRACT: AI-powered brain-computer interfaces (BCIs) are rapidly advancing, offering transformative potential in healthcare and human augmentation. However, AI's ability to decode brain signals raises profound concerns about neuroethical risks and potential harms. When such systems produce erroneous outputs, it is not merely a software defect but a potential violation of cognitive liberty, mental privacy, and neural integrity. The dominant governance response – neuroethics – has largely focused on rights articulation rather than quality as the operational mechanism capable of translating ethical principles into enforceable controls. This gap is compounded by a structural accountability problem across the BCI value chain. While AI capabilities evolve dynamically, existing safety frameworks remain static, and often do not keep pace with technological developments. This article presents a normative doctrinal analysis of EU product safety regulations as they apply to AI-powered BCIs. It examines how quality obligations, risk assessment, and liability should be structured across the BCI value chain in response to the neuroethical risks posed by neural decoding systems. On that basis, the article argues that quality should function as the central legal standard for AI-powered BCIs and proposes a dynamic NeuroAI Governance Framework to operationalize that position.

INTRODUCTION

The ability to decode brain signals is no longer speculative. AI-powered BCIs are deployed today in clinical rehabilitation, consumer wellness devices, and emerging human-augmentation applications. By combining high-density neural signal acquisition – electroencephalography, electrocorticography, and implanted electrode arrays – with deep-learning decoding algorithms, modern BCIs can infer intentions, emotions, motor commands, and imagined speech from neural data [1]. Neural decoding proceeds through a multi-stage pipeline: signal acquisition, pre-processing and denoising, feature extraction, classification, and post-processing via language models. At each stage, AI acts not merely as a filter but as an active interpreter – what has been termed a “semantic amplifier” [21]: transforming sparse, noisy neural data into rich interpretations by drawing on massive prior knowledge about the world [22].

This interpretive role has undergone a fundamental shift. Early (*discriminative*) decoding models relied on pattern-matching and classification against predefined categories – mapping a neural signal to “move left” or “select letter A”. Contemporary (*generative*) decoding models integrate LLMs and diffusion architectures to reconstruct continuous language and visual content from brain activity [22,23]. Recent studies demonstrate that non-invasive EEG- and fMRI-based systems can now reconstruct continuous spoken language at up to 78 words per minute [24,25] and generate high-resolution visual reconstructions from cortical activity [22]. This distinction carries direct legal implications: in

discriminative decoding, a defect is a misclassification; in generative decoding, a defect may be a *hallucination* – the AI completing or distorting a thought based on its prior knowledge rather than the user's actual neural state. The nature of the defect, and thus of the manufacturer's evidentiary obligation, differs accordingly.

BCIs are increasingly conceptualised as the primary frontier at which human and digital cognition converge, giving rise to what researchers term the “Hybrid Mind” [2]. As AI systems become capable of inferring complex cognitive states, they begin to function as *proxy agents* for the human mind: autonomous intermediaries that interpret and act upon neural data. When an AI system misinterprets a user's neural signals, the outcome is potentially a breach of *cognitive liberty*, *mental privacy*, and *neural integrity* – not merely a software bug [3].

Neuroethics has been the dominant framework for addressing these concerns. Scholars have articulated a growing catalogue of neurorights, and international bodies including UNESCO have adopted dedicated ethical guidelines for neurotechnology [4]. Yet ethical principles alone do not determine who bears liability when an AI-powered BCI causes harm. Soft-law instruments carry no sanctions and provide no mechanism for allocating responsibility across complex value chains. Quality – encompassing both the reliability characteristics of AI decoding algorithms and the management procedures – is the only operational mechanism capable of translating neuroethical principles into enforceable engineering obligations.

This gap between the sophistication of neurotechnology and the adequacy of legal instruments constitutes what this article terms the *Neuro-Accountability Gap*. It has three structural dimensions. First, a legal dimension: neuroethical norms lack enforceability. Second, an accountability dimension: the opacity, autonomy, and adaptability of AI systems combined with the complexity of the BCI value chain and the "many hands" problem [9] make it exceedingly difficult to establish defectiveness, causation, and the liable actor when harm occurs. Third, a dynamic dimension: current AI governance frameworks are static, while AI capabilities – autonomy, efficacy, generality, and goal complexity – evolve continuously, demanding equally evolving quality controls [27]. The new EU regulatory framework, particularly EU AI Act, PLD 2024/2853, GPSR 2023/988, and MDR 2017/745, has the potential to address it by converting abstract neuroethical norms into legal obligations enforceable through product liability. The principal driver of this conversion is quality: the technical accuracy, robustness, and reliability of AI decoding algorithms, which emerges in this regulatory environment as the manufacturer's primary legal defence.

PRODUCT DEFECTIVENESS AND AI-POWERED NEURAL DECODING

The traditional framework and its limitations

Product liability in the EU has historically rested on Council Directive 85/374/EEC, establishing strict liability for defective products causing damage to health or property. Applied to AI-powered BCIs, this framework encounters three structural problems. First, the opacity of deep-learning neural decoding algorithms renders it practically impossible for an injured party to identify the specific computational step that caused harm [1,13]. Second, AI systems adapt after deployment, meaning the product that caused harm may differ substantially from the product originally certified. Third, BCI-related harm is frequently psychological or cognitive in nature – a form of damage Directive 85/374/EEC was not designed to address.

The expanded definition of defect under PLD 2024/2853

Directive 2024/2853 on Product Liability represents a substantive revision of EU product liability law, specifically calibrated for AI-enabled products [6]. Article 7(1)(b) explicitly includes *reasonably foreseeable use and misuse* as a circumstance relevant to defectiveness – extending liability to foreseeable misapplications of AI neural decoding algorithms. Article 4(6) defines "damage" to include *psychological impairment*, making violations of mental privacy and

biased AI-powered neural decoding actionable under EU product liability law.

Elizabeth Steindl has argued that for most consumer neurodevices the primary risks concern data privacy rather than health, placing them within the orbit of the GDPR [7]. However, where a BCI's AI component produces a distorted representation of the user's mental state and that distortion causes measurable psychological damage, product safety law applies independently of data protection law. Crucially, this applies regardless of whether the decoding model is discriminative or generative: a generative model that "hallucinates" a mental state causes harm attributable to the manufacturer who trained and integrated it, not to the data controller who operates it.

Illustrative scenario: the mis-decoded speech command

Consider the following scenario. A user with ALS relies on a communication BCI that reconstructs imagined speech via generative neural decoding models. During a period of acute fatigue – a foreseeable edge-case condition – the algorithm generates a plausible but incorrect semantic output and transmits an unintended workplace instruction, resulting in financial harm and severe psychological distress. Under Article 7 PLD 2024/2853, the manufacturer faces a defectiveness claim: the product failed to provide the safety users were entitled to expect (Art. 7(1)), and the documented psychological impairment qualifies as damage under Article 4(6). Under Article 9 PLD, given the technical opacity of AI-powered neural decoding, the court may *presume* defectiveness if the claimant shows the product caused damage and defectiveness is plausible. To rebut this presumption, the manufacturer must produce: (a) accuracy logs showing performance under fatigue-like signal conditions; (b) out-of-distribution validation records; (c) evidence that an abstention or confirmation mechanism was implemented. For a generative model, the manufacturer must additionally demonstrate that the decoding output was constrained by the user's actual neural state – for example, through calibration logs, uncertainty quantification records, or conformal prediction coverage guarantees. Without such documentation, the manufacturer has no viable defence.

The presumption of defectiveness (Article 9 PLD)

Article 9 PLD 2024/2853 introduces a presumption of defectiveness where a claimant faces *excessive difficulty* due to technical or scientific complexity. For AI-powered BCIs, courts may readily invoke this presumption given the opacity of AI neural decoding algorithms. The presumption applies across four conditions: (1) the defendant fails to disclose relevant evidence; (2) the product does not comply with mandatory product safety requirements under EU law;

(3) an obvious malfunction caused the damage; or (4) proving defectiveness is excessively difficult due to technical complexity. Condition (2) is particularly significant: where an AI system should have complied with EU AI Act quality obligations but did not, non-compliance itself triggers the presumption. However, a critical gap remains: for non-high-risk BCIs, non-medical BCIs, AI agents, or GPAI-integrated decoding systems, the EU AI Act and GPSR impose no specific quality requirements – leaving manufacturers without clear standards to demonstrate and injured parties without a straightforward basis for invoking the presumption. Closing this gap requires explicit quality standards for all neural decoding systems, irrespective of their medical classification. To rebut it, manufacturers must produce logs, validation records, accuracy metrics, and post-market surveillance data. A manufacturer that has not documented its AI model's accuracy rates, false-positive and false-negative distributions, and robustness under edge-case conditions is left without a legal defence when Article 9 is invoked.

PRODUCT QUALITY AS THE MANUFACTURER'S LEGAL SHIELD

The Precautionary Principle (Article 2(5) GPSR)

Article 2(5) GPSR 2023/988 codifies the Precautionary Principle: the absence of full scientific proof does not prevent protective measures where there are reasonable grounds to believe a product could be harmful. For AI-powered BCIs, this creates a structural presumption of potential risk – because the long-term neurological and psychological effects of AI neural decoding systems, including *algorithmic shadows* (systematic biases in the decoded representation of a user's mental state) [8], and in generative models the risk that prior training data shapes the decoded output more than the user's actual neural activity, are not yet empirically established – that manufacturers must actively rebut.

Quality as legal justification and state-of-the-art defence

In the absence of harmonised standards for AI-powered neural decoding – whether discriminative or generative – demonstrable technical accuracy, reliability, and robustness constitute the state-of-the-art benchmark under Article 6 GPSR. Quality ceases to be a commercial attribute and becomes a *legal justification*: the objective mechanism by which a manufacturer moves its product from “potentially hazardous” to “safe with an acceptable level of risk.”

Article 10 PLD 2024/2853 retains the development risks defence, but to invoke it, the manufacturer must demonstrate it applied the highest available quality standards in data curation, model training, validation,

and post-market surveillance. Quality therefore functions as the *primary* legal standard of care – though it interacts with other defences including adequate warnings, instructions for use, and compliance with harmonised standards where available.

Reducing the zone of reasonably foreseeable misuse

The safe product assessment under Article 6 GPSR takes into account behaviour under *reasonably foreseeable misuse*. For BCIs, foreseeable misuse conditions include acute stress, emotional distress, cognitive impairment, and pharmacological alteration – all of which may generate neural signals substantially different from training data. For generative decoding systems, these conditions also increase the risk that the model will “fill in” a plausible but incorrect mental state from its prior knowledge, amplifying the hallucination risk. High-quality AI neural decoding systems can be engineered to recognise signal degradation and respond with appropriate caution: suspending output, requesting confirmation, or escalating to human oversight. The more robust the quality framework – including out-of-distribution detection, abstention mechanisms, and graceful degradation – the more credibly the manufacturer demonstrates compliance with the state-of-the-art standard.

WHO IS RESPONSIBLE? LIABILITY IN THE BCI VALUE CHAIN

The many-hands problem

The AI-enabled BCI involves: (1) a hardware manufacturer; (2) AI algorithm developers; (3) a GPAI model provider; (4) a system integrator; (5) a deployer/operator; and (6) the end user. This multi-stakeholder structure creates the “many hands” problem of AI liability [9]: when harm occurs, each actor contributes to the conditions producing it, yet no single actor bears unambiguous responsibility. The shift from discriminative to generative decoding deepens this problem: where a GPAI foundation model contributes its “world model” to the decoded output, the boundary between the model provider's contribution and the system integrator's contribution becomes legally opaque. Multi-agent BCI deployments further amplify this complexity: as Hammond et al. identify, multi-agent AI systems introduce distinct failure modes – miscoordination, conflict, and collusion – alongside emergent agency and network effects that compound attribution challenges and may make harm causation irrebuttable without robust quality logging [28].

Liability under PLD 2024/2853 and the EU AI Act

Article 4 PLD extends liability to importers, authorised representatives, and fulfilment service providers. Where

the AI neural decoding system qualifies as a *high-risk AI system* under Annex III of the EU AI Act – including BCIs used in medical applications and, where not prohibited under Article 5, emotion-recognition applications in regulated contexts – the AI provider bears primary obligations including conformity assessment, technical documentation, and accuracy and robustness requirements (Article 15 AI Act). It should be noted that Article 5(1)(f) AI Act prohibits emotion recognition in workplace and educational contexts; the high-risk classification under Annex III applies to emotion recognition in other regulated contexts. This distinction is directly relevant to consumer BCIs marketed as stress or affect monitors.

A significant regulatory lacuna remains: consumer wellness BCIs processing EEG signals to infer stress or cognitive states may fall below the high-risk threshold. Yet their potential for harm to mental privacy and cognitive self-determination may be substantial. This article argues that any AI system that performs neural decoding of brain signals should, as a minimum, be subject to an enhanced risk assessment under the AI Act, irrespective of its ostensible purpose, because the long-term consequences of algorithmic mediation of neural activity remain scientifically undetermined [10]. This argument is reinforced by the generative decoding paradigm: a consumer wellness BCI that integrates an LLM to infer mood or intent from EEG data is functionally equivalent, in terms of interpretive power and risk, to a medical-grade speech decoding system.

GPAI models and post-deployment learning

The EU AI Act does not establish quality requirements for GPAI models as such. If a BCI manufacturer integrates a GPAI foundation model that produces a hallucination – a semantically plausible but neurally unfounded output – the allocation of liability between the hardware manufacturer, GPAI provider, and system integrator remains unresolved. This problem is compounded by post-deployment learning [11], which can cause the deployed model to diverge significantly from the certified one. Quality frameworks – continuous monitoring, algorithmic logging, testing, model versioning – are the primary available mechanism for closing this liability gap and enabling proportional fault apportionment.

TOWARDS A NEURO-AI GOVERNANCE FRAMEWORK

The regulatory gap

The current EU regulatory architecture leaves AI-powered BCIs that do not qualify as medical devices or high-risk AI systems under no affirmative quality requirements: GPSR is risk-oriented but mandates no

specific quality level; the AI Act's quality obligations cover only high-risk systems; MDR applies only to medical devices. Consumer neurotechnology BCIs fail through all three nets. This gap is particularly acute for generative neural decoding applications – wellness BCIs, consumer EEG-based mood monitors, and LLM-integrated communication tools – which combine the interpretive power of foundation models with the intimacy of neural data, yet currently attract no mandatory quality obligations. Moreover, existing regulatory frameworks are static, while AI capabilities evolve dynamically. What is needed is an agile governance approach in which quality obligations and risk assessments co-evolve in response to changing AI capabilities – a Risk-Quality Feedback Loop rather than a fixed compliance checklist.

Framework components

The proposed NeuroAI Governance Framework addresses this gap through six integrated components:

(1) *BCI Risk Taxonomy* – four tiers: (i) medical BCIs (MDR); (ii) MDR-quasi BCIs; (iii) high-risk AI BCIs (AI Act Annex III); (iv) other neurotechnologies applying AI-powered neural decoding. Each tier attracts proportionate quality obligations. The taxonomy distinguishes between discriminative and generative decoding within each tier, with generative systems attracting stricter quality obligations due to their hallucination risk.

(2) *Risk Assessment Methodology* – incorporating AI opacity, adaptability, and emergent behaviour, alongside neurotechnology-specific harm categories: mental privacy violation, algorithmic shadow effects, cognitive liberty infringement, and psychological harm. Critically, the methodology must also account for AI capabilities. Following Kasirzadeh and Gabriel [27], AI agents can be classified by their qualitative capabilities: autonomy, efficacy, impact, generality, and goal complexity. The higher these capabilities, the greater the risk – and therefore the more stringent the quality obligations triggered. For multi-agent systems, risk assessment must additionally incorporate the failure modes identified by Hammond et al. [28]: miscoordination, conflict, and emergent collusion that static taxonomies cannot capture. This capability-based dimension is absent from current static regulatory frameworks.

(3) *Quality Models* – minimum performance benchmarks across: (a) data quality (curation, diversity, annotation); (b) AI algorithm quality (accuracy, robustness, OOD performance, hallucination rate, fidelity of interpretation for generative models); (c) AI product quality (integration, explainability); (d) neural decoding process quality (pre-processing, real-time performance, graceful degradation). Quality characteristics and AI capability assessments must

directly inform risk assessments – creating a dynamic feedback loop rather than a one-time conformity check.

(4) *Quality Controls* – implemented at each stage of the BCI lifecycle: *research and pre-clinical testing*, design, development, conformity assessment, market placement, post-market surveillance, and end-of-life. Research-phase controls are particularly critical for consumer neurotechnologies where claims of efficacy or safety are not always backed by independent clinical evidence [20]. These controls embed neuroethical principles as operational engineering requirements and generate the documentary record necessary for liability defence under PLD 2024/2853. Controls are determined by both regulatory requirements and applicable standards (ISO 14971, IEC 62304, ISO/IEC 23894), and must include at minimum: documented accuracy thresholds, OOD detection metrics, hallucination rate ceilings, and post-market drift monitoring.

(5) *Risk-Quality Balance Scale* – proportionality principle: the higher the risk tier, the more stringent the quality requirements. The governing logic is bidirectional: high risk demands stricter quality; and conversely, higher AI capabilities (autonomy, generality, goal complexity) require a higher risk tier assignment. This scale operationalises the Precautionary Principle as a quality imperative.

(6) *Accountability and Liability Distribution Rules* – specifying conditions under which product liability may attach to each actor depending on the source of the defect that caused harm (hardware manufacturer, AI algorithm developer, GPAI model provider, system integrator, deployer) and the mechanism for proportional liability allocation in cases of joint causation. Additionally, the GPSR safety requirements framework should be explicitly adapted to AI-enabled systems: product safety assessments must account for mental safety alongside physical safety, and must incorporate AI-specific risk factors including hallucination, drift, and post-deployment capability expansion.

DISCUSSION

The findings above carry three implications for BCI researchers, manufacturers, and policymakers.

First, the legal framework is shifting faster than industry practice. PLD 2024/2853 entered into force in December 2024 with a transposition deadline of December 2026. Manufacturers who do not begin building quality documentation infrastructures now – accuracy logs, OOD validation records, drift detection systems – will face an evidentiary deficit when the first litigation arises. This urgency is amplified for manufacturers integrating generative neural decoding models: existing logging practices designed for

discriminative classifiers do not capture the hallucination risk profile of LLM-integrated systems.

Second, quality and risk must be governed in symbiosis, not in isolation. The NeuroAI Governance Framework is grounded in a Risk-Quality Feedback Loop: quality controls reduce risks, but expanding AI capabilities generate new risks that demand updated quality controls. This dynamic is absent from current EU regulatory frameworks, which treat risk assessment and quality requirements as separate, static exercises. For AI-powered BCIs, where post-deployment learning and increasing model autonomy continuously alter the risk profile, only a co-evolutionary approach – in which quality metrics and risk assessments are continuously updated in response to changing AI capabilities – can ensure sustained safety and accountability.

Third, the regulatory gap for consumer neurotechnologies is real and urgent. The NeuroAI Governance Framework proposed here is a first step towards filling it. Its operationalisation requires collaboration between legal scholars, neuroscientists, and standards bodies including CEN/CENELEC, ISO/IEC JTC 1/SC 42, and IEEE – which has already developed dedicated neuroethics and BCI standards roadmaps. Without such collaboration, the translation of neuroethical principles into enforceable engineering obligations will remain incomplete.

CONCLUSION

AI-powered BCIs represent one of the most legally consequential product categories of the coming decade. The new EU regulatory framework, anchored by EU AI Act, PLD 2024/2853 and GPSR 2023/988, is beginning to supply the missing legal architecture: the expanded definition of defectiveness now encompasses psychological harm caused by AI misinterpretation or hallucination of neural signals; the presumption of defectiveness shifts the evidentiary burden to manufacturers; and the Precautionary Principle creates a structural presumption of risk under scientific uncertainty.

In this regulatory environment, quality is the primary legal defence available to the product manufacturer. The quality of AI-powered BCIs is not just a matter of product safety: it engages fundamental rights – mental privacy, cognitive liberty, and neural integrity – that demand protection through enforceable engineering obligations, not merely ethical guidelines. For manufacturers integrating generative neural decoding models, this means documenting not only classification accuracy but fidelity of interpretation: the degree to which the decoded output reflects the user's actual neural state rather than the model's prior knowledge. Crucially, governance of AI-powered BCIs cannot be static. As AI capabilities expand – in autonomy, efficacy, and generality – risk profiles

evolve, and quality obligations must evolve with them. The NeuroAI Governance Framework, grounded in a dynamic Risk-Quality Feedback Loop, represents a first attempt to supply that regulatory clarity.

REFERENCES

- [1] Cao Z. A review of artificial intelligence for EEG-based brain-computer interfaces and applications. *Brain Sci Adv.* 2020;6(3):162-170.
- [2] Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy.* 2017;13(1):5.
- [3] Bublitz JC. Novel neurorights: From nonsense to substance. *Neuroethics.* 2022;15:7.
- [4] Lighthart S, Bublitz JC, Alegre S. Neurotechnology: We need new laws, not new rights. *Nature.* 2023;620(7976):950-952.
- [6] Directive (EU) 2024/2853 on liability for defective products [2024] OJ L 2024/2853.
- [7] Steindl E. Consumer neuro devices within EU product safety law. *Comput Law Secur Rev.* 2024;52:105945.
- [8] Li T. Algorithmic destruction. SSRN. 2022. doi:10.2139/ssrn.4066845
- [9] European Parliament. Report on liability for artificial intelligence (A9-0178/2022). 2022.
- [10] Farahany NA. *The Battle for Your Brain.* St. Martin's Press; 2023.
- [11] OECD. Recommendation on AI, Principles 1.3-1.4. 2019.
- [12] Ienca M, Malgieri G. Mental data protection and the GDPR. *J Law Biosci.* 2022;9(1):lsac006.
- [13] Maiseli BJ, Abdalla AT, Massawe LV, Mbise M, Mkocha K, Nassor NA, Ismail M, Michael J, Kimambo S. Brain-computer interface: Trend, challenges, and threats. *Brain Inform.* 2023;10(1):20.
- [14] Chandler JA, van der Loos C, Boencke S, Buchman DZ, Baudry JS, Illes J. BCIs and communication disorders. *Front Hum Neurosci.* 2022;16:841035.
- [15] Gilbert F, Ienca M, Cook M. How I became myself after merging with a computer. *Brain Stimul.* 2023;16(3):783-789.
- [16] Haag L, Starke G, Ploner M, Ienca M. Ethical gaps in closed-loop neurotechnology. *NPJ Digit Med.* 2024;8:510.
- [17] Magee P, Ienca M, Farahany N. Beyond neural data: Cognitive biometrics and mental privacy. *Neuron.* 2024;112(18):3017-3028.
- [18] Lighthart S, Ienca M, Meynen G, et al. Minding rights: Neurorights foundations. *Camb Q Healthc Ethics.* 2023;32(4):461-481.
- [19] Regulation (EU) 2023/988 on general product safety [2023] OJ L135/1.
- [20] Maia de Oliveira Wood G, Berger L, Jarke J, Barnard G, Gremsl T, Dolezal E, Staudegger E, Zandonella P. The protection of mental privacy in the area of neuroscience. European Parliament. 2024. doi:10.2861/869928
- [21] Ienca M. Towards a governance framework for brain data. *EMBO Mol Med.* 2022;14(9):e15337.
- [22] Takagi Y, Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity. *CVPR.* 2023. doi:10.1109/CVPR52729.2023.02151
- [23] Tang J, LeBel A, Jain S, Huth AG. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neurosci.* 2023;26(5):858-866. doi:10.1038/s41593-023-01304-9
- [24] Horikawa T. Mind captioning: Evolving descriptive text of mental content from human brain activity. *Sci Adv.* 2025;11(45):eadw1464. doi:10.1126/sciadv.adw1464
- [25] Willett FR, Kunz EM, Fan C, Avansino DT, Wilson GH, Choi EY, Kamdar F, Hochberg LR, Druckmann S, Shenoy KV, Henderson JM. A high-performance speech neuroprosthesis. *Nature.* 2023;620(7976):1031-1036. doi:10.1038/s41586-023-06377-x
- [26] Regulation (EU) 2024/1689 laying down harmonised rules on AI (AI Act) [2024] OJ L 2024/1689.
- [27] Kasirzadeh A, Gabriel I. In conversation with artificial intelligence: aligning language models with human values. *Philos Technol.* 2023;36(2):1-24. doi:10.1007/s13347-023-00606-x.
- [28] Hammond L, Chan A, Clifton J, et al. Multi-Agent Risks from Advanced AI. Cooperative AI Foundation, Technical Report #1. arXiv:2502.14143. 2025. doi:10.48550/arXiv.2502.14143.

Epidural micro-ECoG provides a robust interface to the cortex over 4 years in non-human primates

C.A. Gkogkidis¹, M. Schuettler¹, M. Ulloa¹, J.W. Ryou², K.P. Purpura², J.L. Baker²

¹ CorTec GmbH, Freiburg, Germany

² Laboratory of Cognitive Neuromodulation, Brain and Mind Research Institute,
Weill Cornell Medicine, New York, USA

E-mail: job2037@med.cornell.edu

ABSTRACT: Implantable brain-computer interfaces (BCIs) are on the verge of transitioning from the research and investigational phase to clinical readiness stage. Micro-electrocorticography (μ ECoG) is one possible choice of recording modality and shows promising long-term results in both epi- and subdural implantations with direct clinical translational impact. In this paper, we present longitudinal data from epidural implantation in non-human primates. The results show that neural patterns stay comparable over the course of multiple years, and that task- and condition-specific cortical signal dynamics can be distinguished. The findings satisfy two of the most important criteria for μ ECoG-based chronic BCIs: long-term stability of neural signal quality and the presence of potentially decodable information.

INTRODUCTION

Chronically implantable brain-computer interfaces (BCIs) are getting growing attention worldwide [1, 2, 3]. During the past 20+ years various technologies were used to electrically interface with the brain, showing a broad range of performance and longevity. The use of electrocorticography (ECoG) as a recording modality has become a viable option for long-term recordings from the cortical surface. Its use has markedly increased in a miniaturized form termed μ ECoG (usually considered electrode arrays with inter-contact spacing smaller than 10 mm), due to the better trade-off between invasiveness, spatial resolution and coverage, and without the need to penetrate the cortex. In addition, μ ECoG electrode arrays may be placed below or above the dura mater, i.e., either sub- or epidurally, the latter with the aim of further minimizing the surgical intervention. While both epi- and subdural μ ECoG electrode placement has advantages and disadvantages regarding signal quality, implantation risk, and long-term performance, the impact of electrode placement remains to be fully characterized [4]. Here, we report on neural signals obtained with silicone elastomer-based μ ECoG electrode arrays implanted epidurally in two non-human primates (NHPs) for up to 4 years, while experiments in one animal are ongoing. We report on the reliability and robustness of the recorded signals obtained while the animals performed a task and during non-engagement and drowsy periods,

thus demonstrating that the basic requirements for BCI applications are met: longitudinal signal stability and ability to capture neural cortical dynamics with high fidelity and measurable differences between conditions. These findings have direct implications for the use of μ ECoG electrode arrays in preclinical and clinical BCI research and applications.

MATERIALS AND METHODS

Design & Manufacturing of μ ECoG Grids: The employed μ ECoG grids (arrays) were fabricated in cleanrooms (CorTec GmbH, Freiburg, Germany) applying a method first described in detail in [5] and briefly explained in two modifications for the two designs used here in the following paragraphs.

(1) *Fetz Array:* A 130 μ m layer of liquid MED1000 (Nusil, Carpinteria, USA) silicone diluted 1:1 with n-Heptane was deposited to a mechanical carrier by spin-coating. After onset of curing, a perforated 15 μ m Parylene-C foil (SCS, La Haux-de-Fonds, Switzerland) was laminated on top. Another 130 μ m layer of diluted silicone was deposited by spin-coating, embedding the parylene foil. A 25 μ m of Pt90Ir10 foil (Goodfellow, Cambridge, UK) was laminated to the stack and a laser (picosecond pulse width, UV wavelength) was used to cut the perimeter of electrode contacts, integrated meandering tracks and weld pads. Excess PtIr foil was manually removed. A final silicone layer (130 μ m) was deposited. This silicone layer was removed from the electrode contact sites and the weld pads using the laser. Subsequently, the outer contour of the grid was cut with the laser. The Fetz Array (Fig. 1A) is a strip-like μ ECoG grid of 2.5 mm width and 28 mm length. It has 8 electrode contacts of 0.3 mm diameter and is about 0.4 mm thin.

(2) *4x8 Square 45° Grid:* This grid was produced similar to the Fetz Array, but instead of a parylene layer another PtIr metal layer was integrated that allows higher contact density. The 0.4 mm thin grid covers an area of 24.5 mm x 20 mm and has 32 electrode contacts of 1 mm diameter Fig. 1B).

(3) *Interconnection/Wiring:* The two electrode array types were lifted from the mechanical carrier and 75 μ m diameter Pt90Ir10 wires with 7 μ m polyesterimide insulation (Alleima, Sandviken, Sweden) were spot-

Animal Model & Cephalic Implant: This study was carried out in strict accordance with the NIH Guidelines for Use of Animals in Research, under a protocol reviewed and approved by the Weill Cornell Medical College Institutional Animal Care and Use Committee (IACUC). The animals were cared for by the Research Animal Resource Center (RARC) at Weill Cornell Medicine (WCM).

A

8-ch Fetzer Array

9L

ch 1

8B

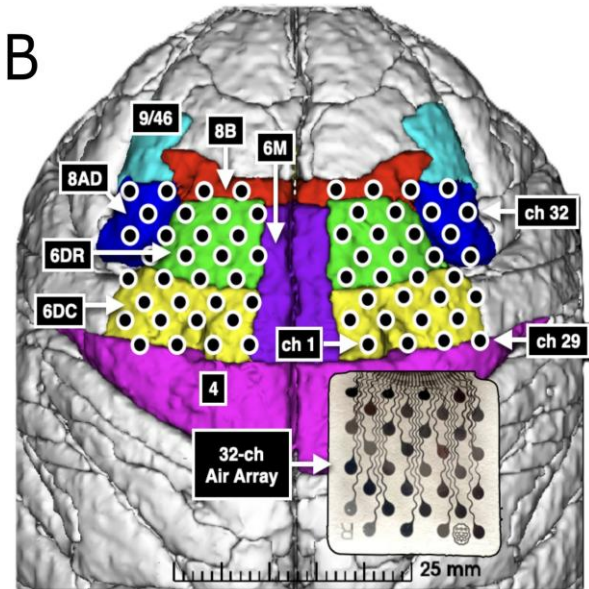
6DR

6M

6DC

ch 8

10 mm



Behavioral Task, Recording Setup & Data Analysis
The animals were trained to perform various visuomotor reaction time tasks for liquid rewards, as done in prior studies [6, 7]. Briefly, each trial of the vigilance task was

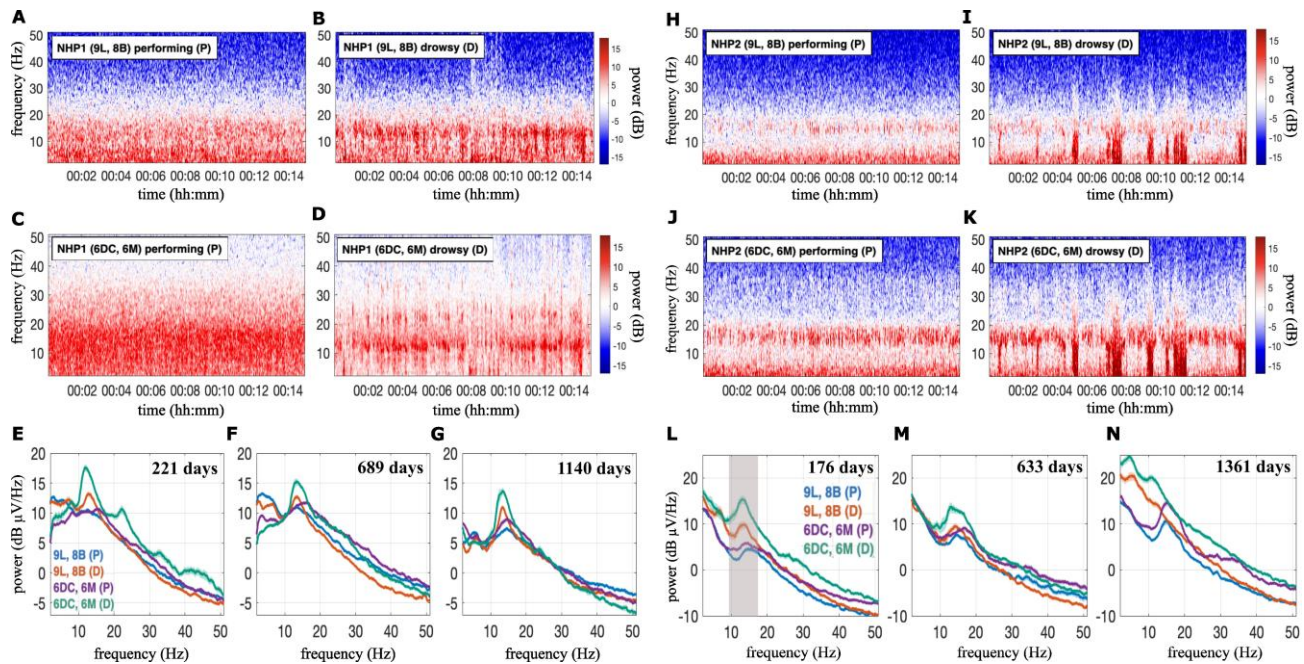


Figure 3: (A-D) Spectrograms of the epidural μ ECoG signal recorded over more frontal (9L/8B, contact 2 in Fig.2A) and more parietal/pre-motor areas (6DC/6M, contact 8 in Fig.2A) in NHP1, while the animal performed the vigilance task (performing (P), subpanels A, C) and while the animal exhibiting periods of drowsiness, eye closures, and disengagement from the task (drowsy (D), subpanels B, D), 221 days following implantation. (E-G) Average power spectral profile for the data shown in A-D at 221, 689 and 1140 days after implantation. (H-K) The same as in A-D but for NHP2, with electrode contacts at the transition of adjacent areas (9L/8B and 6DC/6M, respectively). (L-N) The same as in E-G, but for NHP2 and for 176, 633 and 1361 days after implantation. For results shown in E-N, note the largely consistent spectral profile patterns for each condition (P and D) across the implantation period, however, with varying levels, especially on day 1361 for NHP2 in which the animal exhibited drowsiness in while performing the task.

initiated by the animal first touching and then holding its hand over an infrared (IR) switch (Crist Instruments Company Inc., Hagerstown, MD, USA). The switch, attached to the front of the animal's behavioral chair, was linked to custom behavioral task programs designed in MonkeyLogic [10]. This setup controlled the visual appearance of a central gray-scale Gabor-patch (3 deg of visual angle) embedded in a $1/f$ noise pattern on an LCD screen placed 100 cm in front of the animal. The animal had to maintain fixation on the Gabor within an invisible constraint window as the spatial phase of the Gabor changed (every 300 ms) for a variable number of phase transitions (lasting 1–5 s) before changing orientation (from 0 deg to 11.25 – 90 deg). The change in orientation was the 'GO' cue for the animal to remove its hand from the IR switch, within 1 s, to receive a liquid reward. Experimental sessions typically lasted until the animal was satiated (typically within 1 to 3 h). Broadband neuronal activity was recorded with an RZ2-4 data acquisition system (TDT – Tucker Davis Technologies Inc., Alachua, FL, USA) at a sampling rate of 1-25 kHz while running TDT's Synapse software. A customized vibration sensor (801S) was attached to the behavioral chair to detect large body movements, which did at times lead to motion artifacts that were recorded in the neuronal signals (see Fig. 3B, D, I, K). Behavioral performance data and body movements were simultaneously recorded along with neural signals. The Chronux toolbox [11] was

used to generate multi-taper estimates of the power spectral density (PSD) of the μ ECoG signals. Time-frequency spectra were created to detect qualitative and quantitative differences between task-related and engaged/fatigued conditions as shown in Fig. 3. Plots showing average PSD profiles for engaged task-related and non-engaged/fatigued conditions were created for three time points (NHP1: 221, 689 and 1140 days post implantation; NHP2: 176, 633 and 1361 days) as shown in Fig. 3.

RESULTS

Electrophysiology results: Viable signals were recorded in both animals up to 4 years post-implantation and we obtained large-scale cortical μ ECoG activity (Fig. 3 and Fig. 4) during the experimental sessions. The animals typically performed well at the start of each session (Fig. 3A, C, H, J) and then performance degraded as time on task increased, due to satiety and boredom (Fig. 3B, D, I, K), at which time the animals exhibited signs of fatigue and drowsiness, which included extended eye blinks and that proceeded full eye closures. As shown in Fig. 3 for data snippets lasting ca. 15 minutes, oscillatory power within the alpha band (ca. 9-16 Hz) was significantly increased during non-engaged/drowsy conditions, as compared to when the animals were engaged and performing the task correctly.

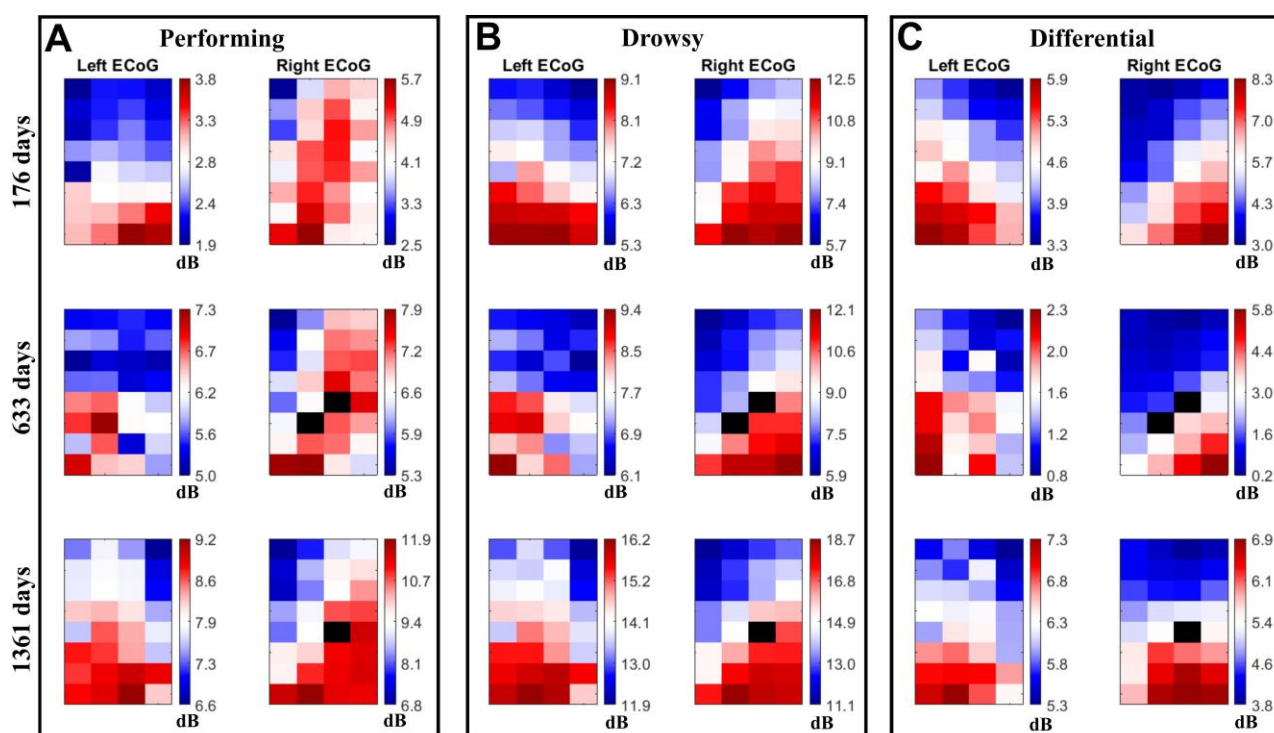


Figure 4: Spatial distribution of alpha band (9-16 Hz) spectral power changes shown for the 3 investigated time points in NHP2 (176, 633 and 1361 days after implantation) for left and right hemisphere ECoG electrode arrays and for (A) performing condition, (B) drowsy condition and (C) the differential of performing and drowsy condition. Note the varying levels of band power/activity across the left and right arrays, yet largely consistent spectral profiles and differentials across the investigated time points. Alpha band spectral power was estimated from between 9-16 Hz depicted by the gray-shaded area in Fig. 2L. Black rectangles depict dysfunctional electrode contacts. The orientation of heatmaps corresponds to the orientation of electrode arrays on the cortex of NHP2 as shown in Fig. 2B.

PSD profiles for NHP1 (Fig.3E, F, G) show slightly varying magnitude with a tendency of decreasing signal power across the three time points, however, with clear differences between the two investigated conditions (alert/performing versus drowsy) preserved.

Results for NHP2 (Fig.3L, M, N) show similar variability over time. Here, clear differences between the two conditions are seen for the first two time points (176 and 633 days), but the difference is less pronounced for the third time point (1361 days). Overall, the signal quality and spectral content under the two conditions were stable across the investigated time periods. The data shown in Fig.3 were obtained from electrode contacts placed over frontal (9L/8B) and pre-motor areas (6DC, 6M) and are representative of the recording and signal quality across both electrode arrays. Animal NHP1 tended to move more in the behavioral chair. These movements were captured by the motion sensor and also produced broadband electrical disturbances as seen in the spectrogram in Fig.3I, K (starting around 00:05). These periods were excluded from the analysis.

To investigate signal stability and preservation of spatial distribution of alpha band activity across the implantation period and across the two μ ECoG arrays for NHP2, activity was plotted as heatmaps for the performing and drowsy conditions as well as the differential of the two conditions (Fig.4). Despite varying activity levels for the three investigated time points (176, 633 and 1361 days

after implantation) a largely preserved spatial pattern for the performing (Fig.4A) and drowsy condition (Fig.4B), and most importantly a clear and robust preservation of spatial pattern for the differential (Fig.4C) can be identified for both left and right ECoG electrode arrays. These results further demonstrate μ ECoG signal stability on individual electrode contacts and across the footprint of the whole μ ECoG electrode arrays.

Technical results: Post-op reconstruction of the μ ECoG electrode array contacts confirmed that they were positioned in the epidural space. In NHP1, all 8 electrode contacts were functional and provided a good electrical coupling to the brain until the animal was euthanized according to the study protocol, and the array was removed. Studies with NHP2 are ongoing and high-quality neural signals are still being obtained from the implanted μ ECoG arrays, now more than 4 years post-implantation. Notably, in NHP2, only one out of 64 implanted electrode contacts has stopped working (high impedance) after ca. 14 months of implantation, with another one being temporarily defective at time point 633 days after implantation. We anticipate a technical failure in the one contact which will be further investigated after study finalization, which is possibly related to an experimental rejuvenation protocol applied to the individual contacts, as described by O'Sullivan and colleagues [12].

DISCUSSION

For many applications, μ ECoG grids provide an alternative to penetrating intracortical electrode arrays, for which multiple failure modes have been identified; a loss of channels (capability of sensing unit activity) over time is consistently reported in the literature [13] along with local neuronal reconfiguration, which includes neuronal death, glial scar formation, and persistent local inflammation. The signal quality and longevity of the brain interface provided by the μ ECoG electrode arrays observed in this study is consistent with findings of other groups using ECoG, e.g., that of Yang and colleagues [14] who also used electrodes produced with the same materials and methods. Although Yang et al. report data from almost 2 years in a subdural setting from two NHPs, our study demonstrates more than double the longevity for similar electrode arrays, here implanted epidurally. Davis and colleagues [15] reported on 5 years of stable use of subdural silicone-elastomer ECoG strips in a spinal cord injured patient and another study reported that cortical modulations could be observed more than 18 months post-implantation with subdural ECoG arrays in one NHP [16]. Vansteensel and colleagues [17] were able to collect successfully data from an ALS patient over the course of 7 years using the same implant technology as Davis et al. [15]; however, signal quality declined towards the end of the study which was associated with brain atrophy, “although alternative explanations are plausible” as suggested by [17]. Recently, longitudinal data from a chronic epidural electrode-based BCI system were reported [18]. The data from two patients implanted for 32 and 14 months, respectively, showed high stability over time regarding various metrics like RMS, SNR and spectral distribution [18]. Other studies investigated the impact of the dura mater on the signal amplitude and concluded that the attenuation of signals is stronger for smaller ECoG electrodes [19], similar to the electrode arrays used in this study.

These and other studies were recently incorporated into a comprehensive review by Branco and colleagues [4], which investigated the differences between epi- and subdural recordings in humans and non-human primates (NHPs). The authors reveal considerable variability in the results from the various studies, and concluded that 1) epidural recording signals attenuated more compared to subdural signals; 2) they are both subject to issues of complications and decodability of signals over time [4]. Similar to the meta-findings by Branco et al., the results in this paper show a longitudinally stable and preserved alpha band spatial pattern on the one hand and a varying signal amplitude level with a tendency of decreasing signal strength on the other. In addition, the human dura is thicker compared to that of NHPs which has to be taken into account when interpreting the results regarding signal strength in general and longitudinal signal behavior in the light of human BCI applications. However, while the decodability of the signals collected in this study were not assessed, our findings suggest that the information content in the neural signals captured

during the two behavioral conditions are different enough to be decodable. This is supported by data from one of the animals in this study where a linear classifier was used to detect the drowsy state from the μ ECoG signals and to then trigger central thalamic DBS that restored performance [20]. Of note, the studies in NHP2 are ongoing, and the complication rate is nearly zero, but standard histology has yet to be conducted to examine any potential tissue changes linked to surgical- or implantation-related complications. Furthermore the results shown in this study comprise only three days within the almost 4-year period. Although data shown across the whole μ ECoG arrays shows very well preserved spatial patterns over the investigated time points, a more comprehensive analysis will be necessary to fully characterize the longitudinal behavior of μ ECoG signals in this dataset. In summary, the longitudinal results presented in this work add to the available data in the epidural implantation realm and demonstrate the viability of implantation of μ ECoG electrode arrays above the dura as an alternative to subdural implantation for chronic μ ECoG-based brain-computer interfacing devices and applications.

CONCLUSION

Consistent with the experience collected in scientific literature about implanted ECoG grids, the study here used μ ECoG grids with 1.0 mm and 0.3 mm electrode contact diameter provide a robust electrical coupling to the cortex of two NHPs using epidural recordings. This permitted neural signal sensing with high fidelity to detect condition-specific brain oscillation changes especially in the alpha band over the observation time of ca. 4 years. The results complement the existing scientific literature reporting on chronic epidural electrode array implantation and its suitability for long-term μ ECoG-based implantable BCI devices.

ACKNOWLEDGEMENTS

This work was supported by National Institute of Neurological Disorders and Stroke grant NIH-NIHDS NS111019 and the Daedalus Fund for Innovation at Weill Cornell Medicine.

REFERENCES

- [1] Kawala-Sterniuk A, Browarska N, Al-Bakri A, Zygarlicki J, et al. “Summary of over Fifty Years with Brain-Computer Interfaces—A Review,” *Brain Sciences*, vol. 11, no. 1, Jan. 2021, doi: 10.3390/brainsci11010043.
- [2] Maiseli B, Abdalla AT, Massawe LV, Mbise M, Mkocha K, Nassor NA, et al. “Brain-computer interface: trend, challenges, and threats,” *Brain Inf.*, vol. 10, no. 1, p. 20, Aug. 2023, doi: 10.1186/s40708-023-00199-3.
- [3] Patrick-Krueger KM, Burkhart I, Contreras-Vidal JL. “The state of clinical trials of implantable brain-

- computer interfaces,” *Nat Rev Bioeng*, vol. 3, no. 1, pp. 50–67, Jan. 2025, doi: 10.1038/s44222-024-00239-5.
- [4] Branco MP, Geukes SH, Aarnoutse EJ, Ramsey NF, Vansteensel MJ. “Nine decades of electrocorticography: A comparison between epidural and subdural recordings,” *European Journal of Neuroscience*, vol. 57, no. 8, pp. 1260–1288, 2023, doi: 10.1111/ejn.15941.
 - [5] Schuettler M, Stiess S, King BV, Suaning GJ. “Fabrication of implantable microelectrode arrays by laser cutting of silicone rubber and platinum foil*,” *J. Neural Eng.*, vol. 2, no. 1, p. S121, Feb. 2005, doi: 10.1088/1741-2560/2/1/013.
 - [6] Baker JL, Ryou JW, Wei XF, Butson CR, Schiff ND, Purpura KP. “Robust modulation of arousal regulation, performance, and frontostriatal activity through central thalamic deep brain stimulation in healthy nonhuman primates,” *Journal of Neurophysiology*, vol. 116, no. 5, pp. 2383–2404, Nov. 2016, doi: 10.1152/jn.01129.2015.
 - [7] Janson AP, Baker JL, Sani I, Purpura KP, Schiff ND, Butson CR. “Selective activation of central thalamic fiber pathway facilitates behavioral performance in healthy non-human primates,” *Sci Rep*, vol. 11, no. 1, p. 23054, Nov. 2021, doi: 10.1038/s41598-021-02270-7.
 - [8] Calabrese E, Badea A, Coe CL, Lubach GR, Shi Y, Styner MA, et al. “A diffusion tensor MRI atlas of the postmortem rhesus macaque brain,” *NeuroImage*, vol. 117, pp. 408–416, Aug. 2015, doi: 10.1016/j.neuroimage.2015.05.072.
 - [9] Schiff ND, Giacino JT, Butson CR, Choi EY, Baker JL, et al. “Thalamic deep brain stimulation in traumatic brain injury: a phase 1, randomized feasibility study,” *Nat Med*, vol. 29, no. 12, pp. 3162–3174, Dec. 2023, doi: 10.1038/s41591-023-02638-4.
 - [10] Hwang J, Mitz AR, Murray EA. “NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB,” *Journal of Neuroscience Methods*, vol. 323, pp. 13–21, Jul. 2019, doi: 10.1016/j.jneumeth.2019.05.002.
 - [11] Bokil H, Andrews P, Kulkarni JE, Mehta S, and Mitra PP. “Chronux: A platform for analyzing neural signals,” *Journal of Neuroscience Methods*, vol. 192, no. 1, pp. 146–151, Sep. 2010, doi: 10.1016/j.jneumeth.2010.06.020.
 - [12] O’Sullivan K, Orazem M, Otto KJ, Butson CR, Baker JL. “Electrical rejuvenation of chronically implanted macroelectrodes in nonhuman primates,” *J Neural Eng*, vol. 21, no. 3, p. 10.1088/1741-2552/ad5703, Jun. 2024, doi: 10.1088/1741-2552/ad5703.
 - [13] Bjånes DA, Kellis S, Nickl R, Baker B, Aflalo T, Bashford L, et al. “Quantifying physical degradation alongside recording and stimulation performance of 980 intracortical microelectrodes chronically implanted in three humans for 956–2130 days,” *Acta Biomaterialia*, vol. 198, pp. 188–206, May 2025, doi: 10.1016/j.actbio.2025.02.030.
 - [14] Yang H, Fukuma R, Namima T, Okuda K, Nishi A, Iwata T, et al. “Longitudinal multitask wireless electrocorticography data from two fully implanted nonhuman primates,” *Sci Data*, vol. 13, no. 1, p. 63, Dec. 2025, doi: 10.1038/s41597-025-06359-w.
 - [15] Davis KC, Wyse-Sookoo KR, Raza F, Meschede-Krasa B, Prins NW, Fisher L, et al. “5-year follow-up of a fully implanted brain–computer interface in a spinal cord injury patient,” *J. Neural Eng.*, vol. 22, no. 2, p. 026050, Apr. 2025, doi: 10.1088/1741-2552/adc48c.
 - [16] Degenhart AD, Eles J, Dum R, Mischel JL, Smalianchuk I, Endler B, et al. “Histological evaluation of a chronically-implanted electrocorticographic electrode grid in a non-human primate,” *J. Neural Eng.*, vol. 13, no. 4, p. 046019, Jun. 2016, doi: 10.1088/1741-2560/13/4/046019.
 - [17] Vansteensel MJ, Leinders S, Branco MP, Crone NE, Denison T, Freudenburg ZV, et al. “Longevity of a Brain–Computer Interface for Amyotrophic Lateral Sclerosis,” *New England Journal of Medicine*, vol. 391, no. 7, pp. 619–626, Aug. 2024, doi: 10.1056/NEJMoa2314598.
 - [18] Larzabal C, Bonnet S, Costecalde T, Auboiroux V, Charvet G, Chabardes S, et al. “Long-term stability of the chronic epidural wireless recorder WIMAGINE in tetraplegic patients,” *J. Neural Eng.*, vol. 18, no. 5, p. 056026, Sep. 2021, doi: 10.1088/1741-2552/ac2003.
 - [19] Bundy DT, Zellmer E, Gaona CM, Sharma M, Szrama N, Hacker C, et al. “Characterization of the Effects of the Human Dura on Macro- and Micro-Electrocorticographic Recordings,” *J Neural Eng*, vol. 11, no. 1, p. 016006, Feb. 2014, doi: 10.1088/1741-2560/11/1/016006.
 - [20] Baker JL, Toth R, Deli A, Zamora M, Fleming JE, Benjaber M, et al. “Regulation of Arousal and Performance of a Healthy Non-Human Primate Using Closed-Loop Central Thalamic Deep Brain Stimulation,” in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, Apr. 2023, pp. 1–6. doi: 10.1109/NER52421.2023.10123754.

Hacking BCI Ethics

M. Welter¹, F. Lotte¹

¹Inria Center at the University of Bordeaux/LaBRI, Bordeaux, France

E-mail: marc.welter@inria.fr

ABSTRACT: In the age of surveillance capitalism, the exploitation of personal data has become ubiquitous. Although, this can be beneficial to personalize user experience, such data is also regularly exploited for manipulation and harm, e.g. to influence voting behaviour or even incite violence. As the emergence of Brain-Computer-Interface (BCI) technologies can grant unprecedented access to private and mental data, BCI users need to be protected against unethical exploits of their brain data. To be able to defend against attacks, it is helpful to be aware of possible attacks a priori. To this end, we propose to adopt an adversarial approach inspired by ethical hacking in cybersecurity to hack BCI ethics. By taking the hypothetical perspective of unethical actors, e.g. surveillance capitalists, we can determine concrete ethical vulnerabilities in BCI systems and work on patches to fix them.

INTRODUCTION

It has been suggested that consumer Brain-Computer-Interfaces (BCIs) that record and process data from their user's brain could soon be adopted by the general public.[1] At the same time, we live in an era of unprecedented and ubiquitous algorithmic data exploitation. In unison with (proto-)authoritarian governments, data appropriating businesses have created a new economy based on surveillance and the exploitation of personal user data.[2] In this economic model, often described as surveillance capitalism[3] or, more recently, technofeudalism[4], data collection and analysis algorithms are used to influence user behaviour, e.g. by personalizing advertisements in order to push product sales.

However, besides such relatively innocuous commercial goals, these algorithms are also used more and more in political contexts, e.g. to influence voting behaviour during political elections.[5] Although, data-driven algorithms have the potential to improve user experience through personalization or by offering access to online communities, it has been shown that, in practice, surveillance capitalist platforms tend to promote hateful and divisive messages by design and as such, pose an existential threat to democracy and social cohesion.[2] The danger posed by these algorithms (and their owners) becomes even more evident, if one considers the definition of 'the art of using technological tools to alter the cognition of human targets' as 'cognitive warfare', e.g. by NATO.[6] Taking this definition literally, one can argue that, given the omnipresence of technological surveil-

lance today, e.g. through smartphones, microphones and cameras, human cognition and thought are under constant assault from surveillance capitalists and governments. Indeed, the human brain has been called the 'battlefield of the 21st century'.[6]

For now, the weaponized surveillance algorithms used to influence human cognition, mostly rely on behavioural data, e.g. clicks and likes. However, in many cases it can remain challenging to infer cognitive and affective processes from non-physiological data. For example, we can consider the case of a person watching or listening continuously to video or audio streams on a digital platform. This person might be highly attentive and engaged towards the stimuli, but might not exhibit any overt behavioural correlates of their appreciation. Time spent is commonly used as a surrogate for engagement, however, this is highly flawed, because a stream can just be running without anyone watching/listening. Therefore, surveillance capitalists often implement automated pop-up windows after long viewing times, so that users need to engage with the platform (e.g. click on a button) from time to time and generate explicit feedback that is then used by the algorithms. In contrast, physiological computing could allow the implicit extraction of relevant mental states from the user without requiring explicit user feedback.[7]

Unsurprisingly, companies and governments have started to realize the potential of physiological computing and BCI which is evident in their increasing economic investments (estimated around \$33.2 billion in 2021 [8] regarding such technologies.[9]

Given the increasing popularization of wearable physiological computing devices that are often embedded in surveillance capitalist ecosystems, e.g. Apple's walled garden which already includes a patent application for wireless in-ear headsets with integrated EEG electrodes [10], it seems highly likely that in the near future (neuro-)physiological data will be commonly exploited by surveillance entities in order to fine-tune their algorithms for nefarious purposes. Therefore, ethical reflections on future use of physiological computing systems in general and BCIs in particular are critical in order to defend individual humans and democratic societies from the threats posed by surveillance capitalism.

To this end, we propose an adversarial approach inspired by ethical hacking in order to hack BCI ethics. Ethical hacking is a concept from cybersecurity where an agent

plays the role of an unethical adversary and tries to discover system vulnerabilities in order to be able to better protect the system against real attacks. Indeed, simulating attacks during authorized penetration testing of computer systems by ethical hackers has become a highly important technique to mitigate security risks. [11]

Analogous to ethical hacking in cybersecurity, we propose that adopting adversarial roles can be beneficial to mitigate ethical risks that arise from Neurotechnology. First, we will begin with a brief overview of the state-of-the-art of BCI ethics. Afterwards, we will describe the framework of hacker ethics that constitute the ethical foundation of this work. Then, we will define the general methodology of applying ethical hacking to BCI ethics. Next, we will illustrate our approach in a hypothetical case study by applying this method on a type of passive BCI that is currently developed and of particular interest to surveillance capitalists and authoritarian governments. After elaborating on the ethical risks posed by this BCI system, we will also discuss how these risks could be patched in order to protect its users.

ETHICAL BACKGROUND

BCI Ethics:

The emergence of BCI technology is accompanied with ethical risks concerning individual users and societies as a whole. Although, until now, the development of BCI ethics has been outpaced by its commercialization (Boonstra, 2025), the need for ethical reflection of BCI has been recognized (see [12] for a review on BCI ethics). Ethical issues in BCI concern concepts such as privacy [13], agency[14], inclusivity[15], among others.

Various techniques to identify and assess ethical risks arising from BCI technology have been developed (e.g. [13], [16]). Notably, Krol & Zander also proposed an adversarial approach to BCI ethics in which they call for the development of unethical BCI systems under controlled and ethical laboratory conditions.[1] While we believe that this proposal is promising and should be further explored, it suffers from two major limitations. First, this method does not allow to explore the potential ethical risks of technology that is not quite ready to be implemented yet. However, potential ethical risks should ideally be considered before a technology can be physically realized. Secondly, some BCI systems might be too unethical for researchers to get approval from an ethics committee and test them in laboratory conditions, e.g. if they are used for torture. In the following, we will present an alternative adversarial method inspired by ethical hacking to be used for ethical thought experiments regarding BCI.

Hacker Ethics:

Although, originally ‘hacker’ used to refer to creative tinkers in a more general way[17], it is nowadays more commonly used to describe actors that access or exploit computer systems unauthorized.[18]

On an individual scale, such actors can be malicious, e.g. so-called black hats[11], however, hacker culture in general is closely aligned with liberalism. Although, there is no unitary code of hacker ethics, liberal ideas, such as freedom, privacy, equal opportunity, meritocracy, etc., can be considered general moral commitments for many hackers.[19]

As such, hackers are in direct ideological opposition to surveillance capitalists and authoritarian governments.¹ Indeed, hackers and hacker organizations, such as the Chaos Computer Club in Germany² are some of the most prolific critics of surveillance capitalism and authoritarian governments.

This paper follows in the liberal tradition of hacker culture and its ethical values. Unfortunately, these values are currently under attack from surveillance capitalists in cahoots with authoritarian government.[3] In the next section, we will, thus, discuss how principles from ethical hacking can be used to defend these values and ourselves against this assault in BCI.

ETHICAL HACKING FOR BCI ETHICS

General Method:

This section describes the general method to hack BCI ethics. This method consists of the following steps that will be further described in the next paragraphs: 1. Designing a BCI system. 2. Embedding this system in a social context. 3. Adopting an adversarial role to find ethical exploits in this system. 4. Develop ethical patches for this system.

BCI System 1. Designing the System:

In order to hack BCI ethics, we first need a BCI system to hack. However, at least for now, few consumer BCI system that function reliably exist. As such, it remains challenging to truly hack BCIs currently.

Nonetheless, novel technologies emerge from existing economic and political systems, and as such, we can hypothesize how hypothetical BCIs could be maliciously exploited by unethical actors in these systems. Moreover, a plethora of science fiction media include fictional BCIs[21], and can be used as inspirational source in order to design hypothetical BCI systems to hack.

Although, this reliance on BCI thought experiments might seem like a limitation, flaws in a product should ideally be fixed before its public release. Therefore, it is important that we³ consider the ethical risks of BCI, before the technology becomes widely adopted with a high risk of unethical use.

Still, even hypothetical BCI systems need to be relatively realistic in order to be useful and should be neither logically nor physically (according to current scientific un-

¹Unsurprisingly, hackers are mostly portrayed negatively in mainstream media.[19]

²Not to be confounded with the Chaos Computer Club in France which was a false flag operation to survey hackers by security agencies.[20]

³Especially those of us who develop BCI.

derstanding) impossible. However, a BCI system does not necessarily need to work in order to be used unethically, e.g. as is the case with polygraphy-based lie detection in some countries, e.g. USA.[22]

2. Example: Given the worldwide loneliness epidemic and general disappointment with the current state of dating apps, we can imagine that a BCI dating system could be used to optimize romantic matches based on neural markers of interest, arousal, etc.

Embedding the BCI System 1. Technology in context:

Technology does not exist in a neutral void, but is always embedded in political and economic, and thus, ethical, contexts. Therefore, in order to be useful, our hypothetical BCI system also needs to be embedded in such contexts.

2. Example:

Given the increasing popularity of far-right politicians worldwide, we could, for example, imagine how a BCI system could be exploited in a (post-)fascist regime.

Adopting an Adversarial Role 1. Ethical Exploits:

Given a BCI system and a social context, we can then consider which unethical actors would want to use this BCI system and what their aims, e.g. political, economic, etc., are. Afterwards, one can adopt the role of such an unethical actor and consider how one would exploit the BCI system from their perspective. Here, we can also take inspiration from the ethical vulnerabilities of already existing technology. For example, although BCIs are still a very niche technology, smartwatches have become quite common and many of these devices record physiological data such as cardiac activity which can contain sensitive personal information.[23]

2. Example:

Given access to neural data and the ability to match potential romantic partners, we can imagine that a fascist regime could use neural data to identify neuro-divergent users for eugenic purposes.

Ethical Patches 1. Fixing ethical vulnerabilities:

As with computer vulnerabilities, discovered ethical vulnerabilities should also be patched. Unfortunately, patching ethical vulnerabilities can be even more complicated than patching software, because ethical exploits often concern societal systems as a whole. Still, similarly to ethical hackers in cybersecurity, we think that it is crucial to disclose the discovered vulnerabilities as a first step to develop patches. Only when one (and ideally, society as a whole) is aware that a flaw exists, can one start to fix it. As such, transparency, communication, regulation and education are often critical to patch ethical risks. Indeed, the various ethical patches proposed in this article are deeply intertwined. Transparent technologies permit educated users to better apply their critical reasoning skills, as well as vote for politicians proposing ethical laws and regulations (at least in liberal democracies).

2. Example:

Fixing ethical vulnerabilities in a fascist society can be more than complicated and pretty dangerous. From a liberal perspective, fascist societies are inherently unethical, as such we could imagine liberal hacktivists to engage in cyberattacks against such a BCI used for eugenics, e.g. by spoofing fake brain signals in order to confound its processing algorithms and protect the users. After describing the general method to hack BCI ethics, we will now illustrate it with a case study.

Case Study:

BCI System In this case study, we will describe a hypothetical BCI system whose real-world adoption by consumers would come with severe ethical risks. As we already briefly mentioned in the introduction, passive BCIs could improve user experience by personalizing these experiences, e.g. by personalizing media or product presentation based on a user's neural states. This BCI would measure neural correlates of user engagement, e.g. attentional markers such as neural alpha activity, in order to decode a user's appreciation.[24] This BCI system would then function similar to a Like button, but without requiring explicit user behaviour and be used to train algorithms and make predictions.

Embedding the System in a socio-economic context

Given the current surveillance capitalist economic system in which data are precious commodities, we believe that such hypothetical like decoding BCI systems are probable to be developed in the near future. Therefore, we embed this hypothetical BCI in our current human state of affairs, i.e. we consider how such a technology could and likely would be unethically used if it would arise in our current world.

Adopting an adversarial role 1. Unethical Actors:

Various parties could profit from the unethical use of the hypothesized user experience decoding BCI system and the personal data it records. Potential bad actors range from petty criminals to organized crime, as well as companies and governments. Although, regular criminals could be interested in personal data collected by this BCI, e.g. in order to blackmail a victim for based on neural markers of sexual orientation, we believe that, in the current geopolitical context, surveillance capitalist companies and authoritarian governments pose deeper ethical risks, due to their scale and global power. In particular, surveillance capitalist and governments are typically funding, developing, providing and regulating new technologies and as such have some moral responsibility in their regards⁴. Therefore, we will focus on how these entities could exploit a user experience decoding BCI for their unethical purposes.

2. Hacking:

⁴As do researchers working on these technologies.

In this section, we will, now that we have adopted the role of unethical surveillance capitalist and governmental actors, elaborate on possible ethical vulnerabilities for these actors to exploit if our fictional user experience decoding BCI would exist already. Both types of actors are already heavily exploiting personal data for economic and political gains. Although, purely behavioural data can already be sufficiently informative for many purposes, physiological data from various sources, e.g. health records, fitness trackers, etc. are being more and more integrated in surveillance capitalist data pipelines.[25] As such, most of the ethical risks arising from user experience decoding BCIs are not completely new, but they complement and enhance the ethical risks that are already present in our contemporary world.

In the following, we will discuss a few specific ways how a user experience decoding BCI could be used by surveillance capitalists and unethical governments to manipulate and repress humans.⁵

a) Manipulation:

As mentioned in the introduction, personal data is used by surveillance capitalists and governments to influence human cognition and behavior through the algorithmic micro-targeting of individuals. We imagine that a user experience decoding BCI measuring neural correlates of personal likes and dislikes could improve micro-targeting, by optimizing how personalized stimuli presentation and messaging affects a user's brain. Next to capitalist interest of selling products to consumers, this BCI system could also be used for propaganda and political influence, e.g. by using its data in micro-targeting to affect the outcome in political elections. Furthermore, surveillance capitalist algorithms have been used to promote hate and violence, e.g. on Facebook in the case of the genocide in Myanmar[26], and we can imagine that access to neural data could be used to optimize the evoked negative emotions, e.g. dislike, and consequently, evoke even more violent behavior in the future.

b) Repression:

Besides manipulation, a user experience decoding BCI could also be used in repressing 'undesirable' humans and ideas. For example, neural responses evoked by stimuli featuring attractive persons of different genders can be informative about sexual orientation[27] which could then be used to optimize the existing repression of LGBTQ+ persons who in some countries still risk severe punishment, e.g. the death penalty. Moreover, brain activity evoked by political stimuli can similarly be informative about a user's political beliefs.[28] Such information could then be used by companies to repress activists or unions, or by authoritarian governments to repress political opposition. Nowadays, USA immigration is already checking immigrant social media and personal computing devices and negotiating access to biometric

data from EU citizens. As such, we can also imagine future immigration to include neural data collected from user experience decoding BCIs in their assessments.

Even more extremely, personal (meta-)data and surveillance capitalist technologies are already used in AI-assisted targeting to determine human targets in military contexts, e.g. currently by the IDF in Gaza.[29] Such targeting systems assign a numerical value to individuals, often based on all kinds of (mostly) correlational data and meta-data. Then, an arbitrary threshold is defined and each individual above this threshold is considered a legitimate military target by the system. It seems plausible that neural data would also be used by such systems. And that, for example in the context of the war in Gaza again, neural data relating to liking innocent objects, e.g. water melons which have been adopted a pro-Palestinian symbol, could increase one's correlational score determined by an AI-assisted targeting system. And therefore, the probability to be killed in a military attack.

Finally, we further hypothesize a user experience decoding BCI could be unethically used to personalize torture methods and optimize the amount of harm caused in the victim. Various art stimuli have been used for torture prisoners in the past, e.g. surreal art during the Spanish Civil War, or heavy metal music by the USA in Guantanamo Bay. As a working user experience decoding BCI could also decode negative user experience, e.g. dislike, the technology could be abused to optimize the stimuli presentation in order to cause the most suffering in the tortured victim.

Ethical Patches The first step to patch the ethical flaws discussed above, is to disclose their existence. Consumers in particular need to be made aware of the ethical issues of technology they use, so they can make an informed decision whether they want to use it or not. This is especially important, as both governments and companies increasingly implement opt-out instead of opt-in data sharing procedures which does not allow for informed consent, e.g. in the case of the highly vulnerable electronic health records in Germany or AI-training in Whatsapp respectively, and prey on the uninformed and disinterested.⁶ Instead, consumers need to be educated and informed about the power of technologies and how they are embedded in current power structures. Many BCI researchers work in academia and are educators as well. As such, the community is in a privileged position to promote critical reasoning skills in face of technological innovation. Furthermore, in order to truly understand technology, it can not be a black box. Therefore, data processing pipelines need to be transparent and interpretable, so that consumer know what is done with their personal data.

Finally, regulations and legal frameworks can be used to protect consumer rights (if they are enforced with enough deterrence effect), e.g. data ownership and pri-

⁵This is not meant to be an exhaustive list of all possible ethical exploits of this BCI system, and we encourage interested readers to themselves adopt an adversarial role and search for further ethical exploits that we have missed.

⁶Unsurprisingly, electronic health data is also planned to be easily and legally shareable with companies.

vacy in the case of passive BCI. Indeed, various governance frameworks for Neurotechnologies have been developed by regulatory and government organizations and researchers (see [30] for a scoping review). Notably, international bodies such as UNESCO[8] have developed ethical guidelines for neurotechnology. Furthermore, legal acts, such as the EU AI Act, are currently being adopted by governments in order to regulate AI systems, including BCI, and protect their users. Given the structural power imbalance between citizens and governments, as well as between consumer and companies, we believe that legal protection of consumers and civil rights is of utmost importance in order to protect individuals against unethical applications of BCI technology. Still, (over-) regulation of BCI could stifle urgent technological innovation, e.g. medical treatments, and as such, consensus on BCI governance remains inexistent.[30]

DISCUSSION

In this paper, we propose adopting the role of unethical actors for thought experiments in BCI ethics in order to defend against ethical exploits.

Although, we are aware that bad actors could potentially be inspired by our hypothetical examples and use them to exploit BCI systems in the real world, security by obscurity is generally not considered adequate protection.[31] In contrast, ethical hacking approaches have been widely embraced by the cybersecurity community in order to defend systems against malicious attacks. Similarly, we believe that concepts from ethical hacking in cybersecurity can be used to inform BCI ethics. By taking the perspective of unethical actors, e.g. surveillance capitalists, we can investigate how and for which purposes BCI technology could be exploited. This knowledge about potential ethical vulnerabilities can then be employed to develop safeguards for users and promote the development of ethical BCI systems. Therefore, BCI education and dissemination are crucial for consumers to be aware about its potential risks. As such, we can consider the existence of a moral duty for the BCI community to inform the public about potential benefits and risks of the technologies they develop. However, given the existing power structures in which technologies are embedded, knowledge alone might not be sufficient to protect their users. In these cases, legal protections need to be passed and enforced to discourage bad actors from exploiting a technology. However, as legislative bodies and surveillance capitalists often share common interests, e.g. the surveillance of consumers and citizens, the effectiveness of BCI regulations can be questionable. For example, although the EU AI Act grants some protection to consumers, many loopholes that could be exploited unethically remain.[32] Furthermore, the recently proposed EU Digital Omnibus closely aligns with big tech lobbyist positions and weakens some of the previously established legal protections and widens the allowed exploitation of personal data, including for AI training.[33]

Given that the main bad actors in surveillance capitalism are governments and powerful companies, it can seem hopeless to try and defend against the cognitive war they are conducting against us. However, not all is lost yet. For now, it is still possible to criticize unethical behaviour from both companies and governments and vote for politicians one considers ethical in elections (at least in most EU countries) which can result in laws one would consider ethical. For example, one can still have hope that large-scale activism has contributed towards the current rejection of the proposed EU chat control which would have severely weakened consumer cybersecurity under the guise of ‘child protection’ by EU parliament.[34] This suggests that political actions such as raising awareness through dissemination can still be used to protect against attacks on liberal values, such as privacy.

CONCLUSION

Given the potential public adoption of BCI in the near future, we are faced with potential existential ethical risks that threaten liberal values, democracy and even human lives. In a world where powerful actors have great financial and political interest to use BCI unethically and exploit user brain data for their own goals, e.g. manipulation and repression, it is crucial to determine how such actors could profit from such use, ideally before consumers get caught in exploitative structures, e.g. surveillance capitalist ecosystems. To this end, we proposed an ethical hacking approach towards BCI ethics which takes an adversarial perspective to discover ethical vulnerabilities in BCIs from the point of view of bad actors and patch these flaws.

REFERENCES

- [1] Krol LR, Zander TO. Create no evil? a call to cross ethical boundaries. In: 3rd Neuroergonomics Conference. 2021.
- [2] Dawson J. Who controls the code, controls the system: Algorithmically amplified bullshit, social inequality, and the ubiquitous surveillance of everyday life. *Sociological Forum*. 2023;38.
- [3] Zuboff S. The age of surveillance capitalism (. 2019).
- [4] Durand C. *Techno-féodalisme*. La Découverte: Paris (2020).
- [5] Trauthig I, Valenzuela S, Howard P. The role of generative ai use in 2024 elections worldwide. 2025.
- [6] Claverie B, Cluzel F du. The cognitive warfare concept. 2022.
- [7] Welter M, Bouneau A, Lotte F, Ward T. Towards curating personalized art exhibitions in virtual reality with multimodal electroencephalography-based brain-computer-interfaces. *Art & Perception*. 2023;11(2).
- [8] UNESCO. The risks and challenges of neurotechnologies for human rights. 2023.
- [9] Drinkall J. Capitalist telepathics, psychic debt and the search for collective intelligence. *TOPIA*. 2022;45.

- [10] Azemi E et al. Biosignal sensing device using dynamic selection of electrodes. US Patent Application Number: US 2023/0225659 A1. 2023.
- [11] Yaacoub JPA, Noura HN, Salman O, Chehab A. A survey on ethical hacking: Issues and challenges. 2021. arXiv: 2103.15072 [cs.CR]. [Online]. Available: <https://arxiv.org/abs/2103.15072>.
- [12] Lang A, Coin A, Dubljevic V. A scoping review of the academic literature on bci ethics. Apr. 2024.
- [13] Mecacci G, Haselager P. Identifying criteria for the evaluation of the implications of brain reading for mental privacy. *Science and Engineering Ethics*. 2019;25(2).
- [14] Haselager P. Did i do that? brain-computer interfacing and the sense of agency. *Minds and Machines*. 2013;23(3).
- [15] Livanis E, Voultsos P, Vadikolias K, Pantazakos P, Tsaroucha A. Understanding the ethical issues of brain-computer interfaces (bcis): A blessing or the beginning of a dystopian future? *Cureus*. 2024;16.
- [16] Mecacci G, Haselager P. Five criteria for assessing the implications of nta technology. *AJOB neuroscience*. 2019;10(1).
- [17] Levy S. Hackers: Heroes of the computer revolution. Doubleday: New York (1984).
- [18] Bellaby R. An ethical framework for hacking operations. *Ethical Theory and Moral Practice*. 2021;24.
- [19] Coleman E, Golub A. Hacker practice: Moral genres and the cultural articulation of liberalism. *Anthropological Theory - ANTHROPOL THEORY*. 2008;8.
- [20] Warusfel B. Contre-espionnage et protection du secret : Histoire, droit et organisation de la sécurité nationale en france. Panazol: Lavauzelle (2000).
- [21] Allison B. Why bci-fi? In: *Proceedings of the 10th International Brain-Computer Interface Meeting*. 2023.
- [22] Iacono W, Ben-Shakhar G. Current status of forensic lie detection with the comparison question technique: An update of the 2003 national academy of sciences report on polygraph testing. *Law and Human Behavior*. 2018;43.
- [23] Sifaoui A, Eastin M. “whispers from the wrist”: Wearable health monitoring devices and privacy regulations in the u.s.: The loopholes, the challenges, and the opportunities. *Cryptography*. 2024;8:26.
- [24] Welter M, Lotte F. Ecological decoding of visual aesthetic preference with oscillatory electroencephalogram features—a mini-review. *Frontiers in Neuroergonomics*. 2024;5.
- [25] Gidaris C. Surveillance capitalism, datafication, and unwaged labour: The rise of wearable fitness devices and interactive life insurance. *Surveillance & Society*. 2019.
- [26] International A. The social atrocity: Meta and the right to remedy for the rohingya. London, 2022.
- [27] Ziogas A et al. Deep learning in the identification of electroencephalogram sources associated with sexual orientation. *Neuropsychobiology*. 2023;82.
- [28] Galli G et al. Early eeg responses to pre-electoral survey items reflect political attitudes and predict voting behavior. *Scientific Reports*. 2021;11:18692.
- [29] Dorsey J, Bo M. Ai-enabled decision-support systems in the joint targeting cycle: Legal challenges, risks, and the human(e) dimension. *International Law Studies*. 2026;107(1).
- [30] Robertson L, Higgins NL, Maier MJ, Carter A, Gardner J. Ethical governance strategies for the responsible innovation of neurotechnologies: A scoping review. *Journal of Bioethical Inquiry*. 2025;22.
- [31] Scarfone K, Jansen W, Tracy M. Guide to general server security. recommendations of the national institute of standards and technology. 2008. [Online]. Available: <https://csrc.nist.gov/pubs/sp/800/123/final>.
- [32] Wachter S. Limitations and loopholes in the eu ai act and ai liability directives: What this means for the european union, the united states, and beyond. *SSRN Electronic Journal*. 2024.
- [33] Observatory CE, LobbyControl. Article by article, how big tech shaped the eu’s roll-back of digital rights. 2026. [Online]. Available: <https://corporateeurope.org/en/2026/01/article-article-how-big-tech-shaped-eus-roll-back-digital-rights>.
- [34] Krempel S. Setback for the commission: Eu meps let chat control fail. *Heise Online*. 2026.

Low-Power and Compact Time-Domain Analog Feature Extraction for ECoG Motor Intent Decoding

Ilan Jaoui¹, Joe Saad², Emmanuel Hardy¹, Lucas Struber³, Tetiana Aksenova³, Franck Badets¹

¹Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France

²Univ. Grenoble Alpes, CEA, List, F-38000 Grenoble, France

³Univ. Grenoble Alpes, CEA, Leti, Clnatec, F-38000 Grenoble, France

E-mail: ilan.jaoui@cea.fr

ABSTRACT: Power consumption has become a key constraint in designing implantable Brain-Computer Interfaces (BCIs). For systems with a large number of channels, analog-to-digital conversion and Feature Extraction (FEx) are often the most power-hungry stage of the decoding pipeline. To address this bottleneck, we present a BCI processing chain using an ultra low-power compact analog Time-Domain (TD) Band-Pass Filter (BPF) bank to extract features from ElectroCorticoGraphy (ECoG) signals. First, we show that the frequency spread of TD BPF have little impact on the decoding accuracy. The proposed analog features are then assessed on a 5-class motor decoding task and achieve a decoding performance well above chance level, indicating that these features preserve discriminating information for multi-class decoding. The same features are then used for binary motor/idle detection task and performed as well as digital features, while consuming $\times 646$ less power, and $\times 75$ vs analog counterparts. In this setting, we also show that our system maintains robust accuracy even at high noise levels, highlighting a favorable trade-off when using lower power Analog ReadOuts (AROs). This paper overall demonstrates that TD Analog Feature Extraction (AFEx) can provide a good power-performance trade-off for binary tasks in low-power implantable BCI systems.

INTRODUCTION

Recent progress in BCI decoding methods has enabled increasingly powerful and complex systems that can accurately detect and decode motor intent for patients after spinal cord injury [1, 2]. These advances pave the way towards complete rehabilitative systems that could restore patients' autonomy. However, such demonstrations often focus on improving decoding accuracy and task complexity, rather than on constraints that are central to day-to-day use in a rehabilitative context, such as miniaturization and power reduction. This gap becomes critical for portable and implantable systems, where the power budget is limited. Integrating the hardware constraints early in the design process is therefore crucial to build practical systems for the patient.

Previous works have explored low-power techniques for BCI design [3], either by optimizing existing algorithms

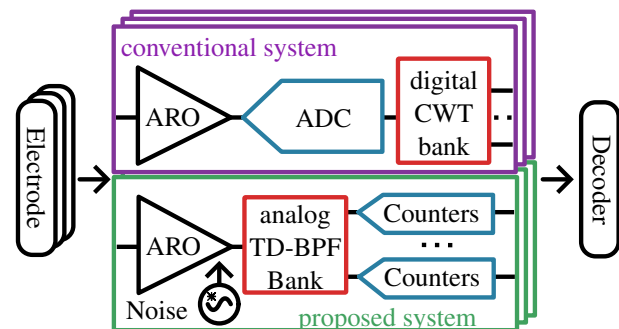


Figure 1: (a) Conventional BCI FEx Pipeline Compared to (b) Proposed Analog TD BPF Bank

and platforms, or by proposing custom Application Specific Integrated Circuit (ASIC) accelerators to reduce power. Such systems have been used for epileptic seizure detection [4], speech [5] and motor decoding [6].

Among invasive neural recording techniques, ECoG offers a trade-off between resolution and invasiveness, while providing stable recordings over time [7] which is required for long-term BCI implants.

In ECoG-based motor applications, motor decoding often relies on detecting changes in activity within specific frequency bands, making spectral representations attractive for such systems [3]. Extracting spectral features through digital signal processing can be done either in an external computer or directly in the implant [3]. This step can be costly in hardware, as it typically relies on Multiply-Accumulate (MAC) operations and memory accesses across channels and bands [8]. This motivates exploring alternatives such as analog approaches for lower-power operation.

Existing approaches reduce power by applying simplifications or by selecting subsets of inputs [9]. However, they remain tied to a continuous FEx mechanism that still consumes a constant amount of power. The AFEx approach, seen in [10] allows for a flat power reduction by extracting features before digitization. This reduces the amount of information to be sampled, as gestated in [11] and is therefore more power and silicon efficient. But [10] still uses regular analog filters, which require large biasing currents, resistors and capacitors to mitigate noise and process variations. This paper focuses on two

complementary directions to demonstrate potential power reduction in BCI processing. First, it introduces a parallel AFEx approach, a compact ultra-low-power TD BPF bank adapted from [12] and based on Injection-Locked Oscillators (ILOs) to estimate the per-band energy over a given time window as features, and compare their performance to Continuous Wavelet Transforms (CWTs) commonly used for BCIs using a baseline classifier. Then, the impact of different noise levels on the AROs on the system accuracy is studied to understand how power can be traded off with noise. The simulated ARO and FEx module are evaluated on real ECoG datasets.

MATERIALS AND METHODS

We evaluate the proposed low-power FEx framework in a realistic simulated setting. Fig. 1 summarizes the two processing flows considered in this work. In the baseline State-of-the-Art (SoA) system, ECoG signals are digitized after amplification by the ARO and features are extracted using digital processing. In our proposed pipeline, spectral features are extracted directly in the analog domain using a TD BPF bank, and only the resulting feature streams are digitized by accumulation. In both flows, the extracted features are fed to a SoA decoder for classification. The systems are evaluated using the same datasets and decoders, with different levels of noise.

FEx:

Most BCI decoding pipelines rely on intermediate representations, also called features, extracted from neural signals before applying a classifier or regressor, commonly known as the decoder. In practice, features summarize signal properties, such as spatial, spectral or temporal characteristics. In this work, we focus on spectral representations extracted from ECoG input signals. For our experiments, features were computed on either 100 ms (8DoF, Runner) or 80 ms (BCI-IV) time frames.

CWT Features: Wavelet-based representations are attractive for brain signal decoding, since wavelets act as time-localized band-pass filters. Moreover, the CWT provides flexible digital filters that can extract any set of frequency components, and this method has often been used for BCI FEx [13, 14].

The CWT relies on a convolution between an input signal and a wavelet. Given a discrete time signal $x[n]$, its CWT component at scale a and time shift τ is defined as:

$$W(a, \tau) = \sum_n x[n] \psi^* \left(\frac{n - \tau}{a} \right), \quad (1)$$

where ψ^* is the complex conjugate of a mother wavelet scaled by a factor of a to obtain a desired frequency component. To obtain features for BCI decoding, we can use the time-frequency coefficients to compute the mean energy at a frequency f_a within a given time window [2].

$$E_{a,T} = \frac{1}{|T|} \sum_{\tau \in T} |W(a, \tau)|^2, \quad (2)$$

where T denotes the time coefficients within the defined extraction window. In this work, the CWT-based energy

features are used as the reference digital baseline, against which we compare the proposed time-domain features extracted using the ILO filter bank.

Analog TD BPF Features: In this paper, a model of analog BPFs based on ILOs, as described in [12], is used. When locked to a reference clock ref , ILOs are forced to run at the ref frequency with a phase shift ϕ proportional to the difference between the natural frequency and the ref frequency. If modulated with an input voltage V_{in} , ILOs change their natural frequency, thus resulting in a voltage-to-time converter of DC gain G_{ILO} , with a low-pass filter property of cutoff frequency f_c . Embedding the ILO in a feedback loop with an integrator of gain G_{INT} results in an ultra-low power, low silicon area, bandpass phase modulated signal at injected signal frequency, of transfer function given in Eq. 3.

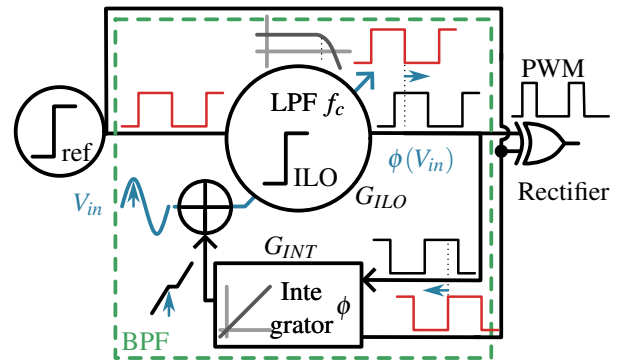


Figure 2: Block diagram of a TD Based ILO BPF and their associated transfer functions representation with the impact of the input signal in blue and the ref clock in red

$$H(s) = \frac{\phi}{V_{in}} = \frac{G_{ILO} \cdot s}{G_{INT} \cdot G_{ILO} + s + \frac{s^2}{2\pi f_c}} \quad (3)$$

The rectified output pulse width of the XOR gate is over-sampled and averaged to estimate the input signal energy in each spectral band. Here, the filter is simulated by discretizing Eq. 3 with a generalized bilinear transform.

To extract multi-band spectral features using these filters in a bank, a logarithmic scale was adopted, which simplifies design by keeping the quality factor constant thus limiting overlap between close bands.

ARO Noise Model:

ARO of BCIs prepare and amplify neural signals so that they can be processed by the FEx stage. They are typically designed to maintain high signal integrity, which necessitates high biasing currents. To reduce power, the signal degradation/power consumption trade-off is explored by introducing increasing levels of noise, and monitoring the decoding performance. To capture this trade-off with a simple model, we consider the input-referred noise of our system to be the sum of a white/thermal term, and a low-frequency/flicker term following a $1/f$ trend as shown in Eq. 4, which matches classical expected noise profiles at low frequencies.

$$\overline{V_i^2} = \overline{V_{thermal}^2} + \overline{V_{flicker}^2} \quad (4)$$

Based on analog amplifier simulations, the overall noise is dominated by flicker noise, as thermal noise density was estimated at $25 \text{ nV}/\sqrt{\text{Hz}}$. To fully cover the range of a typical ECoG signal Power Spectral Density (PSD), the flicker noise will vary from 0.46 to $29.8 \mu\text{V}_{\text{rms}}$ as shown in Fig. 3, plotted here from an electrode average over a session of the Runner dataset. For reference, the estimated noise level from the CINESIC ARO extrapolated noise profile [15] was also drawn. Using Fig. 3, when increasing noise, it can be expected that high frequency oscillations of the signal would be first degraded.

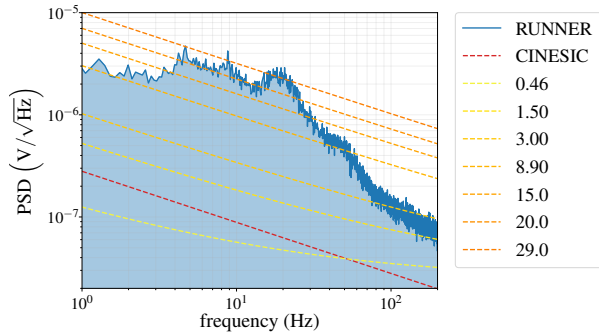


Figure 3: PSD of Simulated ARO Noise labelled in μV_{rms} compared to CINESIC [15] Noise and ECoG Signal Density

The original datasets were acquired with low-noise front-ends (CINESIC [15], SynAmps2 [16]). Their noise contribution will be considered negligible with respect to the noise added in our experiments.

Reference Decoder:

To isolate the impact of the FEx step on decoding metrics, we use a SoA decoder, previously used in a BCI, as baseline to show the impact on different FEx configuration. Classification tasks are the sole focus of this study, thus regressive decoders are omitted.

The model used is an Markov Switching Linear Model (MSLM) that accounts for the temporal structure of the features and state transitions [2]. At each step, the model computes the class likelihood from the feature vector using a linear mapping, followed by a softmax function to obtain observation probabilities. These observations are then combined with a Hidden Markov Model (HMM) with transition probabilities. Decoding is performed by estimating the most probable current state using the posterior. In practice, this temporal prior improves robustness in asynchronous decoding by stabilizing the state transitions. The model is composed of $9600 \times N$ parameters with N the number of classes, which are trained in pseudo-online using a Recursive Exponentially Weighted N-way Partial Least Squares (REW-NPLS) algorithm [17], which incrementally updates the linear model with new data. We used for training the two recommended sessions of the Runner and 8DoF datasets (amounting to respectively 12% and 10%). We collected the accuracy, recall, precision and F1 scores for each session to provide an average and standard deviation over the sessions. For BCI-IV, we followed the original split

yielding a single score.

Benchmark Datasets:

The datasets used in this paper were recorded during Motor Imagery (MI) or Motor Execution (ME) tasks.

8-Degrees of Freedom (8DoF) Virtual Avatar: The first dataset was obtained from a male participant with chronic traumatic quadriplegia following a complete C4-C5 spinal cord injury, and was previously used to develop an adaptive decoder [2]. It was recorded as a part of the “BCI and Tetraplegia” (ClinicalTrials.gov identifier: NCT02550522) clinical trial, approved by the Agency for the Safety of Medicines and Health Products and the ethical Committee for the Protection of Individuals. Two fully implantable WIMAGINE epidural devices were placed over the upper-limb sensorimotor cortices, each integrating a 64-electrode grid [18]. During the virtual-avatar experiment, the participant controlled a first-person avatar and performed motor-imagery commands (hand translation and wrist rotation), providing labelled data for a 5-class decoding task.

Runner 2-Class Task: We also took the Runner ECoG dataset recorded with the same WIMAGINE epidural implants and participant, to evaluate the decoding of binary motor intent. In this task, the participant viewed a third-person virtual avatar and was instructed to alternate between *walk* and *stop* commands. The avatar either walked at a fixed speed or remained stationary [2]. The recordings comprise approximately two hours of data acquired over multiple short sessions of 10 minutes each.

BCI Competition IV - 4: We also evaluated our approach on the BCI-IV dataset, an open-source ECoG dataset recorded from three epileptic patients performing finger motor execution [16]. Each trial is cued by a displayed word indicating which finger to flex, and signals are recorded from 62, 48, and 64 electrodes depending on the subject. This dataset provides a reproducible benchmark to assess robustness across patients, electrode layouts, implantation method, and sampling hardware.

Labels are provided as continuous 1D finger position traces rather than discrete classes. We convert them into one-hot encoded movement labels as shown in Fig. 4. The first step detects movement, the second avoids periodic Idle label during finger flexion, and the third monitors involuntary finger overlaps and either splits the uncertain labels in two if the overlap is contiguous, or removes it if one fully covers the other. For binary experiments, any detected finger movement is treated as a ‘Motor’ state, and the absence of movement as ‘Idle’.

Since the ECoG signals units are not specified, to keep consistence with the other datasets and to simplify the noise injection stage, a global scaling factor of $3 \times 10^{-9} \text{ V}$ per step was applied, thus considering the amplifier [16] to be in AC mode as described in its documentation.

Estimated Power Savings vs SoA:

To demonstrate the potential power savings of the whole pipeline, we will compare the proposed ARO and FEx improvements with SoA systems.

FEx: [12] uses the same AFEx studied here for au-

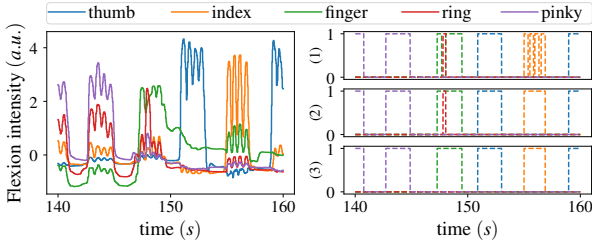


Figure 4: BCI-IV Label Cleaning Steps: (1) Movement Thresholding, (2) Gap Merging, and (3) Overlap Resolution

dio signals with 16 filters filters between 125 Hz and 5 kHz, and reports a power consumption of 0.820 μ W. ILO BPFs power scales with ref frequency, which will be on average $\times 20$ slower, if adapted to ECoG. Thus, the power per extracted feature would be around 3 nW. Only demonstrators extracting motor ECoG frequency band's energy reported in [3] will be studied. The average power consumption from [3] of ECoG based energy band FEx is 30 mW. This is highly due to outliers which are not based on ASIC implementations, but is still somewhat representative of the field. [6] uses digital FEx and consumes 295.4 μ W while extracting 64 features simultaneously, excluding decoding. The system using an AFEx approach [10] consumes 34.6 μ W for 32 features. This makes our system respectively $\times 4 M$, $\times 646$ and $\times 75$ more power efficient per extracted feature than its SoA counterparts.

ARO: To measure the impact of relaxing noise constraints, an extrapolation of the power/noise relation was done in Fig. 5 by harvesting ARO performances from [19]. It roughly follows an inverse power law of parameters mentioned in Fig. 5.

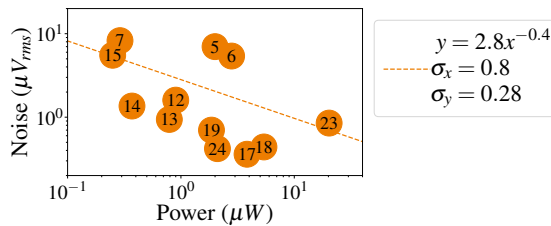


Figure 5: Power vs Noise Trend, with Citation # from [19]

RESULTS

We first evaluate FEx for multi-class decoding, by validating the impact of frequency spacing and then comparing analog BPF-based features to the digital CWT baseline. We finally focus on binary decoding, and evaluate the impact of AROs noise on the decoding accuracy.

Linear vs Log-Spaced Frequency Analysis:

As the analog filter bank extracts log-spaced frequencies, we preemptively evaluate the impact of frequency spacing using CWT-based features, to isolate its effect from the FEx method in Tab. 1. For that, we compare the linearly spaced baseline of 15 frequencies between 10 and

150 Hz in [2], to log-spaced configurations of 8 or 16 frequencies between 10 and 150 Hz or 200 Hz to evaluate the interest of higher frequency bands, used in [1].

Table 1: Frequency Impact on BCI Decoding (Runner Dataset)

Configuration	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)
linear (10-150Hz)	83.9 $\pm$ 5.8	81.5 $\pm$ 7.4	83.3 $\pm$ 5.4	81.6 $\pm$ 7.1
log-16 (10-150Hz)	84.8 $\pm$ 5.6	83.0 $\pm$ 6.8	84.1 $\pm$ 4.9	83.0 $\pm$ 6.5
log-8 (10-150Hz)	84.7 $\pm$ 5.6	82.3 $\pm$ 6.8	84.6 $\pm$ 4.6	82.6 $\pm$ 6.6
log-16 (10-200Hz)	84.6 $\pm$ 5.4	82.5 $\pm$ 6.9	84.0 $\pm$ 5.1	82.6 $\pm$ 6.5
log-8 (10-200Hz)	84.6 $\pm$ 5.8	81.9 $\pm$ 7.2	85.0 $\pm$ 4.5	82.3 $\pm$ 6.9

Across both the runner and 8DoF datasets, the different frequency configurations yielded similar performance distributions across sessions. As shown in Tab. 1, logarithmic provides a slight advantage over linear spacing on the runner dataset. Similarly, Tab. 2 shows that on the 8DoF dataset, the log-16 configuration slightly improved the decoding metrics. As a result, moving from linearly spaced to log-spaced frequencies has little to no impact on decoding performance, and the following results will compare the extraction methods.

Table 2: Frequency Impact on BCI Decoding (8DoF Dataset)

Configuration	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)
linear (10-150Hz)	79.9 $\pm$ 8.5	77.4 $\pm$ 8.8	75.3 $\pm$ 10.3	72.8 $\pm$ 11.6
log-16 (10-150Hz)	81.3 $\pm$ 7.5	78.2 $\pm$ 8.5	75.4 $\pm$ 10.4	73.9 $\pm$ 10.9
log-8 (10-150Hz)	80.1 $\pm$ 8.6	77.4 $\pm$ 8.8	75.1 $\pm$ 10.5	72.8 $\pm$ 11.6
log-16 (10-200Hz)	80.2 $\pm$ 8.3	77.4 $\pm$ 8.7	75.0 $\pm$ 10.5	72.9 $\pm$ 11.4
log-8 (10-200Hz)	79.3 $\pm$ 8.8	76.7 $\pm$ 9.2	73.9 $\pm$ 10.7	71.8 $\pm$ 11.9

Analog vs CWT Features:

We now compare analog log-spaced BPF features to the digital CWT baseline on 5-class and 2-class decoding.

8DoF dataset: Although decoding performance is degraded compared to the digital CWT baseline, all metrics remain significantly above chance level (20%), as shown in Tab. 3. The proposed analog BPF-based features thus preserve discriminative information in multi-class tasks, while loosing finer characteristics that allowed for higher performance classification. Moreover, the log-16 configuration shows a slight improvement compared to the log-8, highlighting non-redundant information across frequency bands. The intersession variance being higher also indicates that while they under-perform on average, there are still sessions where performances closely match. This indicates that further investigating may be needed to fully understand the mechanism behind this information loss.

Table 3: CWT vs Analog BPF BCI Decoding (8DoF dataset)

Configuration	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)
CWT	79.9 $\pm$ 8.5	77.4 $\pm$ 8.8	75.3 $\pm$ 10.3	72.8 $\pm$ 11.6
BPF log-16	69.2 $\pm$ 15.2	67.4 $\pm$ 12.3	68.9 $\pm$ 11.5	61.3 $\pm$ 15.6
BPF log-8	65.3 $\pm$ 16.2	65.6 $\pm$ 11.0	66.6 $\pm$ 12.7	57.9 $\pm$ 15.9

Runner Dataset: On the runner dataset, the analog BPF-based features achieve decoding performance closer to the digital CWT baseline, as shown in Tab. 4. This confirms that binary decoding can be achieved using the proposed analog features.

Table 5: Per-subject BCI-IV benchmark metrics for analog BPF configurations

Configuration	Subject 1 metrics				Subject 2 metrics				Subject 3 metrics			
	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)
BPF log-16	77.52	71.46	76.49	72.77	85.16	82.64	84.53	83.41	86.44	86.51	86.56	86.44
BPF log-8	74.32	70.73	71.42	71.03	84.56	81.92	83.89	82.72	87.96	88.06	88.22	87.95

Table 4: CWT vs Analog BPF BCI Decoding (Runner dataset)

Configuration	Acc. (%)	Rec. (%)	Prec. (%)	F1 (%)
CWT	83.9 ± 5.8	81.5 ± 7.4	83.2 ± 5.4	81.6 ± 7.1
BPF log-16	81.3 ± 6.9	79.4 ± 7.3	81.9 ± 6.3	79.0 ± 7.9
BPF log-8	80.6 ± 5.2	77.7 ± 7.6	80.6 ± 5.4	77.6 ± 7.4

BCI-IV Dataset: To generalize beyond the WIMAGINE implants, we repeated the same analysis on the all of BCI-IV subject’s. As shown in Tab. 5, analog BPF-based features enable above chance level binary decoding.

Noise robustness:

We then evaluate the robustness of AFEx by adding input noise to the analog front-end for both Runner and BCI-IV datasets. As illustrated in Fig. 6, the decoding accuracy remains relatively stable at low noise levels, with a progressive degradation as the $1/f$ noise amplitude increases. This trend is consistent across subjects and datasets, although the noise tolerance varies, indicating subject-dependent sensitivity that likely reflects differences in electrode coverage, signal quality, task type and engagement. Overall, these results support that the proposed TD AFEx remains usable under realistic front-end noise constraints and can enable power driven ARO design trade-offs across heterogeneous ECoG recordings.

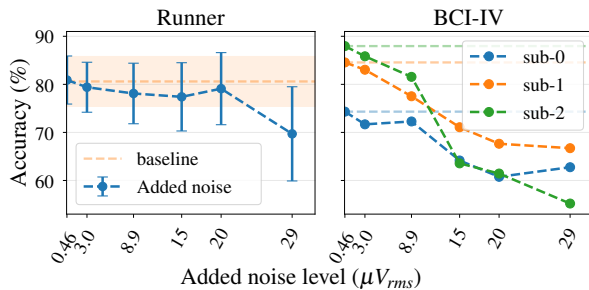


Figure 6: Accuracy of the Analog log-8 BPF Configuration Under Increasing Added $1/f$ Noise Compared to Baseline, with Runner’s Intersession Variance and BCI-IV 3 Subject’s Scores

DISCUSSION

Results show that log-spacing seem to offer a slight improvement ($\sim 1\%$) in decoding performance, which would make sense with classical brain wave sequencing [20]. The number of filter doesn’t seem to have a significant impact. The TD AFEx performs similarly as conventional digital features on binary classification tasks in a BCI flow, while being at least $\times 75$ more power efficient. The AFEx with the baseline decoder shows robustness to higher level of noise than measured in SoA AROs, meaning significant power savings can be made by relaxing noise requirements. Moreover, $1/f$ noise affects more

severely the signal high frequency components as shown in Fig. 3, meaning that these regions may not be crucial for binary motor/idle classification. By increasing the allowed noise 10 times, up to $\times 300$ power savings could be obtained based on Fig. 5 at the cost of a small performance degradation.

The generalizability of the decoding performances results is still to be shown by conducting the same experiments with an increased number of patients.

Given that about two third of the power of [6] is dedicated to the analog front end and FEx and the rest is allocated to decoding, this puts in perspective the potential savings at the scale of a whole BCI system.

CONCLUSION

In this paper, we propose an ILO-based filter bank to extract features from ECoG signals. The decoding showed similar performance as digital CWT features on binary motor intent decoding tasks. The features were also shown to be efficient for a 5-class decoding paradigm, exhibiting significantly above chance level accuracy. This approach consumes $\times 646$ less power than the best existing digital approach and $\times 75$ less power than analog counterparts. Finally, we studied the BCI pipeline robustness to noise on binary motor intent classification, and showed that if we can relax added noise level by $\times 10$ and save up to $\times 300$ power for a low performance cost. By combining these elements, we show the possibility of designing a competitive ultra-low power analog front end for motor intent detector.

ACKNOWLEDGMENTS

The authors thank all the members of the BCI project at CEA-Leti-Clinattec and Grenoble Alpes University Hospital for conducting the clinical trial and collecting the data used here. Research work presented in this paper was also carried out in the framework of the FAMES Pilot Line of the Chips JU, funded by the Horizon Europe grant 101182279 and the ANR NextGen project ANR-22-NEXTG-001 of the France 2030 initiative.

REFERENCES

- [1] H. Lorach et al., “Walking naturally after spinal cord injury using a brain–spine interface,” *Nature*, May 2023.
- [2] A. L. Benabid et al., “An exoskeleton controlled by an epidural wireless brain–machine interface in

- a tetraplegic patient: A proof-of-concept demonstration,” *The Lancet Neurology*, Dec. 2019.
- [3] J. Saad et al., “Comparison metrics and power trade-offs for BCI motor decoding circuit design,” *Frontiers in Human Neuroscience*, Mar. 12, 2025.
 - [4] S. R. Sridhara et al., “Microwatt Embedded Processor Platform for Medical System-on-Chip Applications,” *IEEE JSSC*, Apr. 2011.
 - [5] M. A. Shaeri et al., “33.3 MiBMI: A 192/512-Channel 2.46mm² Miniaturized Brain-Machine Interface Chipset Enabling 31-Class Brain-to-Text Conversion Through Distinctive Neural Codes,” in *2024 IEEE ISSCC*, Feb. 2024.
 - [6] U. Shin et al., “NeuralTree: A 256-Channel 0.227- μ J/Class Versatile Neural Activity Classification and Closed-Loop Neuromodulation SoC,” *IEEE JSSC*, Nov. 2022.
 - [7] C. Larzabal et al., “Long-term stability of the chronic epidural wireless recorder WIMAGINE in tetraplegic patients,” *Journal of Neural Engineering*, Sep. 2021.
 - [8] A. Antoniou, *Digital Signal Processing*, 2nd. McGraw-Hill, 2016.
 - [9] A. Z. Talha et al., “Applications of Brain Computer Interface for Motor Imagery Using Deep Learning: Review on Recent Trends,” *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 2 Feb. 28, 2024.
 - [10] O. Malekzadeh-Arasteh et al., “An Energy-Efficient CMOS Dual-Mode Array Architecture for High-Density ECoG-Based Brain-Machine Interfaces,” *IEEE TCAS*, Apr. 2020.
 - [11] M. Verhelst et al., “Where Analog Meets Digital: Analog-to-Information Conversion and Beyond,” *IEEE Solid-State Circuits Magazine*, 2015.
 - [12] A. Mostafa et al., “17.8 0.4V 988nW Time-Domain Audio Feature Extraction for Keyword Spotting Using Injection-Locked Oscillators,” in *2024 IEEE ISSCC*, Feb. 2024.
 - [13] H. K. Lee et al., “Application of Continuous Wavelet Transform and Convolutional Neural Network in Decoding Motor Imagery Brain-Computer Interface,” in *Entropy*, Dec. 2019.
 - [14] A. Chamanzar et al., “Efficient Hardware Implementation of Real-Time Low-Power Movement Intention Detector System Using FFT and Adaptive Wavelet Transform,” *IEEE TCAS*, Jun. 2017.
 - [15] S. Robinet et al., “A Low-Power 0.7 μ V_{rms} 32-Channel Mixed-Signal Circuit for ECoG Recordings,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, Dec. 2011.
 - [16] G. Schalk et al., “Decoding two-dimensional movement trajectories using electrocorticographic signals in humans,” *Journal of Neural Engineering*, Jun. 2007.
 - [17] A. Moly et al., “An adaptive closed-loop ECoG decoder for long-term and stable bimanual control of an exoskeleton by a tetraplegic,” *Journal of Neural Engineering*, Apr. 1, 2022.
 - [18] C. S. Mestais et al., “WIMAGINE: Wireless 64-Channel ECoG Recording Implant for Long Term Clinical Applications,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Jan. 2015.
 - [19] N. Koo, “Design of Low Power and Low Noise Instrumentation Amplifier for Biopotential Acquisition: A Review,” *IEEE Access*, 2025.
 - [20] G. Buzsáki, *Rhythms of the Brain*. Oxford University Press, Oct. 26, 2006.

Beyond Hemispheric Laterality: Riemannian Metrics as Markers of Motor Imagery Performance in BCI Training

Sainath Murali¹, Aryan Sudeep¹, Raghavendran Lakshminarayanan², Nivethida Thirugnanasambandam^{1, 2, #}

¹Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, India

²Koita Centre for Digital Health, Indian Institute of Technology Bombay, India

[#]E-mail Correspondence: nivethida.t@iitb.ac.in

ABSTRACT: A major challenge in motor imagery based brain-computer interfaces (MI-BCIs) is BCI illiteracy, where many users fail to achieve reliable control despite training. Most MI-BCI decoding approaches rely heavily on hemispheric laterality of sensorimotor rhythms, overlooking a key aspect of BCI learning - the user's ability to generate distinct and stable neural representations during MI. Here, we aimed to identify a quantifiable neural marker of MI by examining whether neural dynamics captured by Riemannian geometry-based EEG metrics relate to users' self-reported MI performance across runs. Participants were categorized as Improved, Worsened, or Unchanged based on perceived BCI accuracy across multiple runs, and three covariance-based metrics - class stability, class distinctiveness, and rest distinctiveness were estimated. Our results show that these metrics are modulated differently across participant groups, suggesting that MI decoding reflects multidimensional neural dynamics beyond hemispheric laterality and that Riemannian metrics may help assess an individual's ability to generate effective MI signals for BCI control.

INTRODUCTION

Brain-computer interfaces (BCIs) enable direct communication between the brain and external devices by translating neural activity into control signals. These systems have shown promise for restoring motor function in individuals affected by neurological disorders such as stroke, spinal cord injury, and neurodegenerative diseases [1, 2]. Among BCI paradigms, motor imagery-based BCIs (MI-BCIs) are widely studied because they are noninvasive and can promote neuroplasticity through effective engagement of motor networks [3]. During motor imagery, users imagine movements without physical execution, producing characteristic modulations in sensorimotor rhythms (SMRs), particularly event-related desynchronization (ERD) in the mu (8–13 Hz) and beta (13–30 Hz) bands [4]. These neural patterns can be decoded to control external devices or provide neurofeedback. Despite advances in signal processing and machine learning tools, MI-BCI performance remains highly variable across individuals. Approximately 10–30% of users fail to achieve reliable control, a phenomenon often referred to as BCI

inefficiency or BCI illiteracy [5]. While many studies have focused on improving decoding algorithms, evidence suggests that variability in user performance often exceeds differences between classifiers [6]. Consequently, MI-BCI control is increasingly viewed as a learnable cognitive motor skill, requiring users to discover and consistently reproduce neural patterns that are both stable and discriminable. This learning process depends on both neurophysiological factors and cognitive aspects such as attention, strategy, and perceived agency [6].

Traditional approaches to evaluating MI-BCI performance rely primarily on classifier accuracy as the measure of neural control. Within this framework, successful BCI operation is interpreted as the classifier's ability to distinguish between neural patterns generated during left and right-hand motor imagery. Improvements in decoding accuracy are therefore typically attributed to stronger hemispheric lateralization, where contralateral sensorimotor regions exhibit greater modulation than ipsilateral regions during imagined movement [7]. Studies combining EEG and fMRI have shown that individuals capable of controlling MI-BCIs often exhibit clear contralateral activation patterns in sensorimotor areas, whereas users lacking such lateralization frequently struggle to achieve reliable control [8].

Classifier accuracy provides only an indirect measure of neural organization and primarily reflects the separability between task-related neural representations. In practical BCI operation, effective control not only requires neural representations of different tasks to be separable, but also these representations to diverge sufficiently from baseline neural activity. If task-related signals remain close to resting-state dynamics, decoding may remain unstable even when class representations are distinguishable. Consequently, understanding how neural representations evolve during training requires metrics that capture multiple dimensions of neural structure rather than relying solely on classifier accuracy.

To address this limitation, Lotte and Jeunet proposed classifier-independent neural performance metrics based on Riemannian geometry [9]. These metrics operate directly on the covariance structure of multichannel EEG signals and quantify key properties of neural representations, including class distinctiveness (separation between

imagery classes), class stability (consistency of neural patterns within a class), and rest distinctiveness (separation between task-related activity and the resting baseline). Because these measures are independent of a specific classifier, they provide a physiologically interpretable description of neural adaptation during MI-BCI training.

In the present study, we use Riemannian geometry-based metrics to investigate how neural representations evolve during MI-BCI training and how these changes relate to users' perceived performance. By integrating run-wise neural metrics with behavioral learning trajectories, we examine whether successful and unsuccessful MI-BCI learning are associated with distinct patterns of class stability, class distinctiveness, and rest distinctiveness. Through this framework, the study aims to provide a multidimensional characterization of neural adaptation during MI-BCI beyond traditional hemispheric laterality-based interpretations.

MATERIALS AND METHODS

Dataset and Participants: EEG data were obtained from the publicly available motor imagery dataset introduced by Cho et al. [10], designed to investigate inter-individual variability in SMR-based BCI performance. The dataset contains recordings from 52 healthy participants (19 females; mean age 24.8 ± 3.86 years) performing left- and right-hand motor imagery across 5–6 runs with 20 trials per class per run. After each run, participants reported their expected BCI performance and motor imagery difficulty. EEG was recorded using a 64-channel Biosemi ActiveTwo system at 512 Hz. Participants with incomplete questionnaire data or unusable EEG recordings were excluded, resulting in a final cohort of 44 participants.

EEG Preprocessing: EEG preprocessing was performed using MNE-Python [11]. Signals were restricted to the 64 scalp electrodes, followed by DC offset correction, linear detrending, band-pass filtering (0.1–50 Hz), and a 60 Hz notch filter. Artifact removal involved RANSAC-based channel repair, Artifact Subspace Reconstruction (ASR), and Infomax ICA. Components were automatically labeled using mne-icalabel [12], and those with artifact probability greater than 90% were removed. Noisy channels were interpolated, and signals were re-referenced to the common average reference.

Riemannian Geometry-Based Neural Metrics: EEG data were segmented into epochs from -1.99 s to 5 s relative to cue onset, baseline-corrected, and cropped to the 0–3 s motor imagery period. Signals were filtered into delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), gamma (30–45 Hz), and broadband (0.1–50 Hz) bands. Neural dynamics during motor imagery were quantified using classifier-independent Riemannian metrics proposed by Lotte and Jeunet [9], implemented using the PyRiemann toolbox [13]. Three metrics—class distinctiveness, class stability, and rest distinctiveness—were computed separately for each frequency band and run.

Table 1: Definitions and equations for the computed Riemannian neural metrics.

Metric	Definition & Equation
Class Distinctiveness	Separation between two task-related neural states (left vs right imagery). $classDis(A, B) = \frac{\delta_R(\bar{C}^A, \bar{C}^B)}{\frac{1}{2}(\sigma_{CA} + \sigma_{CB})}$ where $\bar{C}^A, \bar{C}^B$: geometric means; σ_{CA}, σ_{CB} : intra-class dispersions; δ_R : Riemannian distance.
Class Stability	Consistency of neural patterns within a class. $classStab(A) = \frac{1}{1 + \sigma_{CA}}$ where σ_{CA} : dispersion within class A.
Rest Distinctiveness	Separation between task-related neural activity and the resting baseline. $restDis(A) = \frac{\delta_R(\bar{C}^A, C_{rest})}{\frac{1}{2}(\sigma_{CA} + \sigma_{C_{rest}})}$ where C_{rest} : resting baseline covariance matrix.

Behavioral Performance Trajectories: After each run, participants reported their expected MI-BCI performance as a percentage score reflecting how accurately they believed they controlled the MI-BCI. Behavioral learning across runs was quantified using Ordinary Least Squares (OLS) regression (`np.polyfit`), where the slope represented the longitudinal change in expected performance (% change per run). Subjects were categorized as Improved, Worsened, or Unchanged based on the direction and magnitude of this slope.

Neural Adaptation Trajectories: To examine neural adaptation during training, slopes were computed for each Metric $\times$ Frequency Band combination, yielding 35 neural features per subject. Neural metrics were standardized using population-level z-score normalization, and outliers exceeding $|Z| > 2$ were removed. Slopes were then estimated using OLS regression across runs, with standardized neural metric values as the dependent variable and run index as the independent variable. Positive slopes indicated increasing neural modulation across runs, whereas negative slopes indicated decreasing modulation.

Statistical Analysis: To evaluate differences in neural adaptation between behavioral groups, the Kruskal–Wallis test was used. Statistical significance was validated using a permutation test (5000 iterations), and features with permutation-corrected $p < 0.05$ were considered significant. Significant features were further examined using Mann–Whitney U tests with Holm–Bonferroni correction.

One-vs-Rest Effect Size Analysis: To identify neural features uniquely associated with each behavioral group, a One-vs-Rest (OVR) analysis was conducted. Longitudinal slope distributions for each behavioral group were compared with the pooled distribution of the remaining subjects. Cohen's d quantified effect size, where positive and negative values reflected stronger or weaker run-wise neural trends relative to the remaining participants.

Classifier Benchmarking: To determine whether the behavioral groups were reflected in decoding performance, a Riemannian Minimum Distance to Mean (MDM) classifier implemented in PyRiemann was used. Classification was performed independently for each subject to distinguish left versus right motor imagery, with accuracy estimated using cross-validation across trials.

RESULTS

Results are first presented for stratification of participants based on their behavioral performance and their corresponding motor imagery difficulty trends across runs, thus characterizing their longitudinal learning behavior. This is followed by analyses of neural metric changes, frequency-specific relationships with performance, group-level slope differences, discriminative neural signatures, and classifier-based validation to examine how different dimensions of neural organization relate to MI-BCI learning trajectories.

Stratification of Participants based on Behavioral Performance Trajectories: Participants exhibited heterogeneous behavioral trajectories across the five runs. Based on the slope of reported BCI performance, subjects were categorized into Improved, Unchanged, and Worsened groups. The Improved group showed a progressive increase in performance from approximately 56% in Run 1 to over 72% by Run 5 (Fig. 1A). In contrast, the Worsened group declined from about 72% to 61%, while the Unchanged group remained relatively stable around 60–62%. Subjective imagery difficulty showed a different pattern (Fig. 1B). The Improved group reported moderate difficulty with a temporary improvement around Run 3, whereas the Worsened group reported relatively easy imagery throughout. In contrast, the Unchanged group consistently reported greater difficulty. These observations indicate that perceived imagery ease did not directly correspond to performance improvements. Reported BCI performance reflects subjective evaluation of MI-BCI control, whereas imagery difficulty reflects how difficult participants perceived the motor imagery task itself.

Longitudinal Evolution of Neural Representations: To examine neural adaptation across runs, Riemannian neural metrics were computed using the trial-wise pre-cue baseline as the resting reference. Although resting-state recordings were available, the pre-cue baseline was used to remain consistent with Lotte and Jeunet [9] and better capture run-wise MI training effects. Figure 2 shows longitudinal changes in neural metrics relative to Run 1. The Improved group showed increasing trends across rest distinctiveness, class stability, and left–right distinctiveness metrics, indicating progressively more separable and consistent motor imagery representations. The Unchanged group showed mild declining trends across class stability and left–right distinctiveness metrics, indicating limited longitudinal improvement in neural representations. The Worsened group showed mild increases in left–right distinctiveness and stability metrics but decreasing rest distinctiveness, indicating progressively structured representations with reduced separation from baseline neural activity.

Frequency-Specific Neural Correlates of Motor Imagery: To examine spectral contributions to performance, Spearman correlations were computed between behavioral performance and neural metrics across frequency bands (Fig. 3). Across runs, correlations strengthened particu-

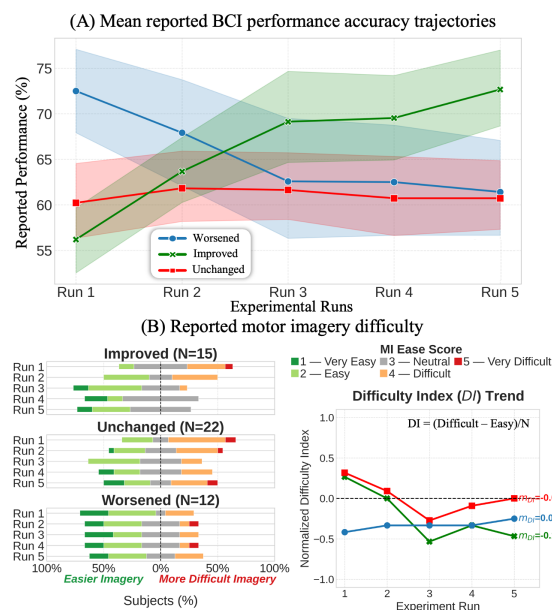


Figure 1: a) Average self-reported expected MI-BCI performance (%) across 5 runs, categorized by the slope of expected MI-BCI performance (Improved, Unchanged, Worsened). b) Distribution of reported motor imagery difficulty ratings and normalized Difficulty Index (DI) trends across 5 runs for each behavioral group, where $N_{difficult}$ and N_{easy} denote the number of difficult and easy responses, respectively. Positive DI values indicate greater perceived imagery difficulty, whereas negative values indicate easier imagery performance.

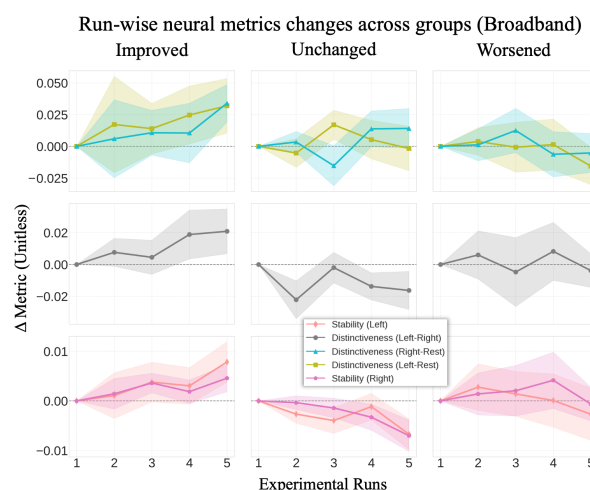


Figure 2: Run-wise changes in broadband neural metrics across behavioral groups relative to each subject's Run 1 baseline. Lines represent median changes across subjects, and shaded regions indicate SEM across subjects.

larly in beta and gamma bands, especially for rest-related distinctiveness and left–right separability. When stratified into behavioral groups, the Improved group showed stronger and more consistent correlations in lower frequency bands, whereas the Unchanged group exhibited weaker and more heterogeneous relationships. In the Worsened group, correlations were primarily observed in higher frequency bands, suggesting that spectral features alone may not fully explain the decline in performance.

Divergent Neural Adaptation Slopes Across Behavioral



Figure 3: Spearman correlation heatmaps between expected MI-BCI performance and neural metric values across frequency bands for a) all subjects and b) behavioral groups. Positive correlations indicate association with higher expected MI-BCI performance. Color intensity indicates correlation strength (Red = Positive, Blue = Negative). Asterisks denote significance (* $p < 0.05$, ** $p < 0.01$).

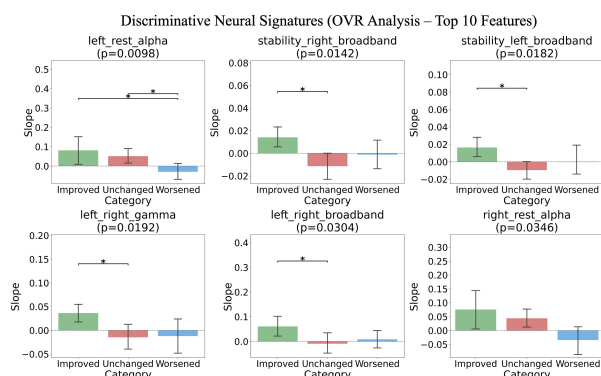


Figure 4: Divergent neural adaptation trends across behavioral groups. Subplots show mean slopes of neural metrics across training runs. p -values (titles) are derived from a permutation-based Kruskal-Wallis test (5,000 iterations); asterisks indicate significant post-hoc pairwise differences (Holm-corrected Mann-Whitney U, $p < 0.05$). Error bars represent ± 1 SEM.

Groups: To quantify longitudinal changes in neural metrics, slopes were computed for each metric across runs. Significant differences in neural adaptation slopes were observed among the behavioral groups (Fig. 4). Notably, alpha-band rest distinctiveness (left_rest_alpha and right_rest_alpha) showed positive slopes in the Improved group but negative or near-zero slopes in the Worsened group. Similarly, broadband stability metrics increased in the Improved group but decreased in the Unchanged group. Differences were also observed in left-right broadband distinctiveness and gamma-band separability, which increased in the Improved group but remained stable or decreased in the other groups. These results indicate that improvement in MI is associated with positive neural adaptation across multiple representational dimensions.

Neural Signatures of Behavioral Learning Trajectories: To identify neural features distinguishing each group, an OVR effect size analysis using Cohen's d was performed on the slopes of neural metric changes across runs (Fig. 5). The Improved group showed strong positive effect sizes for broadband stability and left-right separability,

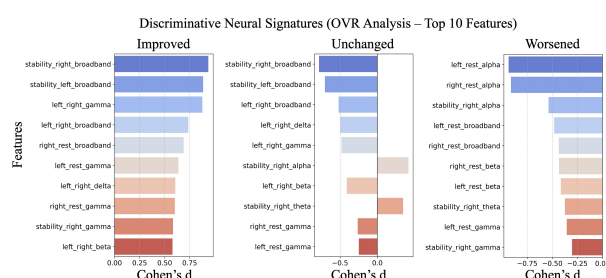


Figure 5: Discriminative neural signatures identified using one-vs-rest (OVR) analysis. Top neural metric slopes distinguishing each behavioral group, ranked by Cohen's d effect size. Positive values indicate stronger association with the target group relative to the remaining participants.

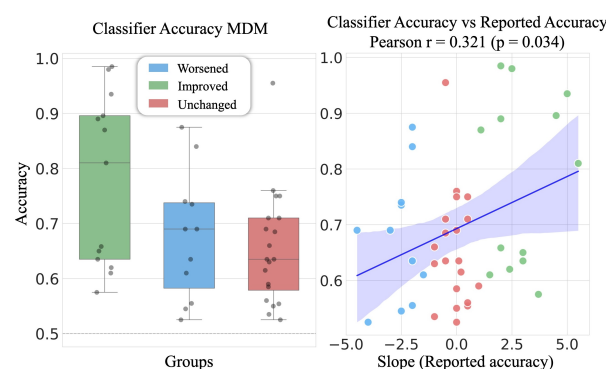


Figure 6: Relationship between classifier performance and behavioral learning trends. Left: Distribution of subject-wise MDM classification accuracy across behavioral groups. Right: Positive correlation between classification accuracy and behavioral performance slope (Pearson $r = 0.321$, $p = 0.034$).

indicating that increased neural consistency and class separability characterize successful learners. The Unchanged group exhibited negative effect sizes for stability and class-separability metrics, suggesting weaker neural structuring despite increased task-rest divergence. In contrast, the Worsened group showed large negative effect sizes for rest-related alpha and broadband metrics, indicating that reduced separation from baseline activity characterizes declining performance.

Classifier-Based Validation of Behavioral Groups: To determine whether these behavioral groups were reflected in decoding performance, a subject-specific Riemannian Minimum Distance to Mean (MDM) classifier was applied to classify left versus right motor imagery. Classification accuracy differed across groups (Fig. 6A). The Improved group achieved the highest median accuracy (~ 0.81), while the Worsened and Unchanged groups achieved lower accuracies (~ 0.69 and ~ 0.63 , respectively). Classifier accuracy was also positively correlated with behavioral learning slope (Pearson $r = 0.321$, $p = 0.034$; Fig. 6B), indicating that stronger behavioral improvement corresponded to greater neural decodability.

DISCUSSION

Motor imagery brain-computer interfaces (MI-BCIs) have traditionally been interpreted through the framework of hemispheric lateralization, where successful control arises

from contralateral modulation of sensorimotor rhythms during imagined movements [7]. In this classical view, improvements in BCI performance reflect increasingly strong lateralized neural activity that enables classifiers to distinguish between motor imagery classes. However, accumulating evidence suggests that this explanation alone cannot account for the variability observed in BCI learning outcomes [9]. In the present study, we examined neural adaptation during multi-run MI-BCI training using Riemannian geometry-based neural metrics and longitudinal slope analysis. Our results indicate that BCI learning cannot be explained solely by hemispheric lateralization but instead reflects the coordinated evolution of class stability, class distinctiveness, and rest distinctiveness. Class distinctiveness is closely related to hemispheric lateralization because it reflects the separability between left- and right-hand motor imagery representations. However, the present results showed that changes in class distinctiveness alone were insufficient to explain behavioral trajectories, as the groups also differed in the evolution of class stability and rest distinctiveness.

Consistent with previous reports highlighting substantial inter-individual variability in BCI performance [9, 14], participants in our dataset exhibited heterogeneous behavioral trajectories. Some individuals showed progressive improvements in classification accuracy across runs, whereas others displayed stable or declining performance. Importantly, these behavioral patterns were accompanied by distinct neural adaptation profiles. Participants in the Improved group exhibited simultaneous increases in class stability, class distinctiveness, and rest distinctiveness across runs. The neural dynamics observed in this group suggest a coordinated strengthening of motor imagery representations. Rest distinctiveness increased across runs, indicating that neural activity during motor imagery became increasingly separated from the resting baseline. At the same time, class distinctiveness increased, reflecting improved separability between the two imagery classes, while class stability increased, indicating more consistent neural patterns within each class. Together, these changes likely improve both the signal-to-noise ratio and the reliability of neural representations used for decoding, consistent with the higher classification accuracy observed in this group.

The Unchanged group exhibited a different neural profile. Participants in this group showed increasing rest distinctiveness but decreasing class distinctiveness and class stability across runs. This pattern suggests that although neural activity diverged from the resting baseline during the task, the neural patterns associated with the two imagery classes became less discriminable and less consistent over time. In other words, participants appeared to engage in task-related neural activity but failed to develop class-specific representations required for BCI control. This profile aligns with the concept of BCI illiteracy described in the literature [9, 14]. A particularly relevant study is the simultaneous EEG-fMRI motor imagery BCI investigation by Zich et al. [8], which identified three cate-

gories of users: BCI literate, illiterate, and pseudoilliterate. In their work, pseudo-illiterate users produced motor imagery activity detectable in fMRI but not EEG, suggesting that neural signals were present but not measurable by EEG due to neurophysiological factors such as dipole orientation. In contrast, BCI-illiterate users showed no clear motor imagery modulation in either EEG or fMRI. Because the present study relies solely on EEG recordings, pseudo-illiteracy cannot be identified. Instead, the Unchanged group likely corresponds to the true BCI-illiterate group, where participants fail to develop neural representations that differentiate motor imagery classes. The increasing rest distinctiveness observed in this group suggests engagement in task-related cognitive activity without recruitment of the neural circuits required for discriminable motor imagery patterns. This interpretation is also consistent with the greater subjective difficulty reported by this group.

The Worsened group exhibited yet another neural adaptation profile. In this group, class stability and class distinctiveness increased across runs, while rest distinctiveness decreased. This indicates that participants were able to generate structured neural patterns differentiating the two imagery classes, but the strength of these signals relative to the resting baseline diminished over time. One possible explanation is that these participants produced class-specific neural activity that was weak relative to baseline activity, reducing the reliability of signals used by the classifier. Unlike the Unchanged group, these users appear capable of generating class-specific neural patterns. Their declining performance may therefore reflect insufficient signal strength, fatigue, reduced engagement, difficulty identifying an effective motor imagery strategy, or recruitment of neural networks not optimally aligned with the intended MI task rather than an inability to perform motor imagery [15]. Such users may benefit from improved training protocols or personalized feedback strategies designed to strengthen task-related neural modulation [9].

Hence, these findings suggest that successful MI-BCI learning requires the simultaneous increase of class stability, class distinctiveness, and rest distinctiveness. Improvements in only one or two metrics appear insufficient for reliable MI-BCI performance. Increasing rest distinctiveness without increasing class distinctiveness produces the neural profile observed in the Unchanged group, whereas increasing class stability and class distinctiveness without rest distinctiveness leads to the pattern observed in the Worsened group. These results suggest that unsuccessful MI-BCI performance may arise from multiple neural failure modes rather than a single underlying limitation.

Recent studies have suggested that MI learning involves large-scale reorganization of functional connectivity and resting-state networks rather than isolated modulation of sensorimotor laterality alone. Resting-state functional connectivity changes have been reported in sensory-motor, visual, and default-mode networks following MI training, indicating that baseline neural dynamics evolve throughout learning [16, 17, 18]. In this context, the present

findings suggest that Riemannian metrics derived from covariance geometry may capture aspects of this broader network-level reorganization through changes in class stability, class distinctiveness, and rest distinctiveness.

Methodologically, this study demonstrates the value of Riemannian geometry-based neural metrics for analyzing motor imagery EEG signals [19]. By examining the covariance structure of multichannel EEG signals, these metrics capture changes in neural representations that are not easily detected using traditional canonical frequency band power-based measures. Analyzing the longitudinal slopes of these metrics across training runs allowed us to identify distinct neural adaptation trajectories associated with different behavioral outcomes.

Finally, classifier benchmarking further supports these findings. Participants in the Improved group achieved higher classification accuracy using a Riemannian Minimum Distance to Mean classifier, and classifier accuracy showed a positive relationship with behavioral learning slopes. This indicates that the behavioral grouping derived from performance trajectories reflects meaningful differences in the separability of neural representations within the EEG signal.

Future research should investigate whether these neural adaptation profiles generalize across different BCI paradigms and training protocols. Understanding these trajectories may help inform the development of adaptive training strategies and personalized feedback systems aimed at improving BCI learning outcomes.

CONCLUSION

This study investigated neural adaptation during motor imagery BCI training using Riemannian neural metrics. Our results show that successful learning is associated with simultaneous increases in class stability, class distinctiveness, and rest distinctiveness, whereas unsuccessful users exhibit partial or conflicting changes in these metrics. In particular, participants in the Unchanged group increased rest distinctiveness but failed to develop class-specific neural representations, consistent with the concept of BCI illiteracy described in previous studies [8, 9]. Overall, these findings demonstrate that MI-BCI learning cannot be explained by hemispheric lateralization alone, but instead depends on the coordinated evolution of class stability, class distinctiveness, and rest distinctiveness during training.

REFERENCES

- [1] Daly JJ, Wolpaw JR. Brain-computer interfaces in neurological rehabilitation. *Lancet Neurol.* 2008;7(11):1032–1043.
- [2] Feigin VL, Owolabi MO, Feigin VL, Abd-Allah F, Akinyemi RO, Bhattacharjee NV, et al. Pragmatic solutions to reduce the global burden of stroke: A world stroke organization-lancet neurology commission. *Lancet Neurol.* 2023;22(12):1160–1206.
- [3] Cervera M, Soekadar S, Ushiba J, Millán JdR, Liu M, Birbaumer N, et al. Brain-computer interfaces for post stroke motor rehabilitation: A meta-analysis. *Ann. Clin. Transl. Neurol.* 2018;5(5):651–663.
- [4] Pfurtscheller G, Lopes da Silva FH. Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clin. Neurophysiol.* 1999;110(11):1842–1857.
- [5] Jeunet C, N’Kaoua B, Lotte F. Advances in user training for mental-imagery-based bci control: Psychological and cognitive factors and their neural correlates. *Prog. Brain Res.* 2016;228:3–35.
- [6] Jeunet C, Lotte F, Batail JM, Philip P, Micoulaud Franchi JA. Using recent bci literature to deepen our understanding of clinical neurofeedback: A short review. *Neuroscience.* 2018;378:225–233.
- [7] Pfurtscheller G, Neuper C. Motor imagery and direct brain-computer communication. *Proc. IEEE.* 2001;89(7):1123–1134.
- [8] Zich C, Debener S, Kranczioch C, Bleichner MG, Gutberlet I, De Vos M. Real-time eeg feedback during simultaneous eeg-fmri identifies the cortical signature of motor imagery. *NeuroImage.* 2015;114:438–447.
- [9] Lotte F, Jeunet C. Defining and quantifying users’ mental imagery-based BCI skills: a first step. *J. Neural Eng.* 2018;15(4):046030.
- [10] Cho H, Ahn M, Ahn S, Kwon M, Jun SC. EEG datasets for motor imagery brain-computer interface. *Gigascience.* 2017;6(7):gix034.
- [11] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. Meg and eeg data analysis with mne-python. *Front. Neurosci.* 2013;7:267.
- [12] Li A, Feitelberg J, Saini AP, Höchenberger R, Scheltienne M. Mne-icalabel: Automatically annotating ica components with iclabel in python. *J. Open Source Softw.* 2022;7(76):4484.
- [13] Barachant A, Barthélemy Q, King JR, Gramfort A, Chevallier S, Rodrigues PL, et al. Pyriemann (v0.10). Zenodo. 2026.
- [14] Vidaurre C, Blankertz B. Towards a cure for bci illiteracy. *Brain Topogr.* 2010;23(2):194–198.
- [15] Singh A, Lal S, Guesgen HW. Small-world EEG network features predict motor imagery BCI performance. *Sensors.* 2021;21(6):2173.
- [16] Zhang L, He H, Sun J, Liu T, Luo C, Li P, Ma L, Yao D. Motor imagery learning modulates functional connectivity of multiple brain systems in resting state. *PLoS One.* 2014;9(1):e85489.
- [17] Ge S, et al. Motor imagery learning induced changes in functional connectivity of the default mode network. *Front. Hum. Neurosci.* 2021;15:730899.
- [18] Stefano Filho CA, de Oliveira A, Soares AB, et al. Motor imagery practice and feedback effects on functional connectivity during brain-computer interface training. *J. Neural Eng.* 2021;18(4).
- [19] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain-computer interface classification by riemannian geometry. *IEEE Trans. Biomed. Eng.* 2012;59(4):920–928.

Physiological-Based Change Point Detection During Sustained Attention: A multimodal EEG and Eye-Tracking Study

Naomi Thomas¹, Dalya Al-Mfarej², James Tung²

¹ Biomedical Engineering, University of Waterloo, Waterloo, Canada

² System Design Engineering, University of Waterloo, Waterloo, Canada

E-mail: naomi.thomas@uwaterloo.ca

ABSTRACT: Distinguishing adaptive cognitive effort from fatigue in real-time is critical for optimising task difficulty in motor rehabilitation. This study applies a data-driven change-point detection framework to 60-minute Attention Network Task (ANT) recordings from 7 healthy participants to detect cognitive state transition by integrating spectral and connectivity EEG features, and eye-tracking features. For each participant, 7 classifiers were trained at every candidate split time to identify the temporal boundary at which pre- and post-transition physiological and behavioural states were maximally discriminable. Multimodal fusion of physiological features achieved a mean ROC AUC of 0.933 ± 0.050 , exceeding behavioural features alone by 0.245 AUC units ($t(6) = 17.89, p < 0.001$) on average. The grand mean physiological-based split occurred at 21.6 ± 7.4 minutes, with substantial inter-individual variability (range: 13.4–35.0 min). Post-split exhibited decreased α / α ratios, increased α power, and pupil enlargement, consistent with compensatory effortful processing rather than passive fatigue. The difference between EEG and behavioural based change-point time ($\Delta = 2.9 \pm 14.4$ minutes) demonstrates that behavioural monitoring alone cannot detect when neural state transitions occur, underscoring the need for individualised physiological monitoring within Challenge Point Framework-based rehabilitation systems.

INTRODUCTION

Effort and fatigue are distinct, yet interrelated, measures critical to inform effective motor rehabilitation. Effort refers to active investment of resources to meet task demands. In the cognitive domain, these resources include attentional allocation, working memory engagement, and processing speed. In contrast, fatigue is defined as a time-dependent state characterized by the progressive depletion of resources, resulting in decreased vigilance and processing efficiency. Hockey's compensatory control model provides a mechanistic account of this relationship: performance may be actively maintained under developing fatigue through increased resource expenditure, but only at the cost of elevated subjective effort and physiological strain [1]. Where primary task performance is maintained, decrements may instead manifest as disrupt-

tion to secondary activities, less efficient strategy use, and/or fatigue after-effects (i.e. prolonged recovery, reduced motivation for subsequent tasks, carry-over impairment), emphasizing that effort and fatigue can co-occur and interact dynamically in a variety of ways [1].

Despite the conceptual distinction between effort and fatigue, they are identified by many of the same physiological signals, and can produce overlapping or opposing patterns. While existing literature has attended to the use of physiological signals to identify and measure effort and fatigue, they are often examined in isolation without assessing their relationship, and are defined inconsistently and often times interchangeably, lacking a unified framework.

The Challenge Point Framework (CPF) provides a structured approach for titrating task difficulty and maximizing motor learning, which has been applied to optimize motor rehabilitation [2]. CPF proposes that optimal learning occurs when task demands are aligned with the individual's skill level and task performance, promoting engagement without overwhelming or demoralizing patients. The point at which optimal learning/recovery occurs is referred to as the challenge point. The optimal challenge point shifts as skill improves, or when fatigued. Within the CPF, effort informs whether task difficulty should be increased (e.g., low effort indicating under-challenge), whereas fatigue indicates a functional limit. Beyond these limits, increasing difficulty may elevate risks, reduce learning efficiency, and/or discourage patients.

Physiological signals, such as electroencephalogram (EEG) and eye tracking (ET) (i.e., gaze and pupillometry), have been proposed to objectively measure cognitive effort and/or fatigue. EEG-based metrics of interest in identifying cognitive states include spectral and connectivity features of different brain rhythms. The meta-analysis by Chikhi et al. [3], summarizes that cognitive workload produces frontal θ increase and α decrease, aligning with Klimesch et al.'s review [4] showing frontal θ and parietal α desynchronisation, jointly indicating effortful processing. Arnau et al. [5], found that θ increases during inter-trial intervals reflecting fatigue-related idling of cognitive control networks, whereas stimulus-locked θ decreases over a session reflecting depleted executive resources and reduced task engagement. This dual be-

haviour means that θ alone cannot distinguish compensatory effort from fatigue without considering its temporal and spatial context. On the other hand, α power has been shown to consistently decrease with active engagement and increase with fatigue, specifically over posterior channels [6]. γ oscillations show the opposite profile wherein power increases with cognitive load, working memory maintenance, and feature binding, but decreases with fatigue and time-on-task [7]. Furthermore, Howard et al. [7] demonstrated a near-linear relationship between γ power and Sternberg memory load at frontal and temporal regions, where γ decreases as time-on-task and fatigue increase [6]. This bidirectional profile makes γ a valuable marker for CPF.

EEG electrode-level connectivity analysis has been found useful in characterizing effort and fatigue. Fronto-parietal θ synchrony, quantified via Phase Locking Value (PLV), has been found to serve as a mechanism for long-range communication between cognitive networks, scaling with working memory demand and executive control [8]. Sauseng et al. demonstrated that θ coherence between frontal and parietal sites increases during executive manipulation of visuospatial information, while fatigue reduces network efficiency and diminishes fronto-parietal synchrony over time [9],[10].

Eye-tracking metrics provide access to arousal and cognitive load indicators through the oculomotor system. Pupil dilation, reliably indexed by task-evoked pupil dilation (TEPD), has been found to scale with information-processing load [11]. Gilzenrat et al. confirmed these findings behaviourally, showing increases in baseline pupil diameter accompanying disengagement from task [12]. On the other hand, the Pupil Unrest Index (PUI) quantifies slow spontaneous pupil oscillation validated as a marker of arousal instability, correlating with EEG δ , θ , and alpha-1 power in sleep-deprived adults and predicting subsequent performance failure [13]. Lastly, gaze entropy captures scanning strategy information such that more focused, lower-entropy fixation patterns correspond to effortful engagement, while dispersed high-entropy patterns accompany fatigue-related attentional drift [14],[15].

The primary objective of this preliminary study is to assess the sensitivity of different physiological features in identifying transitions in cognitive states, which may correspond to (cognitive) effort and fatigue, and to compare them to the sensitivity of established behavioural metrics. The secondary objective is to investigate individual variability in the association between physiological and behavioural measures around state transitions. To investigate, this study employed a temporal change-point detection framework to determine whether and when a reliable transition in neurophysiological state occurs within a sustained cognitive task (Attention Network Task), and to characterise the physiological signatures (i.e., EEG, ET) of that transition.

Rather than dividing the task into fixed pre- and post-periods, classifiers were trained at every candidate split

time to find the temporal boundary at which two states are maximally discriminable. This approach makes no assumption about when the transition occurs, quantifies its timing on a per-participant basis, and directly compares the discriminative value of neural and oculomotor signals against overt task performance, providing a data-driven test of the compensatory effort hypothesis within the CPF context. By identifying distinct neural and physiological signatures of adaptive effort versus fatigue, rehabilitation protocols can be made more precise, dynamic, and responsive to the individual's evolving state. For example, if performance decline is driven by cognitive fatigue, task adaptation may involve reducing working memory demands, simplifying instructions, or pacing attentional load. On the other hand, inadequate cognitive effort can be addressed with increasing task difficulty to help with engagement.

MATERIALS AND METHODS

Participants:

Seven healthy adults (19–36 years; 3 female, 4 male) participated. All self-reported no known neurological or musculoskeletal disorders.

Physiological Recording Hardware:

EEG was recorded using a g.Nautilus-16 wireless system (g.tec Medical Engineering, 24-bit, 250 Hz [16]) from Fp1, Fp2, F7, F3, F4, F8, T7, C3, C4, T8, P7, P3, P4, P8, O1, and O2, with ground and reference placed on each of the participant's mastoids. Signals were streamed and synchronised through the Lab Streaming Layer (LSL) [17], and segmented post-hoc. Pupil diameter and gaze position were recorded using Pupil Labs Neon glasses (Pupil Labs, Berlin) [18], synchronised with task events via LSL [17]. Pre-task resting-state baseline (1-min eyes-open and eyes-closed) established the pupil reference mean for task-evoked pupil dilation and the baseline eyelid aperture for PERCLOS calculation.

Tasks:

Participants completed the Attention Network Task (ANT) continuously for 60 minutes. The ANT presents targets flanked by congruent, incongruent, or neutral stimuli, preceded by alerting, spatial, or no-cue conditions, enabling simultaneous assessment of three attentional networks: alerting, orienting, and executive conflict. Reaction time (RT) and accuracy were recorded trial-by-trial. Resting-state baselines were collected immediately before and after the task.

ANALYSIS

Windowing:

A common 30-second non-overlapping window scheme was applied uniformly across all modalities.

Feature Extraction:

Behavioural data were recorded as timestamped trial-level files. Trials with RT below 100 ms were excluded as anticipatory responses. Response measures, such as

mean, median, and standard deviation of RT across all valid trials; overall accuracy; and trial count (reflecting processing speed) were computed per window. Additionally, condition-specific RT and accuracy were calculated separately for each flanker congruency condition (congruent, incongruent, neutral). Lastly, attentional network scores were computed with three difference scores operationalizing the core ANT networks. The alerting score (RT double-cue - RT no-cue) quantifies the benefit of a temporal warning cue. The orienting score (RT spatial-cue - RT double-cue) quantifies the additional benefit of spatial location information. The conflict score (RT incongruent - RT congruent) quantifies the executive cost of flanker interference. Positive values indicate the expected directional benefit or cost for each network.

Eye-tracking gaze and eyelid signals were smoothed with an 11-point Savitzky–Golay filter (polynomial order 2) prior to feature extraction. Where both left and right pupil channels were available, pupil diameter was computed as their mean. The mean pupil diameter, pupil range (max – min, as a proxy for reactivity), pupil variance, Pupil Unrest Index (PUI, computed as the mean absolute rate of change of pupil diameter $|\Delta d / \Delta t|$), and task-evoked pupil dilation (TEPD, mean pupil diameter minus the pre-task baseline mean) were extracted per window. Gaze entropy was computed as the Shannon entropy (bits) of the 2D gaze position histogram over a 25×25 bin grid, quantifying spatial dispersion of fixation. PERCLOS was computed as the proportion of the 30-second window during which the eyelid aperture was classified as closed. A sample was considered closed if the eyelid aperture fell at or below the 15th percentile of the participant’s baseline eyelid distribution, established during the eyes-open baseline period prior to the task. Blink rate was computed as the number of discrete blink events per minute, where a blink was defined as eyelid aperture falling at or below the 10th percentile of the baseline distribution for between 50 and 400 ms.

Sixteen EEG channels were grouped into five regions of interest (ROIs): Frontal (Fp1, Fp2, F7, F3, F4, F8), Central (C3, C4), Temporal (T7, T8), Parietal (P7, P3, P4, P8), and Occipital (O1, O2). EEG signals were bandpass filtered between 0.5 and 45 Hz using a zero-phase FIR filter (Hamming window, firwin design, order = 1650) and a notch filter was applied at 60 Hz to remove power line noise. Following that, artefact removal was performed using the Automatic and Tunable Artefact algorithm [19], with visual inspection of each recording to ensure signal morphology was preserved, and neural components of interest were not over attenuated. Power spectral density was estimated per ROI using a periodogram with a Hann window over 1–45 Hz. Absolute band power was integrated across five canonical bands (δ , θ , α , β , γ), and relative band power was computed by dividing each band by total 1–45 Hz power. Three spectral ratios were additionally derived as cognitive state indices: θ / α , θ / β , and engagement index ($\beta / [\alpha + \theta]$). All features were computed for each of the five ROIs.

Functional connectivity was estimated using directed and undirected approaches applied at the ROI-electrode level (5 ROIs). Signals were detrended prior to connectivity estimation. Multivariate autoregressive modelling (MVAR) was used to estimate directed (partial directed coherence) and undirected (coherence, phase coherence, magnitude squared coherence) connectivity metrics within each 30-second window using a 4th order model fitted via ordinary least squares. Spectral matrices were band-averaged for θ and α . Additional phase-based and amplitude-based connectivity features were computed directly from the filtered EEG data. For phase-based metrics, the analytic signals for the θ and α bands were obtained via the Hilbert transform and phase estimates were derived over 2-second sub-epochs within each 30-second window. Following that, the phase locking value (PLV), weighted phase lag index (wPLI), and imaginary coherence were computed, and for the amplitude-based metrics, z-scored broadband signals using pairwise cross-correlation and autocorrelation-corrected cross-correlation. All connectivity matrices were summarised per window as the mean off-diagonal value.

Modelling:

The primary objective was to identify, for each participant, the time point at which their physiological state changed the most prominently. For each candidate split time t^* , windows were assigned binary labels (class 0: before t^* ; class 1: from t^* onward), and a battery of 7 classifiers was trained and evaluated to determine the effectiveness of discriminating the two classes (i.e., states). The split time maximising classification performance was taken as the best separation point. This was repeated across six feature sets: eye-tracking (8 features), EEG spectral (40 features: 5 relative band powers and 3 spectral ratios across 5 ROIs), EEG connectivity (16 features: 8 MVAR-based, 6 phase-based, and 2 time-domain), combined spectral-connectivity EEG (56 features), all physiological signals (EEG+eye-tracking; 64 features), and behavioural features (14 features).

All classifiers received z-scored features via a `StandardScaler` pipeline. Classifiers supporting `class_weight` were configured with `class_weight='balanced'`, and sample weights inversely proportional to class frequency were passed to `fit()` to equalise class contributions. Split candidates were bounded by a minimum of 20 windows per class (at least 10 minutes per class). Performance was evaluated using 5-fold stratified cross-validation, with probability scores pooled across folds and a single ROC AUC computed from the full pooled vector. The best split time for each participant, feature set, and classifier was defined as the candidate t^* yielding the highest ROC AUC.

RESULTS

Classification Performance Across Feature Sets:

Table 1 summarises mean ROC AUC and complementary metrics at the best-split candidate, averaged across

all 7 participants and 7 classifiers for each of the six feature sets. Multi-modal fusion of EEG spectral, EEG connectivity, and eye-tracking features achieved the highest mean ROC AUC (0.933 ± 0.050), followed by the spectral and connectivity combination (0.914 ± 0.060) and EEG spectral alone (0.894 ± 0.079). Eye-tracking (0.885 ± 0.113) and EEG connectivity (0.836 ± 0.081) provided strong unimodal performance, while the behavioural feature set achieved markedly lower discrimination at near-chance binary separation from overt performance metrics alone.

Table 1: Mean classification performance at the best-split candidate (mean $\pm$ SD across models). Bal. Acc. = balanced accuracy; Spec. = specificity. Bold = highest value per column.

Feature set	ROC AUC	Bal. Acc.	F1	Recall	Spec.
All modalities	0.933 $\pm$ 0.050	0.827 $\pm$ 0.105	0.922	0.930	0.725
Spec + Conn	0.914 $\pm$ 0.060	0.820 $\pm$ 0.088	0.876	0.871	0.769
EEG Spectral	0.894 $\pm$ 0.079	0.799 $\pm$ 0.105	0.857	0.861	0.736
Eye Tracking	0.885 $\pm$ 0.113	0.802 $\pm$ 0.121	0.852	0.840	0.763
EEG Connectivity	0.836 $\pm$ 0.081	0.723 $\pm$ 0.102	0.833	0.833	0.613
Behavioural	0.663 $\pm$ 0.065	0.579 $\pm$ 0.072	0.568	0.587	0.572

Temporal Change-Point Analysis:

Table 2 presents the best split time for each participant. Within participant SD of best split-times across different inputs (EEG-spectral, EEG-connectivity, EEG spectral and connectivity, ET, all physiological, behavioural) ranged from 5.2 minutes (S004) to 15.5 minutes (S003). Notably, the best-performing modality varied across participants. While ET was optimal for S001 and S002, EEG-informed modelling approaches outperformed ET-only for S003, S005, and S007. This cross-participant variability in timing and optimal modality likely reflects that per-participant model selection and feature weighting is optimal compared to the fixed-modality approach. To quantify the correspondence between physiological and behavioural change-point times, Pearson correlation between each participant's mean EEG-based best-split time (averaged across EEG spectral, connectivity, and spectral and connectivity based models) and their mean behavioural best-split time, was computed. The resulting correlation was low and statistically non-significant ($r = 0.34$, $p = 0.453$, $N = 7$).

Table 2: Best split time (t^* , min) and ROC AUC per participant by modality. All mod. = all physiological signals; EEG = best across spectral, connectivity, and spec+conn; Eye = eye-tracking; Beh. = behavioural.

Subj.	All modalities		EEG		Eye		Behavioural	
	t^*	AUC	t^*	AUC	t^*	AUC	t^*	AUC
S001	12.75	0.936	29.75	0.889	12.75	0.955	46.75	0.717
S002	11.75	0.949	15.25	0.902	11.25	0.984	37.75	0.679
S003	12.25	0.937	45.75	0.940	27.75	0.919	17.25	0.706
S004	23.75	1.000	10.25	0.997	23.75	1.000	14.25	0.779
S005	40.75	0.999	41.25	0.999	36.25	0.990	15.25	0.808
S006	12.25	0.920	12.25	0.920	11.25	0.693	28.25	0.667
S007	10.75	0.970	10.75	0.962	31.25	0.892	11.25	0.698

Changes in Physiological Features at the Change-Point:

Feature values were compared between pre- and post-split phases for all seven participants ($N = 7$) using the

all-modality feature matrix. Note that no feature reached statistical significance ($p < 0.05$) at this sample size.

Pupillometric measures increased post-split: mean pupil diameter (+0.102 mm), TEPD (+0.102), pupil range (+0.113), and PUI (+0.101). PERCLOS showed negligible change (+0.004), inconsistent with drowsiness-related eye closure. All pupillometric changes were directionally consistent with increased arousal and processing load [11][13].

θ -based spectral ratios showed consistent post-split changes across all ROIs. θ/α ratio declined most strongly in the temporal ($1.535 \rightarrow 1.296$, $\Delta = -0.239$) and central ($1.466 \rightarrow 1.248$, $\Delta = -0.218$) regions. θ/β ratios showed similar decreases (frontal $\Delta = -0.261$; central $\Delta = -0.228$). The engagement index $\beta/(\alpha + \theta)$ increased across all ROIs (temporal: +0.100; parietal: +0.087). Relative γ power increased across all five ROIs (Temporal +0.027, Central +0.021, Parietal +0.020, Occipital +0.019, Frontal +0.016). These changes are directionally opposite to cognitive fatigue trends reported in literature, which is characterised by rising θ , increasing θ/α ratios, and declining γ [6][20]. Connectivity changes were small and frequency-specific: MVAR-based θ coherence decreased ($\Delta = -0.007$), θ PLV decreased ($\Delta = -0.006$), while alpha wPLI increased (+0.008).

Figure 1 1, illustrates the dynamics for participant S005. All physiological markers show stable trajectories until ≈ 35 –41 min, at which point γ power, engagement index, and pupil diameter change abruptly. EEG-informed models found the best split to be at 41 mins, whereas the behavioural-based classifier identified an optimal split time at 15 mins. The 26-minute gap between physiological and behavioural change-points suggests performance-only monitoring would trigger CFP challenge adjustment well before the physiological-based feature transition. The difference here, and in others in the sample, suggests behavioural features may not correspond to underlying physiological state transition. Instead, behavioural data may be influenced by learning effects, task-adjustment periods, and/or environmental factors. Eye-tracking and EEG-based models exhibited a less drastic difference with eye-tracking splits ≈ 5 minutes preceding the EEG-based split time.

Feature Importance:

Permutation importance computed at each participant's mean optimal split revealed θ PLV as the most discriminative feature overall (mean importance = 0.0057), followed by PUI (0.0050), central γ (0.0037), α PLV (0.0026), and temporal α (0.0023). Eye-tracking features were prominently represented in the top ten, with gaze entropy (0.0019), pupil mean (0.0018), pupil variance (0.0016), and TEPD (0.0013) all ranking in the top fifteen features across participants. Despite a small session-average pre-post difference ($\Delta = -0.006$), the high ranking of θ PLV suggests its discriminative value reflects a phase-coordination transition at the split boundary rather than a sustained shift.

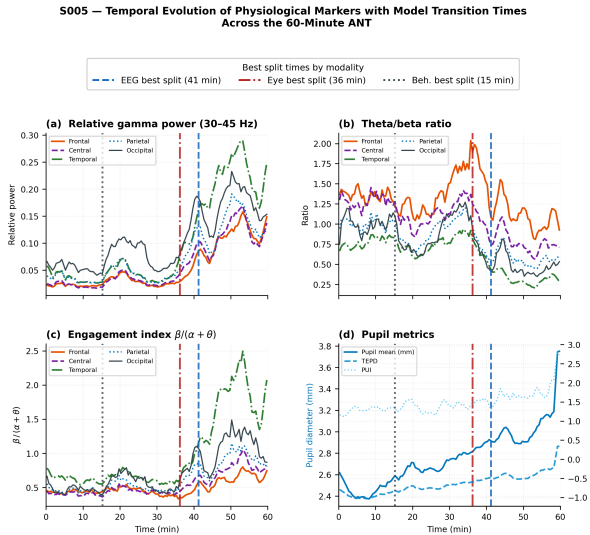


Figure 1: Temporal evolution of key physiological features across the 60-minute ANT for participant S005.

DISCUSSION

Multimodal physiological signals discriminated pre- and post-transition task phases with mean ROC AUC of 0.933, exceeding behavioural features by 0.245 AUC units ($t(6) = 17.89$, $p < 0.001$) across all participants. This is comparable to Trejo et al. [20], who reported 91–100% EEG accuracy when behavioural measures failed to distinguish alertness states for over 30 minutes, and Qi et al. [21], who showed EEG functional connectivity predicted performance impairment (AUC = 0.89) when performance indices were non-significant. The low correlation between EEG and behavioural split times ($r = 0.34$, $p = 0.453$) suggests that more than two cognitive states may be present, and that behavioural-informed change points are not directly reflective of internal physiological state changes. Furthermore, results indicate 5 of 7 (excluding S003 and S005) EEG-based change times precedes behavioural-based change points, indicating that some individuals maintain overt performance by mobilising additional cognitive resources, producing earlier physiological change-points with no corresponding behavioural decline, while others show behavioural change before a clear physiological transition.

Post-split spectral changes were directionally inconsistent with hypothesized cognitive fatigue, characterised by rising frontal θ , increasing θ/α ratios, and declining γ [6]. The inverse was observed: decreasing θ/α ratios (temporal $\Delta = -0.239$; central $\Delta = -0.218$; frontal $\Delta = -0.167$), decreasing θ/β ratios (frontal $\Delta = -0.261$; central $\Delta = -0.228$), and increasing γ power across all five ROIs alongside a rising engagement index. Inconsistent with cognitive fatigue, the observed profile is consistent with the compensatory effort framework described by Hockey [1]. Decreasing θ ratios were consistent with Arnau et al. [5], who showed that stimulus-locked θ decreases with time-on-task reflecting depleted trial-by-trial executive resources, even while inter-trial θ rises. The

net spectral pattern at the optimal split time (~ 22 min) is consistent with an early compensatory phase in which β and γ dominate before tonic fatigue-related θ accumulates, in line with Wascher et al. [22], who found compensatory frontal θ increased most in the early-to-middle task phase. Post-split increases in mean pupil diameter (+0.102 mm), TEPD (+0.102), and pupil range (+0.113) are consistent with expected changes in cognitive workload, as it produces enlarged baseline pupil diameter and a shift from exploitative task-optimised engagement toward more diffuse processing [11][13]. Gilzenrat et al. [12] directly confirmed this finding, showing baseline pupil diameter increased as task utility declined. The positive TEPD, specifically indexing effort exertion rather than demand alone, indicates participants were investing more effort relative to their pre-task baseline, consistent with stable ANT performance across the session. Connectivity changes were smaller in magnitude than spectral and pupillometric effects, but showed a frequency-specific pattern inconsistent with uniform fatigue-related reduction. MVAR θ coherence, θ MSC, and θ PLV decreased post-split ($\Delta \approx -0.006$ to -0.007), while alpha wPLI increased (+0.008). This divergence between declining θ coupling and increasing α phase-lag mirrors findings from Qi et al. [21] and Sun et al. [23]. The modest post-split θ coherence decline may reflect a reduction in the efficiency of the distributed attentional network necessary for ANT conflict resolution.

Inter-individual variability in change-point timing (13.4 – 35.0 min) and the lack of uniformity in optimal modality across participants (Table 2), emphasizes that a fixed-modality, and fixed-timing monitoring strategy is insufficient. Effective CPF-based rehabilitation requires individualized approaches to determine the physiological features indicative of state changes, and continuous monitoring. Feature importance supports this recommendation with the optimal feature-set differing across participants. These findings argue for passive monitoring frameworks in which modality weights and classifier parameters are continuously updated by individual rather than a universal model.

Limitations:

Limitation include that $N = 7$ with insufficient statistical power for significant pre-post feature testing, thus findings should be treated as hypothesis-generating. The 20-window minimum constraint means that state changes within the first or the last 10 minutes are undetectable. Cognitive states such as boredom and within-session learning were uncontrolled, and the change-point framework assumes a single binary transition, which may oversimplify the continuous and multi-dimension nature of cognitive state change, and no ground truth is available against which the detected transition can be validated.

CONCLUSION

A temporal change-point framework applied to 60-minute ANT recordings from 7 participants demon-

strates that multimodal physiological signals can detect individual neurophysiological state transition with mean ROC AUC of 0.933, exceeding overt behavioural performance features by 0.245 AUC units ($t(6) = 17.89$, $p < 0.001$) in every participant. The grand mean split for physiological based models occurred at 21.6 ± 7.4 minutes, with wide individual variation (range: 13.4–35.0 min). The post-split spectral signature: decreasing θ dominance, increasing γ , engagement index, pupil diameter, TEPD and PUI, is consistent with compensatory effortful processing [1],[5],[6],[11],[13] rather than passive fatigue. θ PLV and Central γ power emerged as the most individually reliable discriminative features. S005 exemplifies how physiological and behavioural change-points can be decoupled, with a 26-minute divergence between the EEG- and behavioural-based split-time identification. The low EEG-behavioural correlation ($r = 0.34$) and high inter-individual variability in change-point timing together establish that group-average or performance-based difficulty adaptation will be systematically mistimed for a substantial proportion of individuals, underscoring the need for continuous, individualised physiological monitoring within CPF-based rehabilitation systems [2]. Stroke patients represent the target population, as physiological monitoring for detecting the effort-to-fatigue transition prior to overt performance declines allows for therapists to adapt task difficulty proactively to ensure efficient learning according to the CPF. The wide inter-individual variability in change-point timing observed here (13.4–35.0 min) will likely be greater in stroke populations, reinforcing the need for individualised rather than group-average monitoring. Validation in patient populations, and assessment of whether EEG and oculomotor markers remain discriminative in the presence of lesion-related signal changes, are critical next steps.

REFERENCES

- [1] G. R. J. Hockey, “Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework,” *Biol. Psychol.*, vol. 45, no. 1, pp. 73–93, 1997.
- [2] M. A. Guadagnoli and T. D. Lee, “Challenge Point: A Framework for Conceptualizing the Effects of Various Practice Conditions in Motor Learning,” *J. Motor Behav.*, vol. 36, no. 2, pp. 212–224, 2004.
- [3] S. Chikhi, N. Matton, and S. Blanchet, “EEG power spectral measures of cognitive workload: A meta-analysis,” *Psychophysiology*, vol. 59, no. 6, p. e14009, 2022.
- [4] W. Klimesch, “EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis,” *Brain Res. Rev.*, vol. 29, no. 2, pp. 169–195, 1999.
- [5] S. Arnau, T. Brümmer, N. Liegel, and E. Wascher, “Inverse effects of time-on-task in task-related and task-unrelated theta activity,” *Psychophysiology*, vol. 58, no. 6, p. e13805, 2021.
- [6] S. K. L. Lal and A. Craig, “A critical review of the psychophysiology of driver fatigue,” *Biol. Psychol.*, vol. 55, no. 3, pp. 173–194, 2001.
- [7] M. W. Howard et al., “Gamma oscillations correlate with working memory load in humans,” *Cereb. Cortex*, vol. 13, no. 12, pp. 1369–1374, 2003.
- [8] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, “Measuring phase synchrony in brain signals,” *Hum. Brain Mapp.*, vol. 8, no. 4, pp. 194–208, 1999.
- [9] P. Sauseng, W. Klimesch, M. Schabus, and M. Doppelmayr, “Fronto-parietal EEG coherence in theta and upper alpha reflect central executive functions of working memory,” *Int. J. Psychophysiol.*, vol. 57, no. 2, pp. 97–103, 2005.
- [10] P. Sauseng, B. Griesmayr, R. Freunberger, and W. Klimesch, “Control mechanisms in working memory: A possible function of EEG theta oscillations,” *Neurosci. Biobehav. Rev.*, vol. 34, no. 7, pp. 1015–1022, 2010.
- [11] P. van der Wel and H. van Steenbergen, “Pupil dilation as an index of effort in cognitive control tasks: A review,” *Psychon. Bull. Rev.*, vol. 25, no. 6, pp. 2005–2015, 2018.
- [12] M. S. Gilzenrat, S. Nieuwenhuis, M. Jepma, and J. D. Cohen, “Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function,” *Cogn. Affect. Behav. Neurosci.*, vol. 10, no. 2, pp. 252–269, 2010.
- [13] F. Regen, H. Dorn, and H. Danker-Hopfe, “Association between pupillary unrest index and waking EEG activity in sleep-deprived healthy adults,” *Sleep Med.*, vol. 14, no. 9, pp. 902–912, 2013.
- [14] Z. Cui et al., “Gaze transition entropy as a measure of attention allocation in a dynamic workspace involving automation,” *Sci. Rep.*, vol. 14, p. 23405, 2024.
- [15] S. K. Pokkula, A. Amresh, and J. Duval, “Gaze entropy as a measure of player performance and its correlation with focus and attention,” in *Proc. ETRA '25*, New York, NY, USA: ACM, 2025.
- [16] g.tec medical engineering GmbH, “g.Nautilus Research Headset.” [Online]. Available: <https://www.gtec.at/product/gnautilus-research/>
- [17] C. Kothe et al., “The lab streaming layer for synchronized multimodal recording,” *Imaging Neuroscience*, vol. 3, p. IMAG.a.136, Sep. 2025, doi: 10.1162/IMAG.a.136.
- [18] Pupil Labs GmbH, “Neon Eye Tracker.” [Online]. Available: <https://pupil-labs.com/products/neon/>
- [19] N. Bajaj, “ATAR: Automatic and tunable artifact removal algorithm,” in *Signal Processing Toolkit (spkit)*. [Online]. Available: https://spkit.github.io/guide/notebooks/ATAR_Algorithm_EEG_Artifact_Removal.html [Accessed: Apr. 17, 2025]
- [20] L. J. Trejo, K. Kubitz, R. Rosipal, R. L. Kochavi, and L. D. Montgomery, “EEG-based estimation and classification of mental fatigue,” *PSYCH*, vol. 6, no. 5, pp. 572–589, 2015.
- [21] P. Qi et al., “EEG functional connectivity predicts individual behavioural impairment during mental fatigue,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 2080–2089, 2020.
- [22] E. Wascher et al., “Frontal theta activity reflects distinct aspects of mental fatigue,” *Biol. Psychol.*, vol. 96, pp. 57–65, 2014.
- [23] Y. Sun, J. Lim, J. Meng, K. Kwok, N. Thakor, and A. Bez-erianos, “Discriminative analysis of brain functional connectivity patterns for mental fatigue classification,” *Ann. Biomed. Eng.*, vol. 42, no. 10, pp. 2084–2094, 2014.

Evaluating the Impact of Generic Model Priors on User-Specific Calibration in P300 Brain-Computer Interfaces

Arnav Gupta, Boyla O. Mainsah, Leslie M. Collins

Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States

E-mail: {arnav.gupta, boyla.mainsah, leslie.collins}@duke.edu

The current standard approach for training brain-computer interface (BCI) models is user-specific. Due to high inter-individual variability, there is often a performance gap between generic BCI models trained on data from other users relative to user-specific models. Alternatively, generic models can provide informative priors to reduce BCI calibration time. We investigate the impact of model priors on shrinkage linear discriminant analysis (LDA) with Bayesian sequential updates using a normal-inverse-Wishart (NIW) conjugate prior. Baseline WD and CD P300 classifiers achieved off-the-shelf performance of 0.629 ± 0.022 and 0.612 ± 0.020 area under the curve (AUC), respectively. Incorporating 20% user-specific calibration data (total = 3780 samples) with WD and CD priors yielded a statistically significant increase to 0.683 ± 0.022 AUC with convergence of both models to user-specific performance 0.699 ± 0.023 . Sequential updates of shrinkage LDA (20% data increments) required on average 0.001 seconds with WD/CD priors vs. 0.140 seconds with only user-specific data, demonstrating that generic informative priors efficiently handle domain shifts with reduced computational overhead.

INTRODUCTION

The P300 brain-computer interface (BCI) speller [1] is a widely studied BCI paradigm that provides a means of communication for individuals with severe neuromuscular impairments, such as amyotrophic lateral sclerosis (ALS). In a typical visual P300 BCI speller, items are arranged in a grid interface and rows and columns are sequentially illuminated in a randomized sequence (Fig. 1). When the user's desired target character is illuminated - a relatively rare event - a P300 event related potential (ERP) response is elicited approximately 300 ms after the stimulus is presented. By analyzing segments of electroencephalography (EEG) signals synchronized with stimulus onset, the BCI identifies the user's intended character with a character decoding function (Fig. 2).

However, EEG signals are noisy and non-stationary, making reliable P300 classification challenging. Typically, user-specific BCI decoding models are trained on labeled data collected from the target user during a calibration session. While generic models trained on data from other users offer a potential to reduce BCI calibration time, their performance is often inferior to that of



Figure 1: (Left) Individual using a P300 speller. (Right) BCI spelling interface [2].



Figure 2: (Left) Average electroencephalography (EEG) signals from participant L_03 following target and non-target stimuli. A characteristic positive peak at approximately 270 ms is observed for target stimuli relative to non-target stimuli. (Right) Heatmap of character scores following cumulative P300 classifier score updates; character "V" receives the highest score and is selected as the user's intended character [2].

user-specific models due to inter-user variability and differences in population data domains [3]. Prior work has shown that data domain alignment plays a critical role in model performance, with user-specific P300 classifier models trained on limited data often outperforming generic models trained on larger but less aligned data [2]. The goal of this study is to bridge the gap between calibration-heavy user-specific models and lower-performing generic models in P300 BCIs with transfer learning. This work investigates if the use of informative model priors can mitigate the performance gap with generic models while reducing user BCI calibration time. Integrating incremental amounts of target user data into generic baselines, we quantify the amount of calibration data needed to achieve user-specific performance.

MATERIALS AND METHODS

Dataset: We used data from the bigP3BCI dataset, which contains EEG recordings from previous P300

speller studies with participants with and without ALS [4]. The ALS studies utilized a 6×6 speller grid with a 250 ms inter-stimulus interval (ISI), whereas the non-ALS studies utilized a 9×8 grid with a 125 ms ISI. For consistency, we limited to experiments with the checkerboard stimulus presentation paradigm [5].

Data Domains: Tab. 1 summarizes the study characteristics and the corresponding domain assignments. Population domain definitions include:

- User-specific (US): Training data from the target participant in the held-out test ALS study (Study L).
- Within-domain (WD): Training data from participants with ALS (Study B).
- Cross-domain (CD): Training data from participants without ALS (Studies A and C).

Table 1: Overview of P300 BCI studies used from the bigP3BCI dataset and domain definitions. ALS, amyotrophic lateral sclerosis; CD, cross-domain; US, user-specific; ISI, inter-stimulus interval; WD, within-domain.

Study	ISI (ms)	Grid Size	Population	Domain
A	125	9×8	Non-ALS	CD
B	250	6×6	ALS	WD
C	125	9×8	Non-ALS	CD
L	250	6×6	ALS	US

For baseline WD and CD configurations, a subset of 80,000 epochs was randomly drawn from each respective domain pool to train generic models; results were averaged from generic model training after three independent trials to minimize sampling variance. To account for the discrepancy in ISI between ALS (250 ms) and non-ALS (125 ms), we performed *ISI matching* [2] in the non-ALS data by retaining data from every other stimulus prior to generating the sampled subset. ISI matching is implemented on the stored data and does not alter the original recordings or the feedback delivered during the non-ALS sessions.

For the held-out study, user-specific data contained 3,780 user epochs each for training and testing. User-specific model parameters were updated after adding data sequentially in per-word increments of 756 epochs (20% per word increment), until the maximum amount of training data available.

Preprocessing: EEG signals from electrodes Fz, Cz, P3, Pz, P4, PO7, PO8, Oz were used for signal analysis [6]. Signals were downsampled from 256 Hz using block averaging at 20 Hz. A notch filter was applied at 60 Hz to suppress power line noise, followed by a band-pass filter (0.5–30 Hz) to isolate the P300-relevant frequency band while attenuating slow drifts. No automated artifact rejection was applied. Stimulus-locked epochs were extracted from 0–800 ms following each stimulus onset. Data across channels in epochs were vectorized into a $1 \times D$ feature vector prior to classifier training.

Sequential Classifier Updates: We selected a linear discriminant analysis (LDA) as a classifier model that is amenable to sequential updates. Features are assumed to be multivariate Gaussian variables. The LDA weight vector is defined as:

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad (1)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ are mean vectors of target and non-target classes, respectively; $\mathbf{\Sigma} = p_1\mathbf{\Sigma}_1 + p_0\mathbf{\Sigma}_0$ is the shared covariance matrix; and p_c is the class-specific proportion.

Shrinkage linear discriminant analysis (LDA) was first proposed for P300 classification in [7] to address the ill-conditioned nature of empirical covariance matrices in high-dimensional EEG feature spaces. In shrinkage LDA, the empirical covariance matrix, $\mathbf{\Sigma}$, towards a structured target matrix $\mathbf{T}$ via a shrinkage intensity coefficient $\lambda \in [0, 1]$:

$$\mathbf{\Sigma}_{\text{shrink}} = (1 - \lambda)\mathbf{\Sigma} + \lambda\mathbf{T} \quad (2)$$

In this work, we will update shrinkage LDA weights sequentially via Bayesian inference with a normal-inverse-Wishart (NIW) conjugate prior over the class-specific means and shared covariance [8]. The initial baseline class-specific parameters (mean vector $\mathbf{m}_0$, scaling factor κ_0 , degrees of freedom ν_0 , and scatter matrix derived from the shared covariance, $\mathbf{S}_0 = \nu_0 * \mathbf{\Sigma}_0$, define the prior model. Given an incoming block of data $\mathbf{X} \in \mathbb{R}^{D \times N}$ containing N samples with D -dimensions with a sample mean of $\bar{\mathbf{x}}$, the NIW distribution parameters are updated accordingly:

$$\kappa_N = \kappa_0 + N \quad (3)$$

$$\nu_N = \nu_0 + N \quad (4)$$

$$\mathbf{m}_N = \frac{\kappa_0\mathbf{m}_0 + N\bar{\mathbf{x}}}{\kappa_N} \quad (5)$$

$$\mathbf{S}_N = \mathbf{S}_0 + \mathbf{X}\mathbf{X}^T + \kappa_0\mathbf{m}_0\mathbf{m}_0^T - \kappa_N\mathbf{m}_N\mathbf{m}_N^T \quad (6)$$

The initial class-specific mean and covariance priors were set based on within-domain, cross-domain or user-specific priors (using *scikit-learn* package). κ_0 defines the weight of the initial mean prior while ν_0 defines the weight of the initial covariance prior; we set $\kappa_0 = \nu_0 = 1$. At each data increment step i with N_i samples, shrinkage LDA parameters are updated with the maximum *a posteriori* (MAP) estimates of class-specific means and shared covariance:

$$\boldsymbol{\mu}_{i,c}^{\text{MAP}} = \mathbf{m}_{N_i,c} \quad (7)$$

$$\boldsymbol{\Sigma}_i^{\text{MAP}} = \frac{\mathbf{S}_{N_i}}{\nu_{N_i} + D + 1} \quad (8)$$

RESULTS

P300 Classifier Performance: Performance of trained P300 classifiers was evaluated on test data of participants

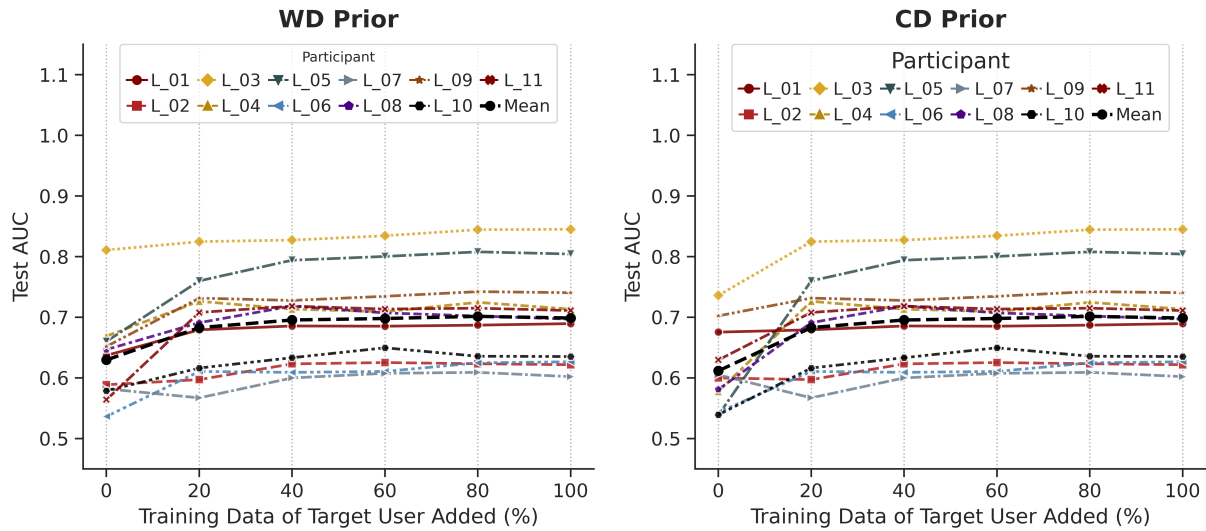


Figure 3: Test area under the curve (AUC) of shrinkage linear discriminant analysis (LDA) P300 classifiers with within-domain (WD) and cross-domain priors as a function of the amount of target user training data added (3780 total samples) for each participant.

in Study L in the bigP3BCI dataset. We analyzed how the incremental inclusion of target user data impacts classification performance across different training data domains. Fig. 3 shows the P300 classifier AUC results for WD and CD models, respectively as a function of the amount of training data from the target user for each participant. Fig. 4 shows the mean AUC values with sequential updates as a function of the amount of training data for generic and user-specific models. The generic WD baseline (0.629 ± 0.022) performs better than the generic CD baseline (0.612 ± 0.020), though this difference was not significant ($p = 0.37$). With incorporation of target user data, performance of both updated models with generic priors converge to that of user-specific models. Both generic models experience a significant initial performance gain with the first sequential update with 20% of target user data (0.683), moderate gains at 40% (0.695) and minimal gains until the full amount of calibration data (0.699).

To model progression of AUC performance in incoming sequential data, a linear mixed-effects (LME) model was fit with amount of calibration data and training data conditions as fixed effects and participant as a random intercept. The model revealed a highly significant main effect of amount of calibration percentage ($\beta = 0.001$, $p < 0.001$), confirming a reliable positive slope across calibration intervals. For trajectory variations from generic baselines, a second LME model restricted to the generic conditions was fit with an interaction term. The interaction between the amount of calibration data and generic training condition was non-significant ($\beta = 0.000$, $p = 0.421$, confirming that both the WD and CD priors follow a similar adaptation path when augmented with user-specific calibration data. Post-hoc paired tests between the initial generic baselines and counterpart sequentially updated models with the full calibration data revealed a highly significant performance gain (WD, $t = -5.286$,

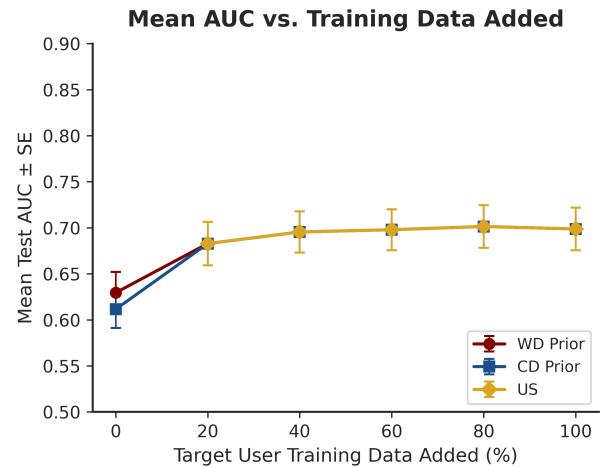


Figure 4: Mean test area under the curve (AUC) $\pm$ standard error (SE) of shrinkage linear discriminant analysis (LDA) P300 classifiers with within-domain (WD) or cross-domain prior and exclusively user-specific (US) data as a function of the amount of target user training data added (3780 total samples). Trajectories of mixed WD/CD converging towards US performance at 20% calibration.

$p = 0.0004$; CD, $t = -3.869$, $p = 0.0031$).

Fig. 5 shows the final classification performance across all conditions, tracking environments. At full calibration (100% added data), both sequentially update models with WD and CD priors achieved similar accuracy as user-specific models (0.699 ± 0.023). This confirms that regularized adaptation allows generic models to effectively overcome initial domain shifts.

Model Training Time: Fig. 6 shows the mean training time on our computing resource across training conditions. The baseline WD and CD models take a longer time due to a larger data pool; though this does not impact the training time of a novel user. Sequential updates with only user-specific data required 140 ms per data in-

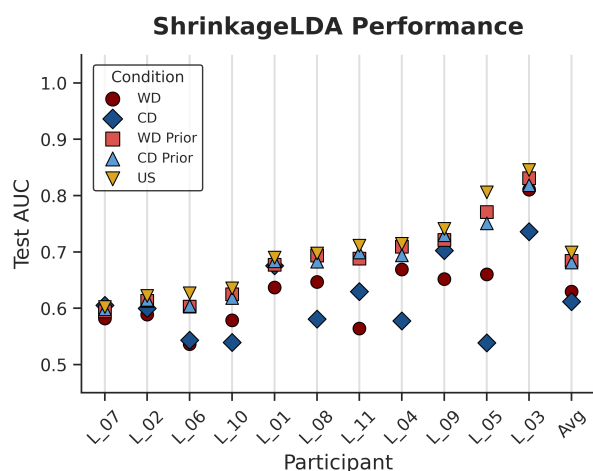


Figure 5: Area under the curve (AUC) of the receiver operating characteristic of shrinkage linear discriminant analysis (LDA) P300 classifiers with 100% calibration data. Markers denote participant test results for the various conditions and the mean performance (Avg). CD, cross-domain; US, user-specific; WD, within-domain.

crement while sequential updates with the WD and CD priors required 1 ms per increment. This represents an significant reduction in computational overhead with sequential updates vs. training on the final pooled data.

DISCUSSION

The analysis of shrinkage LDA with a Bayesian updating framework revealed a clear trend: when utilizing robust regularized linear classifiers, adaptive updates can effectively mitigate the impact of data domain mismatch. While within- or cross-domain data can introduce a persistent performance penalty, the integration of generic priors into local user parameters allows for rapid calibration and steady optimization of the decision space.

The Role of Sequential Calibration: Both within-domain and cross-domain shrinkage LDA models converged to to user-specific performance after a small initial amount of target-user data. This indicates that the NIW prior framework functions as an efficient regularizer. Instead of distorting the classification boundary, the historical prior data provides a robust foundation for the covariance structure, while the target user samples quickly align the class-specific means. This minimizes over-fitting to a limited amount of calibration data without retaining the unaligned structural biases of the baseline generic models.

Implications for BCI Deployment: While our bigP3BCI dataset includes data from individuals with ALS, data from target BCI end user populations are often difficult to obtain and not widely publicly available. In contrast, public BCI dataset mostly include data from individuals without neuromuscular limitations. Models with sequential updates using cross-domain priors achieved similar performance as that within-domain priors and user-specific models, demonstrating that domain

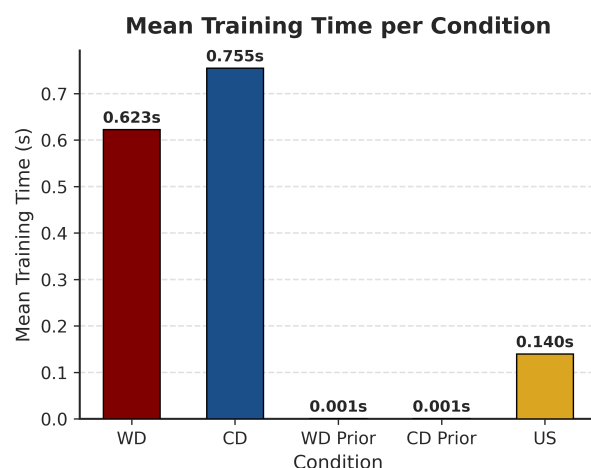


Figure 6: Mean training time of generic models and latency (seconds) of Bayesian sequential updates with generic prior (CD, cross-domain; WD, within-domain or only user-specific (US) data. The Xeon® Gold 6252 central processing unit (24-Core 2.1GHz) was used.

misalignment is not an absolute barrier. A BCI system can exploit generic model priors obtained from non target BCI end user populations, and potentially still achieve user-specific accuracy levels with a relatively short, user-tolerable calibration session.

Furthermore, an incremental data approach in a Bayesian framework offers a computational advantage over training linear models on pooled data. In addition, the shrinkage LDA model parameter updates were implemented in a much shorter amount of time with informative generic priors relative to updates with only user-specific data. Model recalibration mitigates the negative effects of data distribution drift during extended BCI use. A short model recalibration time could allow for more frequent continuous recalibration of BCI decoding models automatically with semi-supervised learning, e.g., [9], in the background without interruption during active BCI use.

Future work includes implementing sequential updates with linear models that enforce feature sparsity, such as stepwise LDA, as well as exploring fine-tuning of generic deep learning models, such as EEGNet [10].

REFERENCES

- [1] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*. 1988;70(6):510–523.
- [2] Gupta A, Liu A, Haines E, Alghamdi R, Collins LM, Mainsah BO. A data-centric analysis of the impact of training data quality vs. quantity on p300 brain-computer interface performance. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Jan. 2026.
- [3] Lin R et al. Assessing the impact of population data domain differences on transfer learning in p300-based brain-computer interfaces. In: *Proceedings of the AAAI*

Conference on Artificial Intelligence. Apr. 2025, 29415–29417.

[4] Mainsah B, Fleeting C, Balmat T, Sellers E, Collins L. bigP3BCI: An Open, Diverse and Machine Learning Ready P300-based Brain-Computer Interface Dataset (version 1.0.0). PhysioNet, RRID:SCR_007345. 2025.

[5] Townsend G et al. A novel p300-based brain-computer interface stimulus presentation paradigm: Moving beyond rows and columns. *Clinical Neurophysiology*. 2010;121(7):1109–1120.

[6] Krusienski D, Sellers E, McFarland D, Vaughan T, Wolpaw J. Toward enhanced p300 speller performance. *Journal of Neuroscience Methods*. 2008.

[7] Blankertz B, Lemm S, Treder M, Haufe S, Müller KR. Single-trial analysis and classification of erp components — a tutorial. *NeuroImage*. 2011;56(1):814–825.

[8] Murphy KP. Machine learning: A probabilistic perspective. MIT Press: Cambridge, MA (2012).

[9] Chen XJ, Collins LM, Mainsah BO. Language model-guided classifier adaptation for brain-computer interfaces for communication. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2022, 1642–1647.

[10] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.

Similar morphology, different timing: Cortico-spinal response delays following median and ulnar nerve stimulation

Markus E. Oberndorfer¹, Bernhard Golser¹, Gernot R. Müller-Putz^{1,2}

¹Institute of Neural Engineering, Graz Brain-Spine-Lab, Graz University of Technology, Graz, Austria

²BioTechMed Graz, Graz, Austria

E-mail: gernot.mueller@tugraz.at

ABSTRACT: Understanding the temporal propagation of sensory signals from peripheral nerves to the central nervous system remains challenging due to limited access to spinal cord activity. In this study, we investigate the capability of electrospinography (ESG) to resolve conduction delays following transcutaneous stimulation of the ulnar (UNS) and median (MNS) nerves. ESG and electroencephalography (EEG) recordings were obtained, preprocessed, and analyzed using latency estimation and cross-correlation methods. Spinal responses exhibited condition-specific latency differences, with UNS stimulation consistently preceding MNS stimulation by approximately 2–3 ms, in agreement with delays predicted from peripheral nerve anatomy and conduction velocity. Comparable delay patterns were observed in cortical EEG responses after averaging across participants, and waveform morphology remained similar between conditions. These findings suggest that ESG provides temporally precise information on ascending sensory signal propagation that may not be detectable at the cortical level alone. The results highlight the potential of ESG as a complementary tool for studying spatiotemporal dynamics within the human sensorimotor system.

INTRODUCTION

Understanding the temporal dynamics of neural signal propagation along somatosensory pathways is essential for characterizing the functional organization of the human nervous system and for developing neurophysiological biomarkers of neurological disorders. Peripheral nerve stimulation provides a controlled and reproducible means of probing these pathways, eliciting afferent responses that ascend from the peripheral nerve through the spinal cord to cortical processing stages within a few milliseconds.

Electroencephalography (EEG) has long been used to non-invasively record cortical somatosensory evoked potentials with high temporal resolution, enabling detailed investigation of sensory processing at the level of the brain. However, a comprehensive understanding of signal transmission along the entire somatosensory pathway requires access to neural activity at earlier stages of processing. In this context, electrospinography (ESG) has re-

cently reemerged as a promising non-invasive technique for capturing spinal cord responses using surface electrodes placed over the spine [1, 2]. The simultaneous acquisition of cortical and spinal signals therefore offers a unique opportunity to investigate conduction timing across hierarchical levels of the somatosensory system. While latency characteristics of responses following median nerve stimulation have been already studied [1–3], comparatively little is known about the temporal properties of responses evoked by stimulation of other upper-limb nerves, such as the ulnar nerve, particularly at the spinal level.

Anatomical and Neurophysiological Considerations: Somatosensory signals from the upper limb are conveyed by large myelinated first-order afferent fibers entering the spinal cord through the dorsal roots and ascending within the dorsal column pathways. These fibers synapse in the dorsal column nuclei of the brainstem, after which second-order neurons project to the thalamus. Subsequent thalamocortical projections transmit the signal to the primary somatosensory cortex. During this ascending transmission from the peripheral nerve to the brain, afferent volleys also engage segmental spinal circuits. These spinal circuits give rise to early spinal somatosensory evoked potentials that can be recorded non-invasively. Following median nerve stimulation, these responses typically manifest as a sequence of characteristic components, including an initial positive deflection (P1) at approximately 9 ms, a prominent negative peak (N1) around 13 ms, and a later positive component (P2) between 20 and 25 ms [1, 3]. This tri-phasic spinal cord potential (SCP) has been consistently described for median nerve stimulation at standard stimulation sites like the wrist [1, 2, 4]. Variations in stimulation location or afferent pathway length are expected to result in corresponding shifts in component latencies while preserving the overall waveform morphology. Detailed physiological interpretations of the individual components have been discussed elsewhere [5, 6] and are beyond the scope of the present study.

Aim and hypothesis: Whether comparable temporal response patterns are present following ulnar nerve stimulation relative to median nerve stimulation, and how dif-

ferences in conduction distance and pathway characteristics influence the timing of spinal and cortical responses, remains largely unexplored. Addressing this question is particularly relevant for understanding pathway-specific timing mechanisms and the morphological consistency of evoked responses. It also contributes to extending non-invasive multilevel electrophysiological approaches beyond well-established stimulation paradigms.

In the present study, we therefore combine high-density EEG and ESG recordings to quantify pathway-specific conduction timing following median and ulnar nerve stimulation. We hypothesize that stimulation of these anatomically distinct but functionally similar peripheral nerves evokes responses with comparable waveform morphology that are shifted in time due to differences in stimulation location and conduction distance. To test this hypothesis, we employ cross-correlation-based latency analysis of participant-specific evoked responses to assess both waveform similarity and relative temporal delay between stimulation conditions.

MATERIALS AND METHODS

Preliminary results of this ongoing study are based on data acquired from six healthy participants (3 male, 3 female; age 24 ± 2.3 years). All participants provided written informed consent prior to participation and were free to withdraw from the study at any time without providing a reason. The study protocol was approved by the Ethics Committee of Graz University of Technology.

Paradigm: Participants were positioned in a forward-inclined seated massage chair with dedicated chest and head supports, allowing relaxation of the neck and back muscles while reducing motion-related artifacts. They were instructed to remain still and to keep their eyes open throughout the recordings in order to reduce contamination of the data by increased alpha-band activity associated with eye closure.

Electrical stimulation was applied to the median nerve at the wrist and to the ulnar nerve at the elbow. The stimulation current amplitude was individually adjusted for each participant. The current was gradually increased until visible muscle contractions associated with activation of the stimulated nerve were observed and subsequently reduced to a level below motor threshold to avoid movements. This procedure ensured mixed-nerve stimulation with predominant activation of sensory afferents. Across participants, the stimulation intensity was 10 ± 2.7 mA for median nerve stimulation (MNS) and 9 ± 3.7 mA for ulnar nerve stimulation (UNS).

For each nerve, 3600 stimuli were delivered in 12 runs of 300 stimuli each, using random inter-stimulus intervals between 700 and 900 ms. In addition to the stimulation runs, three resting-state recordings of five minutes duration were acquired to obtain a comparable number of rest and stimulation epochs.

Between runs, participants were allowed to stand up or adjust their seating position to prevent muscle fatigue and

tension, as electromyographic (EMG) activity represents a major source of artifacts in ESG recordings.

Stimulation and measurement setup: Electrical stimulation was delivered using a DS7R constant-current stimulator (Digitimer Ltd., Letchworth Garden City, UK) in combination with 32-mm Flextrode Plus surface electrodes (Krauth+Timmermann GmbH, Hamburg, Germany). Stimulation electrodes were positioned at the wrist for median nerve stimulation and at the elbow for ulnar nerve stimulation.

EEG and ESG signals were recorded using a BrainAmp Standard amplifier system (Brain Products GmbH, Gilching, Germany) with active electrodes. A total of 64 electrodes were arranged according to a standard EEG montage, while 54 additional electrodes were positioned over the neck and upper back to capture electrospinographic activity (see Fig. 1) For a more detailed description of the ESG electrode montage and layout see [Golser et al., 2026 - accepted manuscript].

The reference electrode was placed on the tip of the nose and the ground electrode on the forehead. Two mastoid electrodes were used for virtual re-referencing, and one electrode placed below the left thorax was used for electrocardiogram (ECG) recording. Furthermore, two electrodes were positioned close to the stimulation sites to enable precise extraction of stimulation onset timing - these electrodes are hereafter referred to as stimulation signal channels. All signals were recorded with respect to a reference electrode located at the tip of the nose and were sampled at 5000 Hz.

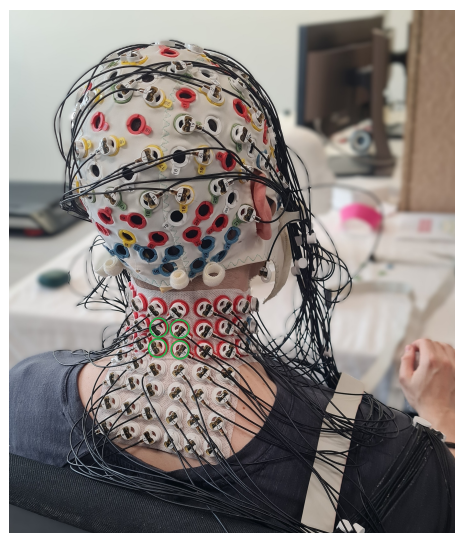


Figure 1: Electrode setup for recording brain and spinal cord activity during median and ulnar nerve stimulation. ESG electrodes marked with green rings are used as ROIs.

Data processing: All data processing was performed in Python using custom scripts in combination with the MNE-Python toolbox [7]. For both EEG and ESG recordings, signals were initially re-referenced to the mastoid electrodes prior to further preprocessing. Although stimulation triggers were logged via Lab Stream-

ing Layer (LSL) [8], hardware-induced delays and temporal jitter limited the reliability of these markers for latency analyses. To ensure accurate onset extraction, anomalous signal segments were automatically identified in the stimulation channels using robust thresholding of sample amplitude and sliding-window variance (median-absolute-deviation based z-scores). Detected samples were merged into bad intervals, and stimulation events within these intervals were excluded from further analysis to ensure reliable temporal alignment for EEG and ESG epoching.

EEG signal processing: First, line noise was removed using a notch filter at 50 Hz and its harmonics, and a band-pass filter between 1 and 1000 Hz was applied. Both filters were implemented as fourth-order Butterworth infinite impulse response (IIR) filters.

Bad channels were identified by visual inspection of the raw data and were interpolated using spline interpolation. Artifact correction was then performed using independent component analysis (ICA). Components were automatically classified by the MNE algorithm, and those associated with eye blinks, muscle activity, or channel noise were removed when the predicted artifact probability exceeded 90%. Finally, the data were re-referenced using a common average reference (CAR).

Cleaned signals were epoched relative to stimulation onset, including a pre-stimulus interval of 100 ms and a post-stimulus interval of 300 ms. For resting-state recordings, fixed-length epochs of identical duration were generated.

ESG signal processing: To attenuate power line interference, a notch filter at 50 Hz was applied. In addition, a high-pass filter at 1 Hz was used to remove slow signal drifts. Both filters were implemented as fourth-order Butterworth infinite impulse response (IIR) filters applied in a forward-backward (zero-phase) manner. Bad channels were identified by visual inspection and interpolated using local spatial averaging of adjacent electrodes. Due to the anatomical proximity of cervical recording sites to the heart, ESG signals were strongly contaminated by electrocardiographic (ECG) activity. To suppress this interference, signal space projection (SSP) was performed using the simultaneously recorded ECG channel [9]. Projection vectors were computed using the MNE-Python function `mne.preprocessing.compute_proj_ecg()`. In all participants, the first projection component accounted for more than 90% of the ECG-related variance and was therefore removed from the data.

Following artifact correction, the data were epoched relative to stimulation onset, including a pre-stimulus interval of 100 ms and a post-stimulus interval of 300 ms. Resting-state recordings were segmented into fixed-length, partially overlapping epochs (40 ms overlap) to increase the number of available rest trials. Epoch rejection was performed using joint probability and kurtosis criteria with an absolute z-score threshold of 3 (corresponding to ± 3 standard deviations). On average, approximately 10% of epochs were excluded from further

analysis.

Region-of-interest definition: For the analysis of evoked responses, regions of interest (ROIs) were defined based on the expected spatial distribution of somatosensory evoked potentials. For EEG, the cortical ROI comprises electrodes over the left motor cortex, namely C3, CP5, CP3, CP1 and P3. In contrast to cortical somatosensory evoked potentials, the spatial distribution of electrospinographic responses is less well established. Therefore, ESG regions of interest were defined using a data-driven approach. After preprocessing and averaging, ESG waveforms were visually inspected to identify electrodes exhibiting the strongest and most consistent deflections after around 13 ms post stimulus, corresponding to the expected arrival time of afferent activity following median nerve stimulation.

Based on this criterion, four centrally located cervical electrodes showing robust early responses were selected as the ESG region of interest. The selected electrodes are indicated by green rings in Fig. 1.

Similarity and latency analysis: For each participant and stimulation condition, preprocessed epochs were averaged to obtain subject-specific evoked responses. Subsequently, signals were averaged across channels within the predefined EEG and ESG regions of interest to derive representative ROI waveforms. All similarity and latency analyses were performed at the participant level to avoid bias related to unequal numbers of trials across participants. To assess whether responses evoked by median and ulnar nerve stimulation exhibited comparable signal patterns with a relative temporal shift, z-normalized cross-correlation analysis was performed. This puts the emphasis on the waveform shape rather than absolute amplitude differences. The normalized cross-correlation between stimulation conditions was computed as

$$R_{xy}(\tau) = \frac{1}{N} \sum_{t=1}^N \tilde{x}^{\text{MNS}}(t) \cdot \tilde{x}^{\text{UNS}}(t + \tau) \quad (1)$$

where τ denotes the temporal lag and N is the number of samples within a predefined post-stimulus analysis window. To focus on early somatosensory responses while avoiding contamination by stimulation artifacts, the correlation analysis was restricted to a time interval from 5 ms to 40 ms after stimulus onset. The lag corresponding to the maximum correlation coefficient

$$\tau^* = \arg \max_{\tau} R_{xy}(\tau), \quad (2)$$

was interpreted as the relative temporal delay between responses evoked by median and ulnar nerve stimulation. The corresponding peak correlation value was used as a measure of waveform similarity.

To ensure physiologically plausible delay estimates, the search range for τ was restricted to ± 10 ms. For visualization, grand-average ROI waveforms were computed by averaging participant-specific responses. Estimated temporal shifts were subsequently used to align ulnar nerve responses to median nerve responses.

RESULTS

After preprocessing, epoching, and averaging ESG and EEG signals across their respective ROI channels, condition-specific evoked response waveforms were obtained for the MNS, UNS, and REST conditions (Fig. 2). A pronounced negative deflection at approximately 13 ms post-stimulus (N1) is observed in the MNS condition of the ESG recording, indicating the arrival of the stimulus-evoked afferent volley at the recording site (Fig. 2 (b)). This component is followed by a subsequent positive deflection. A qualitatively similar response pattern is present in the UNS condition. However, the response exhibits reduced amplitude and a temporal shift relative to MNS. In the MNS EEG recordings, a negative peak at approximately 18 ms post-stimulus suggests the arrival of the afferent signal at the cortical level. Again, a similar signal morphology shows for the MNS and UNS condition in the EEG.

Similarity and delay analysis: Cross-correlation analysis of ESG responses between MNS and UNS conditions across participants indicates a consistent temporal ordering, with the UNS response preceding the MNS response. The mean ESG lag across participants was 3.57 ± 2.23 ms (mean $\pm$ SD), with a corresponding maximum correlation coefficient of 0.50 ± 0.20 . Individual participant values are reported in Table 1. When signals were averaged across participants prior to cross-correlation, the estimated lag between MNS and UNS responses was 2.60 ms (see Fig. 3). For EEG recordings, the mean lag between MNS and UNS responses was 2.43 ± 1.10 ms (mean $\pm$ SD), with maximum correlation coefficients of 0.73 ± 0.15 . After averaging EEG signals across participants, the temporal lag between MNS and UNS response was 3.12 ms.

Spatial distribution of ESG signal: Topographic maps in Figs. 4 and 5 illustrate the spatial distribution of ESG potentials across the electrode array at selected latencies for MNS and UNS, respectively. Prior to stimulation (Fig. 4 (a)), only low-amplitude background activity is visible. At approximately 9 ms post-stimulus (Fig. 4(b)), a focal negative potential emerges on the right side of the electrode array. The N1 component (13 ms) and the subsequent positive P2 component (22 ms) are clearly identifiable in the topographic distributions shown in Fig. 4(c) and (d), respectively.

The topoplots in Fig. 5 show the same time points from the MNS condition minus the calculated delay of 2.60 ms. While the pattern is similar to the MNS condition, it is less pronounced.

DISCUSSION

The ESG and EEG responses observed in this study are broadly consistent with previously reported somatosensory evoked potentials in terms of waveform morphology and latency. In particular, the characteristic negative ESG deflection around 13 ms post-stimulus aligns with

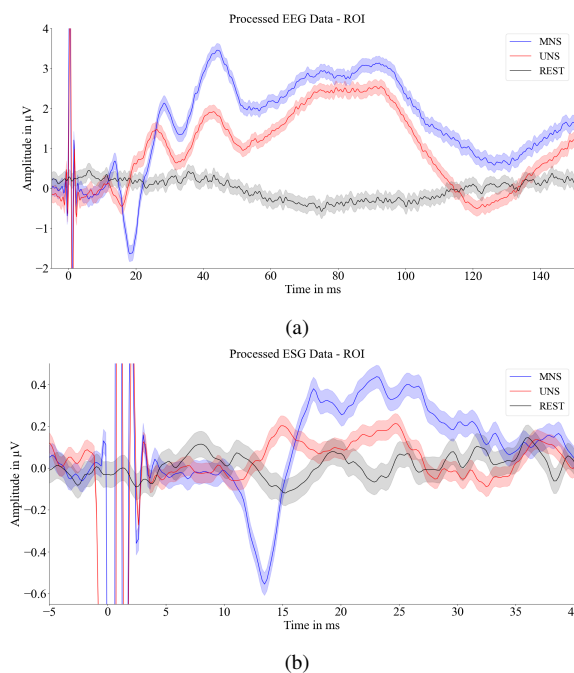


Figure 2: Signal in time-domain for the (a) EEG and (b) ESG recordings averaged over their respective ROI channels. Solid curves represent the mean over all epochs and the shaded surrounding area is the standard error.

Table 1: Participant-specific latency differences between median nerve stimulation (MNS) and ulnar nerve stimulation (UNS) derived from normalized cross-correlation analysis of evoked responses. Positive lag values indicate that responses following UNS occurred earlier than MNS. The maximum correlation coefficient reflects the degree of waveform similarity between stimulation conditions after temporal alignment.

#	ESG		EEG	
	Lag (ms)	max. Corr.	Lag (ms)	max. Corr.
1	7.8	0.41	2.8	0.49
2	4.6	0.53	2.0	0.78
3	0.6	0.86	0.6	0.64
4	3.2	0.26	3.1	0.89
5	2.6	0.46	3.8	0.74
6	2.6	0.47	2.3	0.84

reported spinal components following peripheral nerve stimulation, while the cortical response observed at approximately 18 ms in EEG corresponds to early cortical processing stages. A central finding of this work is the systematic temporal difference between UNS and MNS stimulation at the spinal level. Cross-correlation analysis revealed that ESG responses following UNS stimulation precede MNS responses by 3.57 ± 2.23 ms across participants, with a lag of 2.60 ms when signals were averaged prior to analysis. This temporal shift is physiologically plausible. Assuming an average forearm length of approximately 22 cm and a large-fiber conduction velocity of approx. 70 m/s [10], the expected conduction time difference between proximal (UNS) and distal (MNS) stimulation sites would be approximately 3.10 ms. The experimentally observed delay therefore

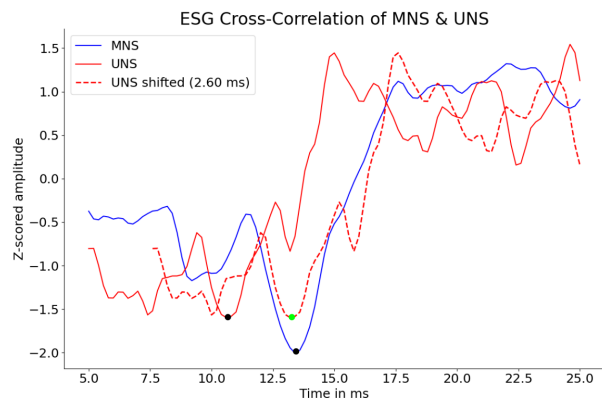


Figure 3: The lag between MNS and UNS signal after z-normalization in the ESG and a shifted UNS to show the overlay to MNS. The timepoints of the MNS and UNS signal that indicate arrival at the spinal cord are marked with a black circle, the green circle indicates the shifted point of arrival of UNS.

closely matches the delay predicted by peripheral nerve anatomy and conduction physiology, suggesting that the ESG recordings reliably capture propagation of the afferent volley along the peripheral nerve towards the spinal cord. Consistent with the latency difference observed at the spinal level, comparable delays were also identified in cortical EEG responses after participant-level averaging. This finding suggests that peripheral conduction differences between stimulation sites are preserved along the ascending sensory pathways and can be detected not only at the spinal level but also at the cortical level. Thus, both ESG and EEG appear sensitive to conduction-related timing differences, indicating consistent temporal propagation of afferent signals from the peripheral nerve to the cortex following MNS and UNS stimulation. After z-normalization, UNS and MNS responses exhibited highly similar waveform morphology in both ESG and EEG recordings, further supported by moderate cross-correlation values. The comparable waveform morphology suggests that both stimulation paradigms evoke similar spinal response dynamics of afferent fibers and engage functionally analogous ascending sensory pathways. The observed latency differences therefore most likely reflect variations in conduction distance between stimulation sites rather than fundamentally different neural processing mechanisms. Across all recordings, the amplitude of UNS responses was consistently lower than that of MNS responses in both ESG and EEG. The underlying cause of this amplitude difference remains unclear. Possible explanations include differences in effective recruitment of sensory fibers due to electrode placement and tissue impedance. Additionally, differences in anatomical depth and orientation of the nerve relative to the stimulation electrodes may influence the resulting signal strength.

Limitations: A limitation of these preliminary findings is related to the stimulation paradigm. Although stimulation intensity was selected to remain below the visible motor threshold in order to preferentially activate sensory

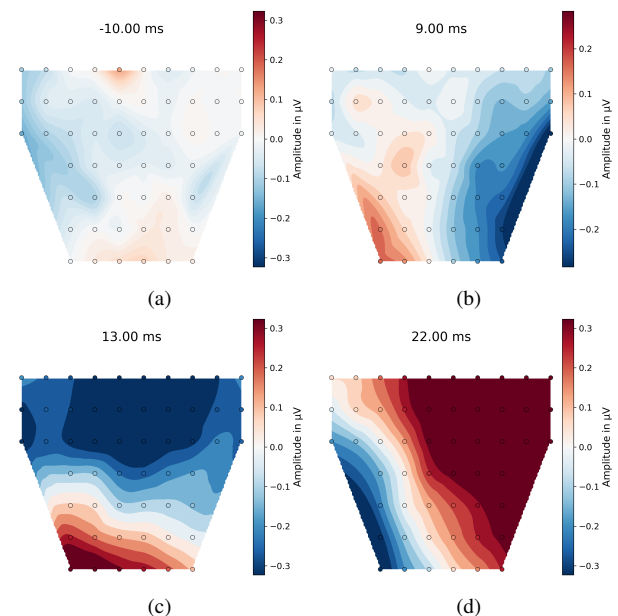


Figure 4: Topoplot of averaged ESG recording during MNS stimulation at (a) 10 ms pre stimulus, (b) 9 ms post stimulus, (c) 13 ms post stimulus and (d) 22 ms post stimulus.

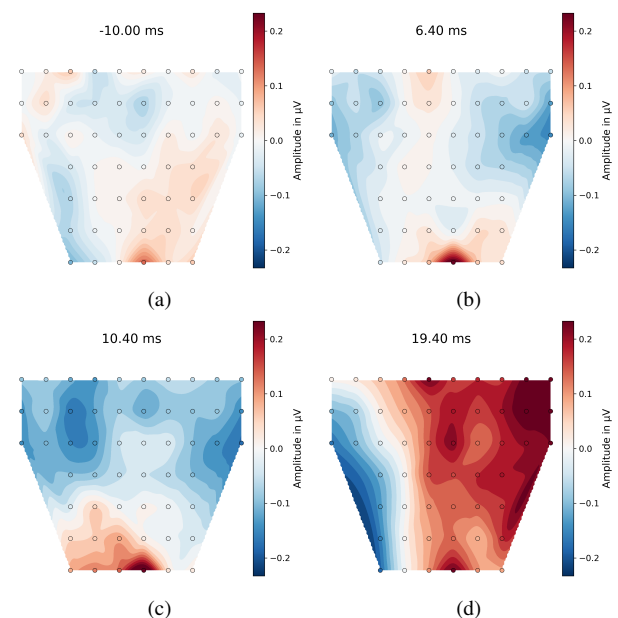


Figure 5: Topoplot of averaged ESG recording during UNS stimulation at (a) 10 ms pre stimulus, (b) 9-2.6 ms post stimulus, (c) 13-2.6 ms post stimulus and (d) 22-2.6 ms post stimulus.

fibers, purely sensory stimulation cannot be guaranteed when stimulating transcutaneously at anatomically complex sites such as the wrist or elbow. Consequently, contributions from motor fibers or muscle artifacts cannot be fully excluded. Furthermore, the current sample size limits statistical inference. Future studies with a larger cohort are required to determine whether observed components, such as the ESG deflection at approximately 13 ms post-stimulus, differ significantly from the REST condition and whether similar condition-specific effects can be robustly demonstrated for UNS stimulation.

Future work: Beyond increasing the number of participants, an important next step is to investigate functional connectivity between spinal and cortical activity [11]. Such analyses could provide insights into bidirectional communication within the sensorimotor system and help identify network-level interactions between brain and spinal cord during sensory processing. Another promising direction is the application of source localization techniques to both EEG [12] and ESG recordings [13]. While EEG source localization is well established, the extension of inverse modeling approaches to spinal recordings remains largely unexplored. Combining cortical and spinal source estimates could enable a more comprehensive spatiotemporal characterization of sensory signal propagation within the central nervous system.

CONCLUSION

The present results demonstrate that spinal evoked potentials recorded with ESG can resolve physiologically meaningful conduction delays associated with different peripheral stimulation sites. UNS stimulation elicited earlier spinal responses than MNS stimulation, with measured delays closely matching values predicted from peripheral nerve anatomy and conduction velocity. Similar latency differences were also observed in EEG recordings after participant-level averaging, and both stimulation paradigms exhibited comparable waveform morphology, suggesting activation of related spinal dynamics. These findings highlight the potential of ESG as a temporally sensitive method for investigating sensory signal propagation in the human nervous system and motivate future studies with larger cohorts and advanced connectivity and source localization analyses.

REFERENCES

- [1] Nierula B et al. A multichannel electrophysiological approach to noninvasively and precisely record human spinal cord activity. *PLoS biology*. 2024;22(10):e3002828.
- [2] Gabrieli G, Somervail R, Mouraux A, Leandri M, Haggard P, Iannetti GD. Electrical spinal imaging: A noninvasive, high-resolution approach that enables electrophysiological mapping of the human spinal cord. *PLoS Biology*. 2025;23(11):e3003116.
- [3] Shimoji K, Willis Jr W. *Evoked spinal cord potentials*. Springer (2006).
- [4] Di Pietro G et al. The n13 spinal component of somatosensory evoked potentials is modulated by heterotopic noxious conditioning stimulation suggesting an involvement of spinal wide dynamic range neurons. *Neurophysiologie Clinique*. 2021;51(6):517–523.
- [5] Jeanmonod D, Sindou M, Mauguiere F. The human cervical and lumbo-sacral evoked electrospinogram. data from intra-operative spinal cord surface recordings. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*. 1991;80(6):477–489.
- [6] Di Lionardo A et al. Modulation of the n13 component of the somatosensory evoked potentials in an experimental model of central sensitization in humans. *Scientific reports*. 2021;11(1):20838.
- [7] Gramfort A et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*. 2013;7(267):1–13.
- [8] Kothe C et al. The lab streaming layer for synchronized multimodal recording. *Imaging Neuroscience*. 2025;3:IMAG.a.136.
- [9] Bailey E, Nierula B, Stephani T, Maess B, Nikulin VV, Eippert F. Evaluating cardiac noise correction approaches for non-invasive electrophysiology of the human spinal cord. *Imaging Neuroscience*. 2025;3:IMAG-a.
- [10] Stetson DS, Albers JW, Silverstein BA, Wolfe RA. Effects of age, sex, and anthropometric factors on nerve conduction measures. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*. 1992;15(10):1095–1104.
- [11] Wimmer M, Kostoglou K, Müller-Putz GR. Measuring spinal cord potentials and cortico-spinal interactions after wrist movements induced by neuromuscular electrical stimulation. *Frontiers in Human Neuroscience*. 2022;16:858873.
- [12] Michel CM, He B. *Eeg source localization*. Handbook of clinical neurology. 2019;160:85–101.
- [13] Oberndorfer ME, Müller-Putz GR. Source localization of simulated neural signals in a cervical spinal cord model. *Journal of Neural Engineering*. 2026.

Volitional neural control with the CorTec Brain Interchange by an individual with chronic stroke

D. Sarma¹, H. Cho¹, A. Robert-Gonzalez², E. Stephan³, M. Bahadori-Nejad⁴, M. Buchheit⁵,
C.A. Gkogkidis⁵, M. Schuettler⁵, K.E. Weaver⁶, S.C. Cramer^{7,8}, C. Moritz^{2,3,9}, J.G. Ojemann¹,
J.A. Herron¹

¹Department of Neurological Surgery, University of Washington, Seattle, WA, USA

²Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA

³Department of Electrical & Computer Eng., University of Washington, Seattle, WA, USA

⁴Department of Bioengineering, University of Washington, Seattle, WA, USA

⁵CorTec GmbH, Freiburg, Germany

⁶Department of Radiology, University of Washington, Seattle, WA, USA

⁷Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA

⁸California Rehabilitation Institute, Los Angeles, CA, USA

⁹Department of Neurobiology and Biophysics, University of Washington, Seattle, WA, USA

E-mail: jeffherr@uw.edu

ABSTRACT: We report the first demonstration of volitional control of a Brain-Computer Interface (BCI) using the CorTec Brain Interchange by an individual with chronic hemiparetic stroke and no prior BCI experience. Subdural electrocorticographic (ECoG) recordings were acquired from sensorimotor cortex to facilitate voluntary control over a velocity-defined Pong paddle. Linear-discriminant and autoregressive ridge-regression decoders trained on mu and beta band-power features were trained on data collected during monthly study visits prior to the BCI testing. Across three short Pong sessions (96 min; 400 rallies) the participant attained 0.53–0.61 mean paddle hits per rally, above the static-paddle physics baseline (0.28; permutation $p \leq 0.013$ for all three sessions) and a stricter pipeline-replay shuffled baseline (~ 0.36 ; $p \leq 0.011$). Participant self-report demonstrated strategy-level insight into their control of motor-imagery and related game performance.

INTRODUCTION

Control of external devices by thought alone has been demonstrated across several recording modalities. These include non-invasive electroencephalography (EEG), intracortical electrode arrays with externalized percutaneous leads to body-external amplifiers [1], wireless implanted intracortical electronics [2], and implanted ECoG arrays with wireless electronics used for long-term communication in a patient with amyotrophic lateral sclerosis (ALS) [3]. Each modality varies in signal quality, coverage, invasiveness, and longevity. It remains to be determined which modality best supports durable deployment and maximally enhances functional status. Sensorimotor ECoG carries spatially specific mu, beta, and broadband high-gamma signals that track motor intent at millisecond resolution and have been used in

human BCIs since the first demonstrations in 2004 [4]. Since then, ECoG has driven neuroprosthetic skill acquisition [5], chronic cursor control in tetraplegia [6], simultaneous BCI and residual motor control [7, 8, 9], and has been used to read out error-related potentials during continuous BCI play [10]. Most of this work relied on research-grade acquisition hardware not amenable to long-term or free-living deployment.

The Brain Interchange (CorTec GmbH, Freiburg, Germany) is an implanted system intended for long-term, research deployment [11, 12]. Using its programming interfaces, our group at the University of Washington developed an open-source real-time stimulation and streaming platform (OMNI-BIC) to support rapid prototyping through modular system integration [11, 13]. The recording performance of the various components of the integrated system have been positively evaluated in a clinical benchmark study with epilepsy patients [12, 13], while separately, the implanted Brain Interchange has been used successfully for >500 days of longitudinal wireless ECoG recording in non-human primates [14]. To our knowledge, no BCI-type demonstration of closed-loop volitional control of a virtual effector with the Brain Interchange has previously been reported in humans.

Stroke is the second leading cause of long-term adult disability worldwide [15], and individuals with chronic upper-limb impairment after stroke are a priority population for restoration of motor capabilities through BCI. Nonetheless, most chronically implanted neural interfaces have been tested by people with tetraplegia or ALS populations, where cortical motor representations are largely intact. Cortical motor representations are context-dependent even in healthy micro-ECoG recordings [16], and potential post-stroke reorganization adds further variability, challenging decoder reliability and cross-session feature stability. We present a case

Table 1: Brain Interchange Specifications - Stroke BCI

Feature	Value
Channels	32 (sense, sense reference, stim, or ground)
ECoG arrays	2 arrays, 2×8 contacts, 10 mm contact spacing
Electrode contacts	2.4 mm Ø, PtIr
Recording sampling rate	32× 1 kS/s @ 16 bit
Bandpass filtering	~5 Hz to 320 Hz
Amplification options	750×, 380×, 190×, 95×
Max. stim. amplitude	5.4 mA
Sense latency: brain to PC	10 ms
Wireless power supply	130 kHz inductive link
Wireless RF data link	2.54 GHz, 5Mbit/s, AES-128 encrypted, ½ -duplex

report of a participant with chronic hemiparetic stroke, no prior BCI exposure and no recent video-game experience, implanted with the Brain Interchange over peri-lesional sensorimotor cortex. He played a velocity-controlled Pong-style game driven by band-power decoders trained from monthly study visits collected prior to the BCI-testing period. We hypothesized that volitional paddle control would be feasible from band-power features over a small responsive-channel set despite chronic post-stroke reorganization.

MATERIALS AND METHODS

Participant

The participant was enrolled and implanted with the Brain Interchange as part of an ongoing NIH-funded clinical study on the use of neural stimulation to enhance stroke rehabilitation (NIH UH3NS121565; clinicaltrials.gov ID NCT06506279). The protocol for the primary study, which received an investigational device exemption (IDE) from the US Food and Drug Administration and approval from the University of Washington Institutional Review Board prior to any study activities, includes a dedicated period at the end of the implanted study window for next-generation therapy development. The right-handed male participant, 52 years old at study enrollment with his most recent stroke 4 years prior, had moderate-to-severe upper-limb impairment (Upper-Extremity Fugl-Meyer 37 at baseline, 47 at the time of BCI experiments following study-provided intensive rehabilitation therapy). BCI activities described here began after the parent study's therapy and monitoring program, approximately nine months after implantation.

The Brain Interchange System

The Brain Interchange System (Fig.1a) is a 32-channel neural interface platform that provides sensing and stimulation access to the brain. In the configuration used in this study, two ECoG electrode arrays were used. The implant (Fig. 1b) is wirelessly powered via an inductive link provided by an external head piece, which aligns to the implant using magnets. The head piece is powered and controlled by the communication unit (com-unit), which also provides a wireless radio frequency (RF) data

link to the implant. A personal computer (PC) running Microsoft Windows communicates with the com-unit and powers it via USB 2.0. The electronics are seated in a recess in the internal table of the skull. The specifications of the system used in this study are listed in Tab. 1. The peri-lesional electrode reconstruction for this participant is shown in Fig. 1c, and a representative gameplay frame in Fig. 1d.

Offline preparation

Monthly study visits prior to the BCI testing period combined a digitized hand grasp task, a standard arm task, and a resting-state recording. Adaptive linear discriminant analysis (LDA) decoders trained on Hilbert-envelope log-band-power features at cued movement vs inter-trial rest (80/20 within-session train/test split) reached 72.8 % binary decoding accuracy on hand-task data and 89.2 % on a separately collected physical-to-imagined transfer block — in which the participant first executed cued physical movements then attempted them as motor imagery within the same session — confirming that decoded motor intent generalized to imagined movement.

The deployed online decoder was a per-channel autoregressive ridge regression (AR; history length 5, $\lambda = 1 \times 10^{-3}$) on mu (8–13 Hz) and beta (13–30 Hz) Hilbert-envelope log-band-power features. High-gamma (70–150 Hz) was extracted by the real-time pipeline but excluded from deployed AR decoders given inconsistent cross-session modulation in this participant. The deployed three-channel set (CH9, CH26, CH28; Fig. 1c) was selected as the most consistent move-vs-rest band-power responders across the prior parent-study visits, accommodating substantial visit-to-visit variance and a feature-shift observed at the monitoring visit immediately preceding the BCI testing period. Smoothed AR variants applied a Gaussian $\sigma \approx 500$ ms to the velocity target. AR outputs were not drift-corrected; sub-

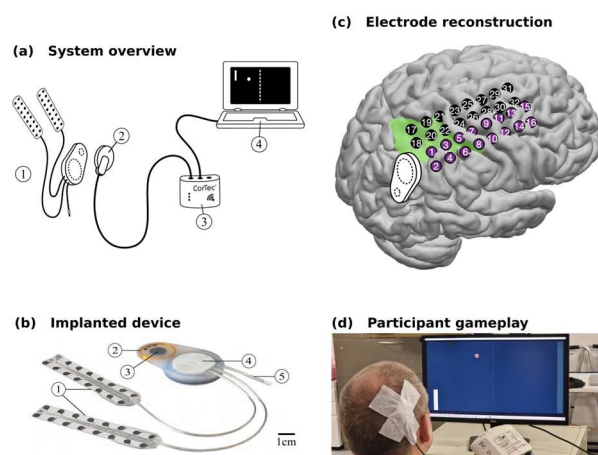


Figure 1: (a) Brain Interchange System, consisting of Implant ①, head piece ②, com-unit ③, and PC ④. (b) Implant with ECoG arrays ①, inductive coil ②, alignment magnet ③, electronics ④, and ground lead ⑤. (c) MRI reconstruction of cortex with electrode contact locations and outlined electronics; green area: stroke-affected region. (d) Gameplay with head piece taped over the implant for additional fixation.

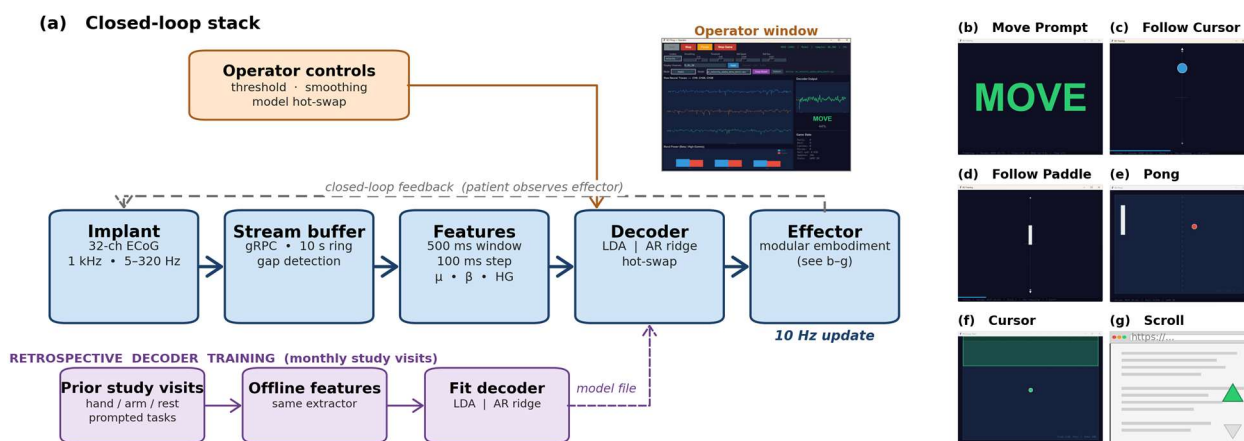


Figure 2: Closed-loop stack and participant interface. (a) Offline training arm feeds saved model into hot-swappable decoder in online arm (100 ms tick, 10 Hz update); operator controls expose threshold, smoothing, and model selection; dashed arc marks closed-loop feedback; inset shows the Operator window. (b–g) Participant interface: Prompt cues MOVE/REST for training-data, tasks of ‘Follow Cursor’ and ‘Follow Paddle’ trajectory tracking with effectors, velocity-controlled Pong, center-out Cursor, and OS-level browser Scroll.

threshold decay toward zero (see §Pong task) provided implicit zero-restoration. LDA decoders were less usable online than AR variants in this participant and were not deployed during Pong play.

Real-time BCI system

The closed-loop BCI system (Fig. 2a) was organized in four stages. A gRPC client connected to OMNI-BIC [11] maintained a 10-second rolling buffer of the 32-channel ECoG stream of macroscale cortical potentials. Every 100 ms (10 Hz update), the most recent 500 ms was passed through a feature extractor (Welch or Hilbert-envelope spectral power in μ 8–13 Hz, β 13–30 Hz, high-gamma 70–150 Hz). The decoder produced a label and confidence value that drove a velocity-controlled Pong paddle on the patient display; an operator-facing window driven from the same tick showed raw neural traces, per-band power bars, and a 30-second rolling history of decoder output (Fig. 2a). All ticks were logged to HDF5 with derived features and decoder outputs.

Analysis pipeline, tick scheduling, and latency

The pipeline ran as a self-paced cycle: after each iteration completed, the next was scheduled to run 100 ms later. Each cycle read the latest neural data, computed features over a 500 ms analysis window, ran the decoder, and updated the paddle. The operator window (Fig. 2a, inset) allowed for three types of adjustments: a decoder-model selector; an exponential-smoothing slider on the decoder output; and a confidence-threshold slider to change a minimum threshold of confidence before modulating the output, with speed proportional to the over-threshold margin. All three were set during initialization from brief participant feedback and left static throughout the retained use-trials; they are reported for transparency and not treated as an experimental variable.

Session protocol

Each in-person session followed the same short protocol. First, a cueing block (≈ 3 minutes) presented MOVE/REST prompts on the patient display, with cue transitions timestamped as ground-truth labels (Fig. 2b). Second, a trajectory-tracking block (≈ 3 –5 minutes)

displayed either a circular cursor (Fig. 2c) or vertical pong-style paddle (Fig. 2d) moving along a fixed trajectory (sinusoidal sweep or step-and-hold) irrespective of decoder output, while the relative decoder state was available in the operator window for tuning. Dynamically following targets was intended as a guide to executing faster motor behaviors than the slow, controlled rehabilitation movements the participant had been performing, more closely mirroring real-time game control. Motor imagery was explained verbally prior to the first BCI session; the participant reported strong familiarity from his youth as a wrestler, where motor-imagery rehearsal had been a routine athletic-training tool. Because arm movements yielded more consistent offline response features than hand movements in this participant, cued retraining used an attempt-to-topple push gesture against the clinician’s monitor for overt rehearsal, with the corresponding motor-imagery instruction being to imagine moving against this resistance for upward paddle motion and disengagement for downward motion. After these trainings, the participant engaged with the Pong game (Fig. 2e); a 2-D cursor task (Fig. 2f) and a browser-scroll controller (Fig. 2g) were also developed but, with limited time available, the participant requested focusing on Pong.

Pong task

The patient display showed a unit-square arena where all positions and velocities are given in arena units (AU) and the arena spans 0–1 AU on each axis. A ball was launched from the right edge at a random angle ($\pm 36^\circ$) and bounced off the top, bottom, and right walls. The paddle was velocity-controlled: the decoder confidence above the positive threshold set a positive target paddle velocity (proportional to confidence – threshold, capped); below the negative threshold the target velocity was negative; sub-threshold output decayed the target velocity toward zero. Position was integrated from velocity per tick. The ball’s speed increased by 0.002 AU/tick on each successful catch up to a cap of 0.08 AU/tick (≈ 0.8 AU/s at the 10 Hz update rate). A ball that

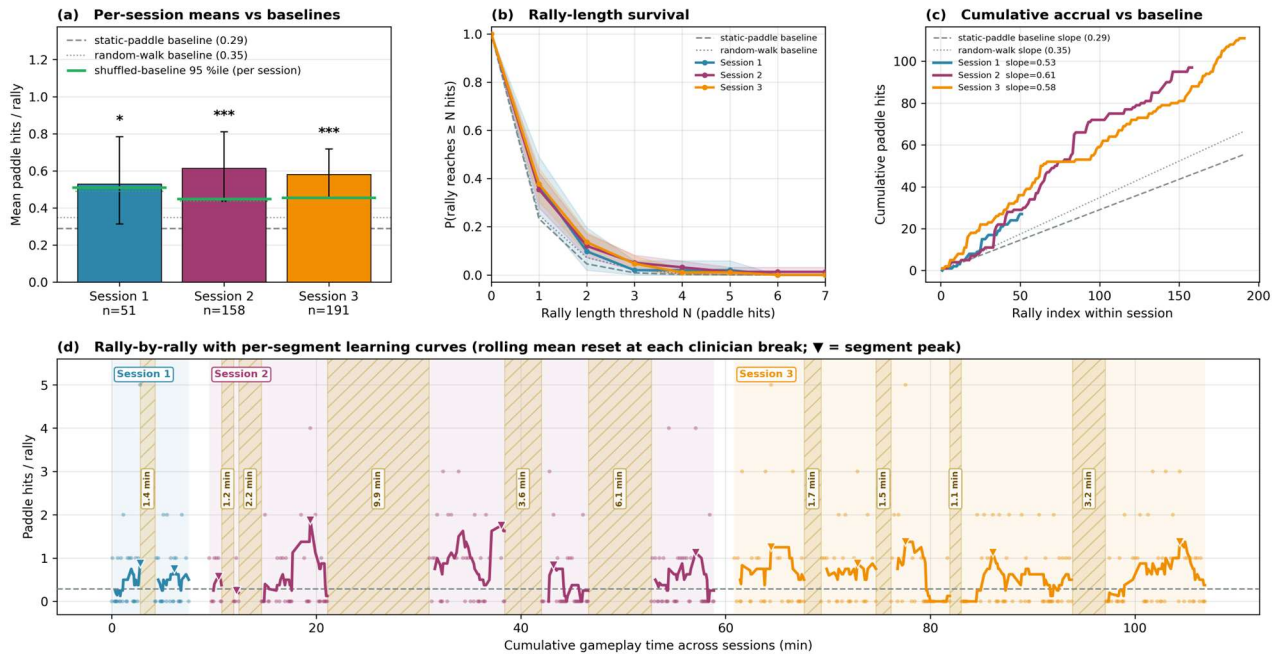


Figure 3: BCI results across sessions. (a) Per-session mean hits/rally with 95 % bootstrap CIs; static-paddle (0.28) and random-walk physics baselines (dashed); asterisks denote permutation $p < 0.05/0.01/0.001$ vs static. (b) Rally-length survival per session against baseline curves. (c) Cumulative hits with baseline slopes (dashed). (d) Rally-by-rally timeline with clinician-initiated breaks (≥ 60 s) shaded; per-segment rolling mean resets at each break, ▼ marks segment peaks.

passed the paddle ended the rally; the next rally launched immediately on the following tick (no explicit inter-trial interval). We define a rally as the span from one ball-launch event to the next miss (or to a game stop), and count paddle hits as the successful events within a rally.

Data analysis

For each session multiple decoders were evaluated. Trial (rally) segments per-decoder containing fewer than 10 rallies were excluded as truncated. Per-session means are reported with 95 % bootstrap confidence intervals after rallies are resampled with replacement 2,000 times. Permutation p-values are the fraction of resampled means that fell at or below the static-paddle baseline. For each decoder $\times$ session cell with at least 10 rallies, 10,000-resample bootstrap confidence intervals are reported alongside the mean hit rate, and inclusion-threshold sensitivity was assessed under minimum rally-count thresholds of 5, 10, 15 and 20. For each inbound rally leg we additionally computed the peak paddle-versus-ball y-position cross-correlation (± 1.5 s lag window) and the terminal positional error at ball arrival.

We observed hit rates against two reference distributions. (i) A physics-matched baseline: 2,000 re-simulations with the paddle held at arena center give 0.28 hits/rally as the geometry-only chance level. (ii) A stricter pipeline-replay baseline: for each retained rally we circularly shift the recorded decoder trace by a random lag (5 s to $T - 5$ s) and re-integrate the paddle through the recorded per-tick smoothing, threshold and control-mode values; unshuffled replay reproduces the recorded paddle at $r = 0.91$, confirming a faithful loop. 1,000 shuffles per trials pooled within session give the pipeline-replay distribution (mean ~ 0.36 hits/rally); observed hit rates exceeding its 95th percentile cannot be attributed to

chance alignment of the actual decoder dynamics with the ball, because temporally shifted versions of the same trace fail to produce comparable rates [17].

RESULTS

Analysis pipeline tick performance

Across all retained rallies, the control loop performed stably. Median tick interval was 124 ms (95th percentile 132 ms), with fewer than 1% of ticks delayed beyond 150 ms. This reflects a 100 ms scheduled delay between iterations plus roughly 24 ms of per-iteration processing (gRPC stream poll, feature extraction, decoder evaluation, paddle command update, and log write). This update rate is comparable to those used in prior ECoG BCI control studies [4, 6, 10].

Pong performance above chance

The participant completed 3 sessions (1/day) totaling 96.1 minutes of gameplay and 400 rallies, achieving 235 paddle hits. Mean hit rates were 0.53 per rally on Session 1 (95 % CI 0.31–0.78, $n = 51$), 0.61 on Session 2 (CI 0.44–0.81, $n = 158$), and 0.58 on Session 3 (CI 0.46–0.72, $n = 191$). All three sessions exceeded the 0.28 physics-imposed static-paddle baseline at permutation ($p = 0.013$, 0.0001 and < 0.0001 , respectively; Fig. 3a). Rally-length survival (Fig. 3b) was above both the static-paddle and random-walk physics baselines in every session, with a peak of nine consecutive hits in Session 2. Cumulative-hit slopes of 0.53, 0.61, 0.58 per rally (Fig. 3c) were each roughly twice the 0.28 static baseline. Mean hit rate did not differ significantly between sessions (all pairwise $p > 0.65$; differences ≤ 0.08 hits/rally), indicating sustained performance across the testing period.

The stricter pipeline-replay baseline (Fig. 4) produced

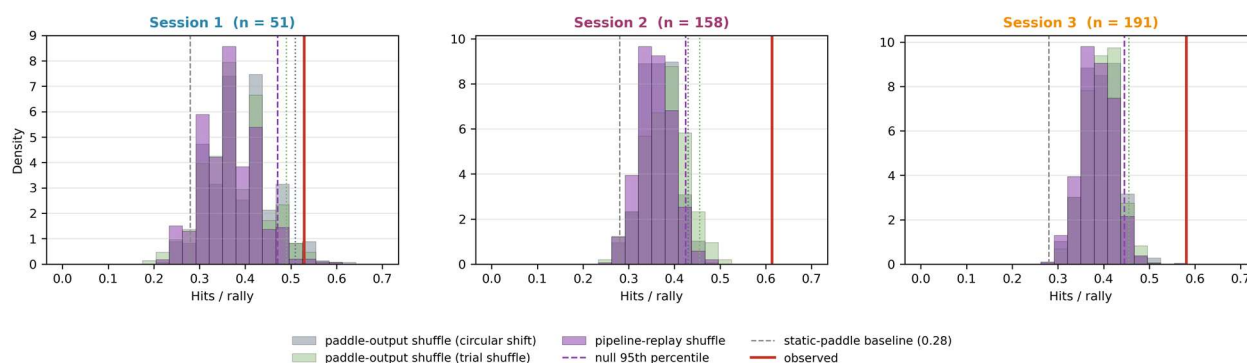


Figure 4: Per-session chance comparisons (Sessions 1–3, $n = 51, 158, 191$ rallies). Density histograms overlay paddle-output shuffle nulls (circular shift in grey, trial shuffle in green) and the stricter pipeline-replay shuffle null (purple), which preserves decoder output and paddle dynamics but destroys their alignment with the ball. Vertical lines mark observed catch rate (red; 0.53/0.61/0.58), 95th percentiles, and the static-paddle physics baseline (0.28). Pipeline-replay $p = 0.011 / <0.001 / <0.001$.

per-session 95th-percentile values of 0.47, 0.42 and 0.45 (pooled mean ~ 0.36 per rally, comparable to the random-walk paddle baseline). Observed hit rates exceeded the pipeline-replay 95th percentile at $p = 0.011$ in Session 1 and $p < 0.001$ in Sessions 2 and 3. A weaker paddle-output shuffle gave consistent 95th percentiles (0.44–0.51 across sessions) and was exceeded at comparable levels. Partitioning rallies by the clinician-initiated breaks (≥ 60 s) reveals a consistent within-segment learning shape in each session (Fig. 3d): performance rises from a low initial level, peaks within the epoch ($\blacktriangledown$ marker), then declines before the next break. In each between-break segment with enough rallies on either side of the peak (9 of 13), pre-peak hit rate exceeded post-peak by 0.43 hits/rally (signed-rank $p = 0.008$), supporting a within-segment post-peak decline.

Decoder choice dominated variability

Of the eight decoder $\times$ session cells with at least 10 rallies, three had lower 95 % bootstrap CI bounds above the 0.28 static baseline. The strongest configuration was the smoothed AR variant on the three best-responding channels used for most of Session 2 (0.78 hits per rally; CI 0.53–1.09; $n = 95$ rallies). Two further AR variants on Session 3 (best-3 and CH9/26/28 triads) tied at 0.64 hits per rally. Across inclusion-threshold sensitivities ($n \geq \{5, 10, 15, 20\}$) the headline per-session means shifted by no more than 0.03, and every session $\times$ threshold combination remained above the static baseline at permutation $p \leq 0.017$.

Paddle tracked position, not direction

Across 628 inbound ball-legs the peak paddle-versus-ball y-position cross-correlation averaged $r = 0.39$ to 0.48, peaking at zero lag. The instantaneous velocity correlation was not distinguishable from zero on any session. Median terminal positional error at ball arrival was 0.22–0.30 arena units, roughly twice the paddle half-height (0.12), consistent with the observed miss rate. The control signal supported coarse positional control but not fine-grained trajectory matching.

Participant-reported experience

The participant had no prior BCI experience and reported “not having touched a video game in years”. He reported during Session 3 that the paddle felt “very responsive

compared to what I’m thinking”. Self-insight followed: noting that the paddle tended to drift up in the final moments before ball arrival, the participant remarked “I think I’m celebrating hitting the ball before it gets there”, and generalized that control was “more the whole brain... it doesn’t really seem to be connected to the movement of my arms,” analogous to focusing on the position of a wrench in space rather than arm movement to achieve a desired position. The participant attributed fluctuating performance to internal state (“a bit of brain fog”) rather than to the system, ultimately finding the design “very achievable... very exciting.”

DISCUSSION

These data establish feasibility of closed-loop ECoG BCI task control with the CorTec Brain Interchange, providing the first such demonstration with a chronic ECoG array fully implanted in a stroke patient. Hit rates exceeded the physics-imposed static-paddle baseline on every session ($p \leq 0.02$) and the more stringent pipeline-replay shuffled baseline (~ 0.36 mean per rally), which preserved the participant’s own decoder and paddle-motion statistics but shuffled their temporal alignment to the ball [17]. Survival curves and cumulative-hit slopes confirm that above-baseline accrual was sustained over dozens of minutes of gameplay rather than concentrated in isolated hits.

Although the paddle mechanism was velocity-defined, the participant’s achieved control was position-dominated rather than fine instantaneous velocity tracking: the paddle tended to hover as the ball arrived, suggesting some coarse control and the ability to track target trajectory. The within-segment rise–peak–decline learning shape visible in Fig. 3d suggests adaptation on a minutes-timescale, with internal-state fluctuation possibly contributing to the within-segment decline; self-report placed the limiting factor on ‘brain fog’ rather than on the system. Longer breaks were initiated by the attending occupational therapist on observation of participant fatigue or strain, not by participant request. In fact, the participant consistently expressed a desire to continue, so breaks are best interpreted as clinician-

determined breaks to prevent fatigue.

Per-decoder rally counts remain small for fine-grained inference, addressed here by 10,000-resample bootstrap confidence intervals per decoder $\times$ session cell and by inclusion-threshold sensitivity checks. Decoder selection prioritized the most consistent neural dynamics available in this individual with chronic stroke. A continuous-velocity decoder tuned to multi-band features, longitudinal stability, and cross-session decoder transfer are the natural next directions for chronic BCI deployment in this population [6, 18].

CONCLUSION

We present the feasibility of closed-loop volitional neural-control using the Brain Interchange for an individual living with chronic stroke. Our preliminary results highlight the rapid development of rudimentary control and motivate further experiments with future participants. We anticipate expanding BCI technology for people with neurological disease and sensorimotor dysfunction will enable a range of neuroprosthetic therapies and interventions to improve recovery.

ACKNOWLEDGEMENTS

We thank the study participant for his time, engagement, and feedback during the BCI testing sessions. Author MB-N is supported by the Weill Neurohub. Financial support for the BCI experiments was provided by CorTec GmbH.

REFERENCES

- [1] Hochberg LR, Bacher D, Jarosiewicz B, Masse NY, Simeral JD, Vogel J, Haddadin S, Liu J, Cash SS, van der Smagt P, Donoghue JP. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* 485(7398):372–375 (2012).
- [2] Sahu C. Neuralink and Its Advantages: Advancements in BCI Technology. *IJSAT-Internat. Journal on Science and Technology*. 2025 Jul 3;16(3).
- [3] Vansteensel MJ, Leinders S, Branco MP, Crone NE, Denison T, Freudenburg ZV, Geukes SH, et al. Longevity of a brain–computer interface for amyotrophic lateral sclerosis. *N Engl J Med* 391(7):619–626 (2024).
- [4] Leuthardt EC, Schalk G, Wolpaw JR, Ojemann JG, Moran DW. A brain–computer interface using electrocorticographic signals in humans. *J Neural Eng* 1(2):63–71 (2004).
- [5] Wander JD, Sarma D, Johnson LA, Fetz EE, Rao RPN, Ojemann JG, Darvas F. Cortico-cortical interactions during acquisition and use of a neuroprosthetic skill. *PLoS Comput Biol* 12(8): e1004931 (2016).
- [6] Wang W, Collinger JL, Degenhart AD, Tyler-Kabara EC, Schwartz AB, Moran DW, et al. An electrocorticographic brain interface in an individual with tetraplegia. *PLoS ONE* 8(2): e55344 (2013).
- [7] Cheung W, Sarma D, Scherer R, Rao RPN. Simultaneous brain–computer interfacing and motor control: Expanding the reach of non-invasive BCIs. *Proc IEEE EMBC* 2012:6715–6718.
- [8] Bashford L, Wu J, Sarma D, Collins K, Rao RPN, Ojemann JG, Mehning C. Concurrent control of a brain–computer interface and natural overt movements. *J Neural Eng* 15(6):066021 (2018).
- [9] Sarma D. Expanding the Reach of Electrocorticographic Brain–Computer Interfaces: A Bimanual Approach. University of Washington (2017). <http://hdl.handle.net/1773/39952>.
- [10] Wilson NR, Sarma D, Wander JD, Weaver KE, Ojemann JG, Rao RPN. Cortical topography of error-related high-frequency potentials during erroneous control in a continuous control brain–computer interface. *Front Neurosci* 13:502 (2019).
- [11] Cho H, Ojemann JG, Herron JA. Open Mind Neuromodulation Interface for the CorTec Brain Interchange (OMNI-BIC). 11th Int IEEE/EMBS Conf Neural Eng (NER) 2023:1–6.
- [12] Ayyoubi AH, Fazli Besheli B, Quach MM, Gavvala JR, Goldman AM, Swamy CP, Bartoli E, et al. Benchmarking signal quality and spatiotemporal distribution of interictal spikes in prolonged human iEEG recordings using CorTec wireless Brain Interchange. *Sci Rep* 14(1):2652 (2024).
- [13] Cho H, Benjaber M, Gkogkidis CA, Buchheit M, et al. Development and evaluation of a real-time phase-triggered stimulation algorithm for the CorTec Brain Interchange. *IEEE Trans Neural Syst Rehabil Eng* 32:3625–3635 (2024).
- [14] Yang H, Fukuma R, Namima T, et al. Longitudinal multitask wireless electrocorticography data from two fully implanted non-human primates. *Scientific Data* 13:63 (2026).
- [15] Krishnamurthi RV, Ikeda T, Feigin VL. Global, regional and country-specific burden of ischaemic stroke, intracerebral haemorrhage and subarachnoid haemorrhage: a systematic analysis of the Global Burden of Disease Study 2017. *Neuroepidemiology* 54:171–179 (2020).
- [16] Kuo CH, Blakely TM, Wander JD, Sarma D, Wu J, Casimo K, Weaver KE, Ojemann JG. Context-dependent relationship in high-resolution micro-EEG studies during finger movements. *J Neurosurg* 132(5):1358–1366 (2019).
- [17] Combrisson E, Jerbi K. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods* 250:126–136 (2015).
- [18] Hammer J, et al. The role of ECoG magnitude and phase in decoding position, velocity, and acceleration during continuous motor behavior. *Front Neurosci* 7:200 (2013).