 10.3217/978-3-99161-093-9-038

Benchmarking Machine Learning strategies to classify Steady-State Auditory Evoked Potentials.

Henrique Lefundes da Silva[1], Lenaïg Guého[1,2], Cyril Plapous[1], Laurent Bougrain[2,3], Patrick Hénaff[4], Rozenn Nicol[1]

[1]NUX, Orange Innovation, F-22300 Lannion, France
[2]Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
[3]Sorbonne Université, ICM, CNRS, Inria, Inserm, Paris, France
[4]Univ 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.

10.3217/978-3-99161-093-9-038

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]

 10.3217/978-3-99161-093-9-038

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).

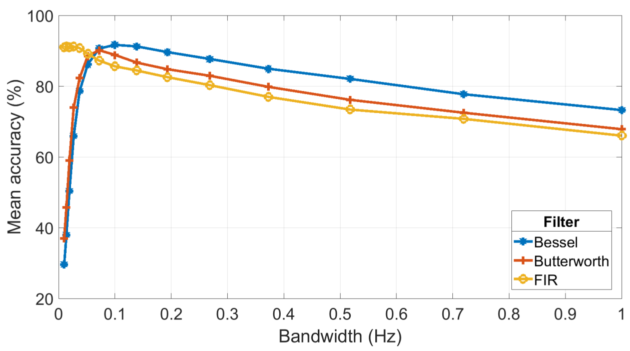


Figure 1: Effect of filter type and bandwidth on classification performance, using 2-second non-overlapping windows. Results are averaged across all subjects.

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.

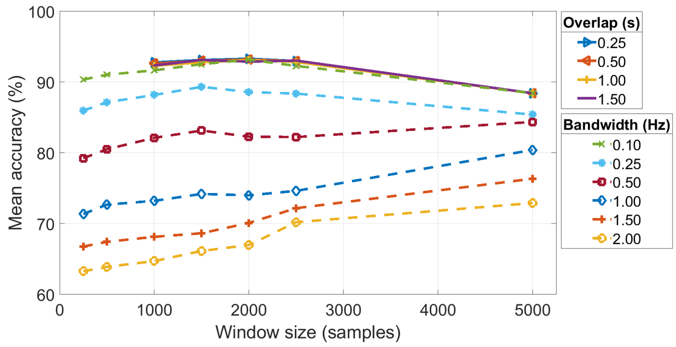


Figure 2: Effect of window size for various overlapping configurations and bandwidths (without overlap), using the optimal Bessel filter (order 4, 0.1 Hz bandwidth). Results are averaged across all subjects.

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).

 10.3217/978-3-99161-093-9-038

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

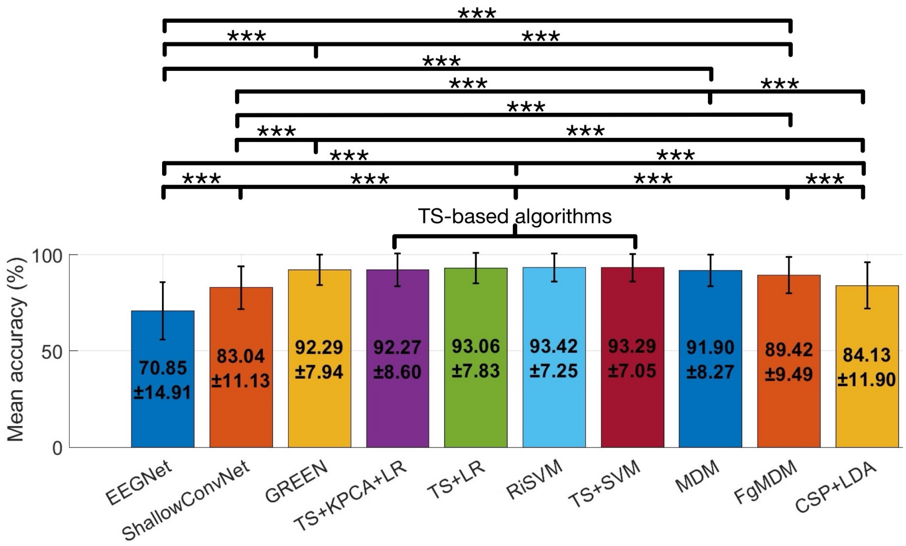
Using an appropriate filter can enhance classification performance by over 15% (Fig. 1). The Bessel filter yields the best results, while the FIR filter performs the worst. However, accurately isolating the modulation frequency is the most influential parameter. Bessel filters offer flat group delay 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 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.

Table 2: Deep Learning Models' Hyperparameters.

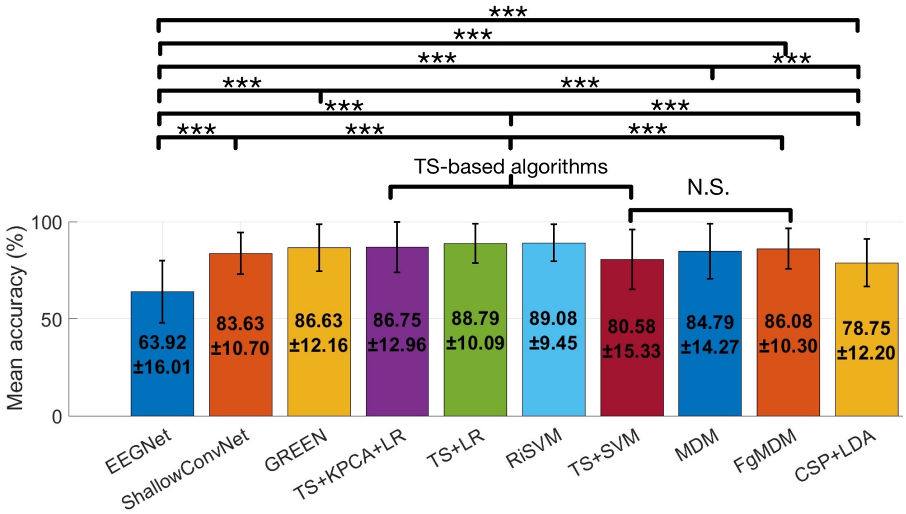
Hyperparameter	Value WS	Value CS
General		
Batch Size	16	16
Loss	BCEWithLogits	BCEWithLogits
Optimizer	Adam	Adam
Epochs	20	30
Learning Rate	0.0005	0.001
EEGNet		
F1	2	2
D	12	12
F2	24	24
kernel_length	32	32
Batch Size	16	16
SCN		
n_filters_time	2	2
n_filters_spat	8	12
pool_time_length	100	100
pool_time_stride	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

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,

10.3217/978-3-99161-093-9-038



(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.

10.3217/978-3-99161-093-9-038

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, Papadopoulo 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] Schirrmeister 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.