NOTE • OPEN ACCESS

Benchmarking brain–computer interface algorithms: Riemannian approaches vs convolutional neural networks

To cite this article: Manuel Eder *et al* 2024 *J. Neural Eng.* **21** 044002

Journal of Neural Engineering

NOTE

Benchmarking brain–computer interface algorithms: Riemannian approaches vs convolutional neural networks

Manuel Eder[1,*], Jiachen Xu[1] and Moritz Grosse-Wentrup[1,2,3]

[1] Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna, Vienna, Austria
[2] Data Science Network @ Uni Vienna, University of Vienna, Vienna, Austria
[3] Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria
[*] Author to whom any correspondence should be addressed.

E-mail: ederm42@univie.ac.at

Keywords: brain-computer interface, electroencephalography, classification, benchmarking, Riemannian geometry, convolutional neural network, machine learning

OPEN ACCESS

RECEIVED
13 March 2024

REVISED
3 June 2024

ACCEPTED FOR PUBLICATION
19 July 2024

PUBLISHED
21 August 2024

Abstract

Objective. To date, a comprehensive comparison of Riemannian decoding methods with deep convolutional neural networks for EEG-based brain–computer interfaces remains absent from published work. We address this research gap by using MOABB, The Mother Of All BCI Benchmarks, to compare novel convolutional neural networks to state-of-the-art Riemannian approaches across a broad range of EEG datasets, including motor imagery, P300, and steady-state visual evoked potentials paradigms. *Approach.* We systematically evaluated the performance of convolutional neural networks, specifically EEGNet, shallow ConvNet, and deep ConvNet, against well-established Riemannian decoding methods using MOABB processing pipelines. This evaluation included within-session, cross-session, and cross-subject methods, to provide a practical analysis of model effectiveness and to find an overall solution that performs well across different experimental settings. *Main results.* We find no significant differences in decoding performance between convolutional neural networks and Riemannian methods for within-session, cross-session, and cross-subject analyses. *Significance.* The results show that, when using traditional Brain-Computer Interface paradigms, the choice between CNNs and Riemannian methods may not heavily impact decoding performances in many experimental settings. These findings provide researchers with flexibility in choosing decoding approaches based on factors such as ease of implementation, computational efficiency or individual preferences.

Abbreviations

BCI	brain–computer interface
MOABB	Mother Of All BCI Benchmarks
EEG	electroencephalography
MI	motor imagery
ERP	event-related potential
SSVEP	steady-state visual evoked potential
CNN	convolutional neural network
RMDM	Riemannian minimum distance to mean
TGSP	tangent space
SVM	support vector machine
LDA	linear discriminant analysis
LR	logistic regression
Cov	Covariance
ERPCov	ERP covariance
XCov	Xdawn covariance
FB	filterbank

1. Introduction

Brain–computer interfaces (BCIs) have evolved to be used for various tasks ranging from medical/assistive technology to the entertainment/gaming industry. These include for example, moving a cursor on a computer screen [1], controlling a mobile robot [2], controlling prosthetic devices [3], neurorehabilitation [4], or decoding speech from neural activity [5]. All BCI tasks and applications come with their own problems and challenges, so in order to properly find, extract and classify task-relevant neural states, several different EEG-based BCI paradigms are used accordingly. Three of the most popular paradigms are MI, P300–based spellers, and SSVEPs [6]. These paradigms work in fundamentally different ways. MI is often used for self-paced (asynchronous) BCIs,

but with many experiments following a cued protocol, while P300-based systems and SSVEPs are usually used for stimulus-paced (synchronous) BCIs [7]. Often, they can be used as potential approaches to solve the same tasks (e.g. spelling tasks, moving a cursor, etc). Depending on the task, BCI systems can implement a single paradigm or use multiple paradigms for hybrid approaches [6].

Given the diversity of experimental paradigms and machine learning pipelines for brain decoding [8], it is natural to ask whether any particular machine learning framework performs best across a broad range of experimental paradigms and datasets. The MOABB—the Mother Of All BCI Benchmarks—is an open-source package for benchmarking BCIs written in Python that allows such a comparison across a broad range of publicly available datasets [9]. While MOABB has been used to compare various machine learning pipelines, two of the most successful machine learning frameworks for BCI decoding, Riemannian decoding [10] and deep convolutional networks [11, 12], have not yet been rigorously compared with each other [13].

In this work, we use the MOABB to fill this gap. In particular, we compare EEGNet [12], a compact CNN, along with a deep CNN [11] and a shallow CNN [11] with state-of-the-art Riemannian methods on multiple publicly available MOABB datasets as shown in table 2. For within-session, cross-session and cross-subject decoding, we do not find significant differences between the two frameworks. All code is available at https://github.com/ederm42/BCI-riemannian-vs-eegnet.

2. Methods

To compare the CNNs [11, 12] with state-of-the-art Riemannian approaches, several literature sources were used to set up various MOABB processing pipelines for each paradigm. The main source used as a reference was 'A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update' [8]. In order to reach proper coverage of different Riemannian methods, additional pipelines were implemented, as found in multiple publications [14–16] and in code published by MOABB [17] and EEGNet [18]. An overview of all included decoding pipelines is shown in table 1.

The Riemannian pipelines use the implementations of Covariance from the pyRiemann Python package [19]. TGSP mapping and all Riemannian algorithms also make use of the pyRiemann package. All classifiers are implementations provided by the scikit-learn Python package [20]. All CNNs, as implemented by the EEGModels Project [18], were trained on an NVIDIA GeForce RTX 3090 using TensorFlow 2.8.0, CUDA 11.2 and cuDNN 8.1.0. The statistical analysis, as well as the creation of the Gardner–Altman estimation plots, was done using the 'DABEST: Data Analysis with Bootstrap-coupled ESTimation' [21, 22] Python package.

2.1. Decoding pipelines

2.1.1. MI

For MI, the focus was on evaluating the left–right imagery paradigm. For all datasets, MOABB default MI band-pass filter for 8–30 Hz was applied, and the EEG data was resampled to 128 Hz. All Riemannian pipelines implemented for MI calculate the covariance matrix first. One pipeline directly uses the RMDM as the classifier, while the others first project the covariance matrix into TGSP and afterward classify using either a SVM (TGSP + SVM) or LDA (TGSP + LDA). For the CNNs, parameters were left on default as implemented by the EEGModels project [18]. These parameters were standardized, validated and optimized for 128 Hz EEG data.

2.1.2. P300

For the P300 datasets, MOABB default P300 bandpass filter for 1–24 Hz was applied and the EEG data was resampled to 128 Hz. The pipeline using RMDM as a classifier use special covariance matrices specifically made for ERP/P300 paradigms, XDawnCovariance (XCov + RMDM), as implemented in the MOABB package. The other pipelinse first projects the XDawnCovariance into TGSP and afterward use a linear SVM (TGSP + SVM) or LR (TGSP + LR) as its classifier. For the CNNs, parameters were left on default as implemented by the EEGModels project [18]. These parameters were standardized, validated and optimized for 128 Hz EEG data.

2.1.3. SSVEP

For all SSVEP approaches, the EEG data was resampled to 128 Hz. A FB approach was used with multiple bandpass filters, centered ± 0.5 Hz around the relevant frequencies. The Riemannian pipelines also estimate the covariance matrix first and afterward use either RMDM or a TGSP projection followed by LR (TGSP + LR). For the CNNs, parameters were left on default as implemented by the EEGModels project [18]. These parameters were standardized, validated and optimized for 128 Hz EEG data.

2.2. Evaluation methods

The evaluation methods offered by MOABB are, depending on the dataset, within-session evaluation, cross-session evaluation, and cross-subject evaluation. MOABB splits the data into training/test trials accordingly and, by default, evaluates using cross-validation. For left–right MI and P300 - the binary classification tasks—ROC-AUC is used as the evaluation metric. The SSVEP paradigm, given that it is a multi-class classification task, accuracy is used as the metric.

Table 1. Description and references of pipelines used for the evaluations. All paradigms were evaluated with three different CNNs—EEGNet, deep ConvNet and shallow ConvNet. For each paradigm, the Riemannian Minimum Distance to Mean (RMDM) and two tangent space mapping approaches (TGSP) were implemented, with the TGSP classifier depending on the paradigm. The TGSP classifiers are either a Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) or Logistic Regression (LR). For the P300 RMDM classifier, the XDawnCovariance (XCov) is used to estimate the covariance matrices.

Paradigm	Pipeline	References
MI	EEGNet	[12]
	ConvNet (deep)	[11]
	ConvNet (shallow)	[11]
	RMDM	[8]
	TGSP + SVM	[17]
	TGSP + LDA	[8]
P300	EEGNet	[12]
	ConvNet (deep)	[11]
	ConvNet (shallow)	[11]
	XCov + RMDM	[8, 14, 15]
	TGSP + SVM	[23]
	TGSP + LR	[18]
SSVEP	EEGNet	[12]
	ConvNet (deep)	[11]
	ConvNet (shallow)	[11]
	RMDM	[8, 16]
	TGSP + SVM	[17]
	TGSP + LR	[17]

2.2.1. Within-session evaluation

This method evaluates a decoding pipeline by estimating classification scores for each session of each subject separately. A single session is split into training/test trials, and stratified 5-fold cross-validation is used. Stratified k-fold cross-validation preserves the class ratios for training and test sets.

2.2.2. Cross-session evaluation

When multiple sessions were recorded for a single subject (e.g. on different days), this method is used to evaluate a pipeline. Cross-session evaluation uses leave-one-out cross-validation, where the training set is created with trials of all the sessions of a single subject, except one session, which is used for the testing. This is done for all available sessions within a subject.

2.2.3. Cross-subject evaluation

This method is used to evaluate a pipeline when data from multiple subjects is available within the same dataset. Similarly to cross-session, the cross-subject evaluation uses leave-one-out cross-validation. The data is split into a training set containing trials of all available subjects except one, and the pipeline is evaluated on the trials of the left-out subject. This training and evaluation method can be considered as transfer learning.

3. Results

Overall, 13 different datasets with EEG trials from 80 different subjects were included for the statistical analysis. For each paradigm, an equal amount of subjects was picked from each dataset to avoid biasing the results towards one dataset. This subset was created incrementally, picking subjects by their numerical subject identifier starting with 1. Information about the included datasets and subjects is shown in table 2. While for several MI and P300 subjects multiple sessions were available, no multi-session experiments were recorded for any of the included SSVEP datasets.

Table 2. Included data sets for motor imagery are BNCI2014001 [24], BNCI2014004 [25], PhysionetMI [26], Shin2017A [27], MunichMI [28], for P300 are BNCI2014008 [29], BNCI2014009 [30], BNCI2015003 [31], BI2014A [32], BI2015A [33], and for SSVEP are SSVEPExo [34], Nakanishi2015 [35], MAMEM3 [36].

Paradigm	Datasets	Subjects	Classes	Electrodes
Motor imagery	5	25	2	3–128
P300	5	25	2	8–32
SSVEP	3	30	4–12	6–14
Total	13	80		

3.1. CNN architectures

To explore the impact of training iterations on the performance of the CNNs, the models were compared on their test scores for 100, 200 and 300 epochs. This comparison covered all BCI paradigms and evaluation methods, offering a clear sense of the models' overall performances. As shown in figure 1, the results suggest that while the EEGNet is the best performing model, none of the models show significant

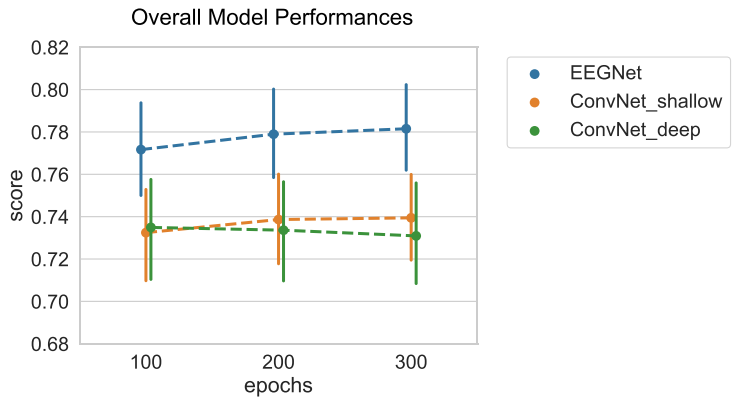

Figure 1. Overall model performance of the EEGNet, shallow ConvNet and deep ConvNet after 100, 200 and 300 training iterations. Each data point represents the mean of the ROC and accuracy scores over all three paradigms. The error bars show the bootstrapped 95% confidence interval.

improvements in test scores with increasing training iterations after 200 epochs.

Given that the EEGNet shows stable, high-performance results across paradigms, evaluations, and all different experimental setups, this architecture, trained on 300 epochs, is a robust and reliable choice for the comparative analysis with the Riemannian approaches.

3.2. CNN vs riemannian methods

All classification scores from all different pipelines over all included recordings were created with MOABB. After that, an analysis of the pipelines of each paradigm (MI, P300, SSVEP) and each evaluation method (within-session, cross-session, cross-subject) was done. To compare the pipelines, the two-sided 95% bootstrap confidence intervals (10 000 permutations) and the p-value for the permutation tests (10 000 permutations) were computed.

To find the approach that performs best over all paradigms, evaluations and possible experimental setups, the individual results were aggregated and afterwards the EEGNet was compared to the deep & shallow ConvNets, the RMDM approach as well as two TGSP mapping approaches, each with different classifiers based on the paradigm as shown in table 1. In the overall analysis, as visualized in figure 2, the EEGNet performs significantly better than the ConvNets. And while it also statistically outperforms the RMDM approach, the classification results of the Riemmanian TGSP mapping approaches are on par with the EEGNet. So, essentially, the CNNs and the Riemannian approach show comparable performance in the overall setting.

Further, the performance of the different pipelines for all specific paradigms and evaluation methods was analyzed. For each paradigm and each evaluation method, the mean scores of the EEGNet were compared to the mean scores of the ConvNets and each of the individual Riemannian pipelines. The results of the comparison between the EEGNet, the ConvNets and the Riemannian pipelines can be seen in table 3. For visualizations (Gardner–Altman estimation plots) of all results, see appendix.

Out of the 30 comparisons shown in table 3, we only found statistically significant differences (at significance level $\alpha = 0.05$, uncorrected) in four cases. The first three of these four cases concern the comparison of EEGNet with the deep and shallow ConvNet classifiers in the MI and P300 settings. Interestingly, we found EEGNet outperforming RMDM pipeline when performing cross-subject decoding in the SSVEP setting. Because we do not find significant differences in this case when comparing EEGNet with the TGSP pipelines, we do not consider these as evidence for superior performance of EEGNet in this setting in general.

To further investigate possible differences between Riemannian methods and the EEGNet, we performed a comparative analysis of each individual classification score, as shown in figure 3. For each classification score, the respectively best-performing Riemannian TGSP mapping pipeline was chosen and directly compared to the performance of the EEGNet. For the comparisons of all pipeline pairs, see appendix.

For the P300 and SSVEP paradigms, both the EEGNet and the Riemannian approaches show consistent, similar performances and low variability for the majority of sessions and subjects. Contrary to this, the MI paradigm shows clear differences in classification performances with higher variability. While there is no discernible pattern when highlighting the different evaluation methods as in figure 3, the cause of this spread can be seen in figure 4.

Highlighting the classification performances on the different datasets, the results show that for MI there are datasets, most apparent being Shin2017A

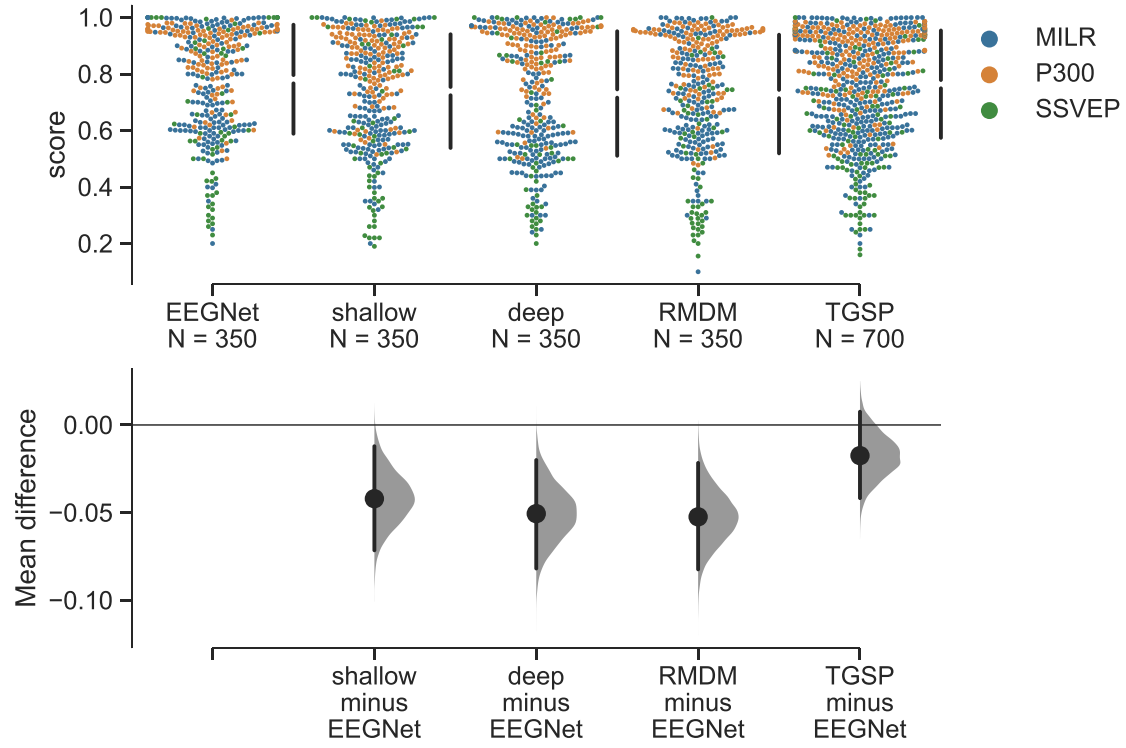

Figure 2. Overall model performance of the EEGNet compared to the shallow & deep ConvNets, the Riemannian minimum distance to mean (RMDM) and tangent space mapping (TGSP) approaches. Each point represents an individual classification score. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines. A negative mean difference indicates the respective pipeline performing worse than the EEGNet.

[27], where the different classification methods do not perform equally well on single subjects/sessions. This deviation indicates that for certain subjects and experimental settings, either of the approaches might be missing features that are relevant for accurate classification.

4. Discussion and conclusions

We found only minor differences in decoding performance between CNNs and Riemannian methods. The only exception to this finding is the cross-subject SSVEP setting, in which all CNNs outperformed the RMDM approach by a large margin. However, the TGSP mapping approach also showed similar performance to the CNNs in this specific setting. Notably, the results of the CNN pipelines show a more pronounced bimodal distribution compared to the Riemannian pipelines. The observation that deep convolutional networks may be particularly useful in transfer learning settings is in line with the results of a recent BCI decoding competition that was also won by a CNN approach [37]. However, we remark that the methods benchmarked here were not explicitly designed for cross-subject decoding. In particular, extensions of Riemannian pipelines for transfer learning have been developed but have not been incorporated [38]. Additionally, a novel type of Riemannian networks based on geometric deep learning [39] has recently found adoption for decoding EEG data. These deep Riemannian networks (DRNs) have shown state-of-the-art performances [40] and improvements to transfer learning and domain adaptation [41] for MI. While we do consider DRNs a promising approach, they have been left out of this analysis due to them not being developed for or validated on P300 and SSVEP datasets. In any case, the results show comparable performance of the two conceptually very different decoding approaches we analyzed. We hypothesize that future significant breakthroughs in new decoding pipelines are less likely to be found for P300 and SSVEP, with more room for advancements in the MI paradigm. For MI, there is the need to find approaches that extract the optimal information on any given experimental setting and subject. Future progress for BCIs may also depend on novel paradigms and techniques for recording brain signals. However, classification score is not the only relevant metric for the practical utility of BCI decoding pipelines. Depending on the particular setting, the computational complexity of training and online decoding may also have to be considered,

Table 3. Results of the statistical analysis.The *mean difference* column shows the average difference in classification, given *n* evaluation scores for each of the pipeline pairs. If the mean difference is negative, the EEGNet performs better. The lower and upper bounds of the 95% confidence interval are given and the rightmost column shows the statistical significance ($p < 0.05$).

Paradigm	Evaluation	Pipelines (EEGNet vs X)	n	mean difference (X−EEGNet)	95% CI	p-value	p<0.05
MILR	WithinSession	shallow	60	−0.0225	−0.0969, 0.0529	0.5490	–
MILR	WithinSession	deep	60	−0.0781	−0.155, 0.0002	0.0500	–
MILR	WithinSession	RMDM	60	−0.0723	−0.1457, 0.0007	0.0598	–
MILR	WithinSession	TGSP+SVM	60	0.0099	−0.0635, 0.0827	0.7922	–
MILR	WithinSession	TGSP+LDA	60	−0.0550	−0.1294, 0.0204	0.1550	–
MILR	CrossSession	shallow	50	−0.0577	−0.1276, 0.0139	0.1166	–
MILR	CrossSession	deep	50	−0.1051	−0.1789, −0.0272	0.0100	✓
MILR	CrossSession	RMDM	50	−0.0516	−0.1205, 0.0181	0.1532	–
MILR	CrossSession	TGSP+SVM	50	−0.0044	−0.0705, 0.0599	0.8916	–
MILR	CrossSession	TGSP+LDA	50	−0.0469	−0.1147, 0.0198	0.1920	–
MILR	CrossSubject	shallow	60	−0.0431	−0.1026, 0.0142	0.1492	–
MILR	CrossSubject	deep	60	−0.0805	−0.1442, −0.0187	0.0130	✓
MILR	CrossSubject	RMDM	60	−0.0219	−0.0782, 0.0354	0.4558	–
MILR	CrossSubject	TGSP+SVM	60	0.0158	−0.0419, 0.0701	0.5838	–
MILR	CrossSubject	TGSP+LDA	60	−0.0357	−0.0932, 0.021	0.2256	–
P300	WithinSession	shallow	45	−0.0657	−0.0994, −0.0326	0.0004	✓
P300	WithinSession	deep	45	0.0019	−0.0264, 0.0297	0.9034	–
P300	WithinSession	XCov+RMDM	45	−0.0218	−0.0565, 0.0115	0.2206	–
P300	WithinSession	TGSP+SVM	45	−0.0206	−0.0554, 0.0109	0.2364	–
P300	WithinSession	TGSP+LR	45	−0.0111	−0.0418, 0.0193	0.4836	–
P300	CrossSession	shallow	30	−0.0113	−0.0353, 0.0151	0.4102	–
P300	CrossSession	deep	30	0.0058	−0.0191, 0.0311	0.6792	–
P300	CrossSession	XCov+RMDM	30	−0.0119	−0.0403, 0.0156	0.4254	–
P300	CrossSession	TGSP+SVM	30	−0.0115	−0.0409, 0.0155	0.4490	–
P300	CrossSession	TGSP+LR	30	−0.0092	−0.0376, 0.0178	0.5350	–
P300	CrossSubject	shallow	45	−0.0255	−0.0775, 0.0264	0.3450	–
P300	CrossSubject	deep	45	−0.0023	−0.0557, 0.0507	0.9384	–
P300	CrossSubject	XCov+RMDM	45	−0.0030	−0.0584, 0.0494	0.9120	–
P300	CrossSubject	TGSP+SVM	45	0.0014	−0.0523, 0.0542	0.9660	–
P300	CrossSubject	TGSP+LR	45	0.0122	−0.0392, 0.0618	0.6466	–
SSVEP	WithinSession	shallow	30	−0.0625	−0.1945, 0.0719	0.3540	–
SSVEP	WithinSession	deep	30	−0.0914	−0.2222, 0.0465	0.1828	–
SSVEP	WithinSession	RMDM	30	−0.0869	−0.2146, 0.0388	0.1908	–
SSVEP	WithinSession	TGSP+LR	30	−0.0034	−0.1284, 0.122	0.9570	–
SSVEP	WithinSession	TGSP+SVM	30	−0.0217	−0.1456, 0.1017	0.7234	–
SSVEP	CrossSubject	shallow	30	−0.0530	−0.1799, 0.077	0.4148	–
SSVEP	CrossSubject	deep	30	−0.0119	−0.1371, 0.1163	0.8420	–
SSVEP	CrossSubject	RMDM	30	−0.2011	−0.3086, −0.094	0.0014	✓
SSVEP	CrossSubject	TGSP+LR	30	−0.0439	−0.1604, 0.0711	0.4514	–
SSVEP	CrossSubject	TGSP+SVM	30	−0.0757	−0.1898, 0.0386	0.1968	–

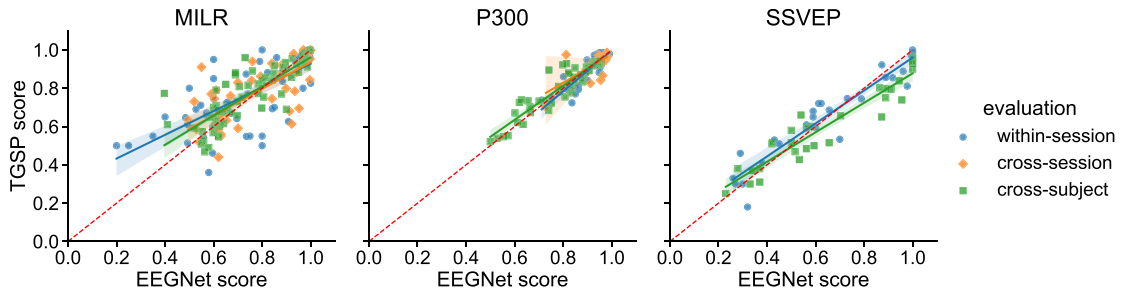

Figure 3. Comparative analysis of classification performances between the EEGNet and the respectively best-performing Riemannian TGSP method. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

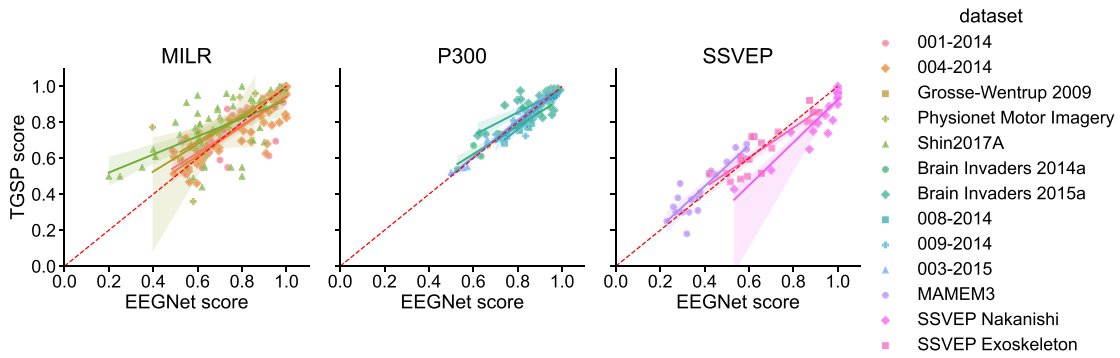

Figure 4. Comparative analysis of classification performances between the EEGNet and the respectively best-performing Riemannian TGSP method. Each point represents an individual classification score. The distinct colors indicate the dataset.

with potential trade-offs between decoding performance and computational complexity. While we did not conduct a systematic runtime comparison, we found that training the EEGNet required substantially higher computational resources than the Riemannian methods.

We further note that our analyses are based on publicly available datasets, which may incorporate sampling biases, e.g. because subjects have been recruited in academic settings. For instance, the number of BCI-illiterate subjects, which have not been excluded in our analysis, may vary across settings [42]. Our results may thus not generalize to future users of BCI in non-laboratory settings, where any of the tested pipelines may exhibit more or less robust performance. We consider it a relevant topic of future research to explore whether certain pipelines are more suitable for particular sub-groups of subjects. Such an analysis probably requires a substantially higher number of datasets and subjects than currently available in the MOABB framework. A broad adoption of passive BCI technologies may render such analyses feasible [43].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/ederm42/BCI-riemannian-vs-eegnet.

Conflit of interests

The authors declare that they have no competing interests.

Authors' contributions

ME designed this research, implemented the computational framework and analysed the data. JX provided scientific guidance and helped supervise the project. MGW conceived the original idea, verified the analytical methods and supervised the project. The authors read and approved the final manuscript.

Appendix

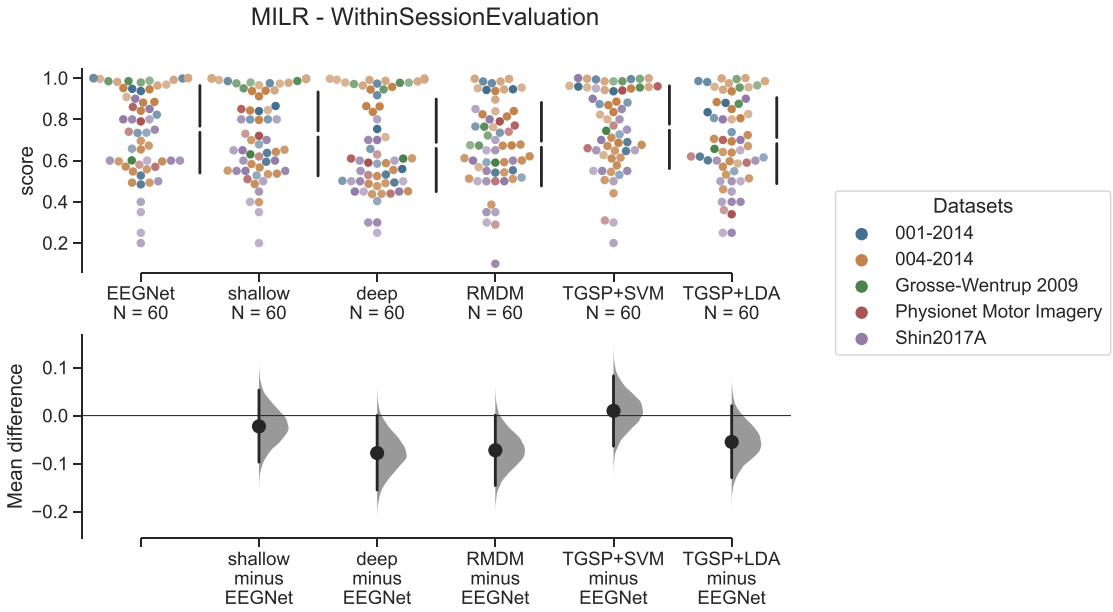

Figure A1. Gardner–Altman estimation plot for Motor Imagery within-session evaluation. Each point represents an individual classification score. The shadings represent individual sessions within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

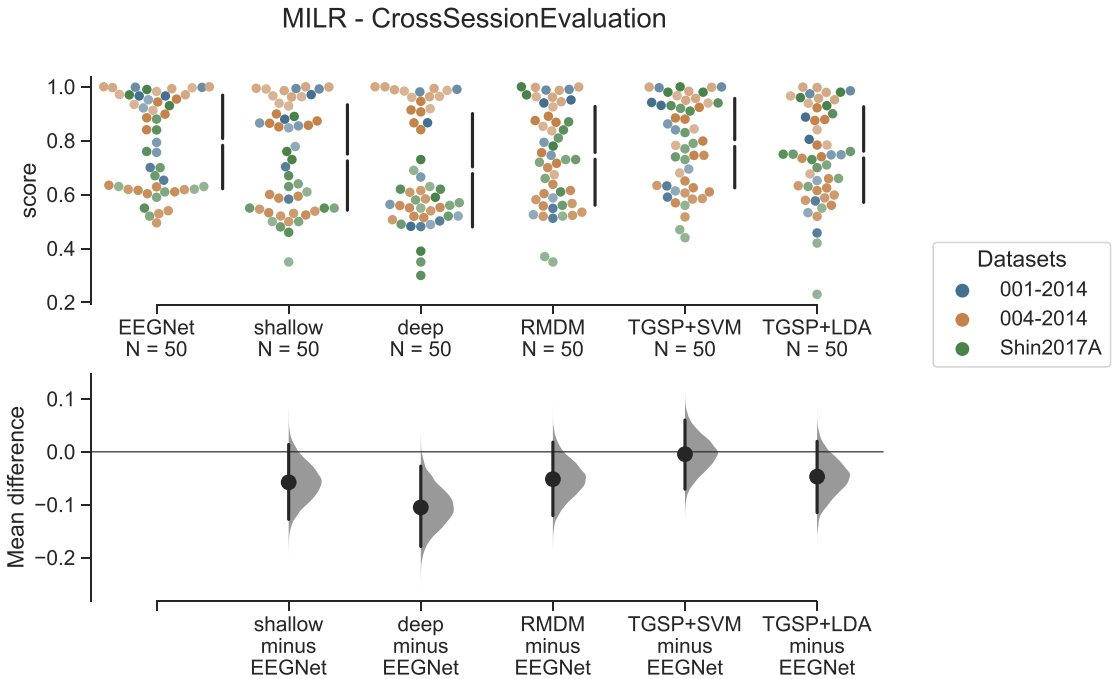

Figure A2. Gardner–Altman estimation plot for Motor Imagery cross-session evaluation. Each point represents an individual classification score. The shadings represent individual sessions within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

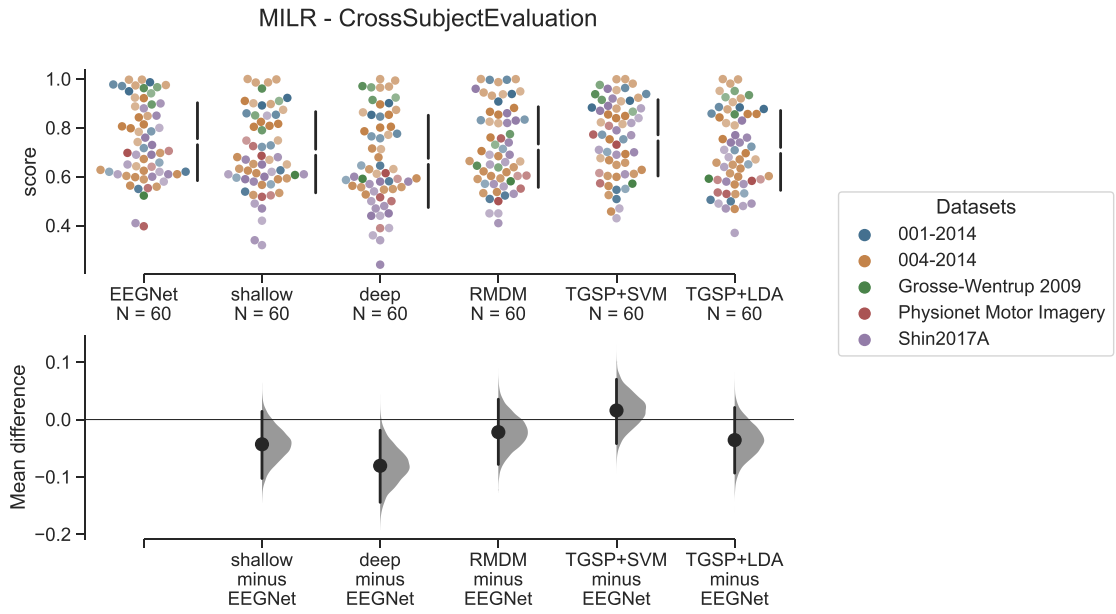

Figure A3. Gardner–Altman estimation plot for Motor Imagery cross-subject evaluation. Each point represents an individual classification score. The shadings represent individual subjects within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

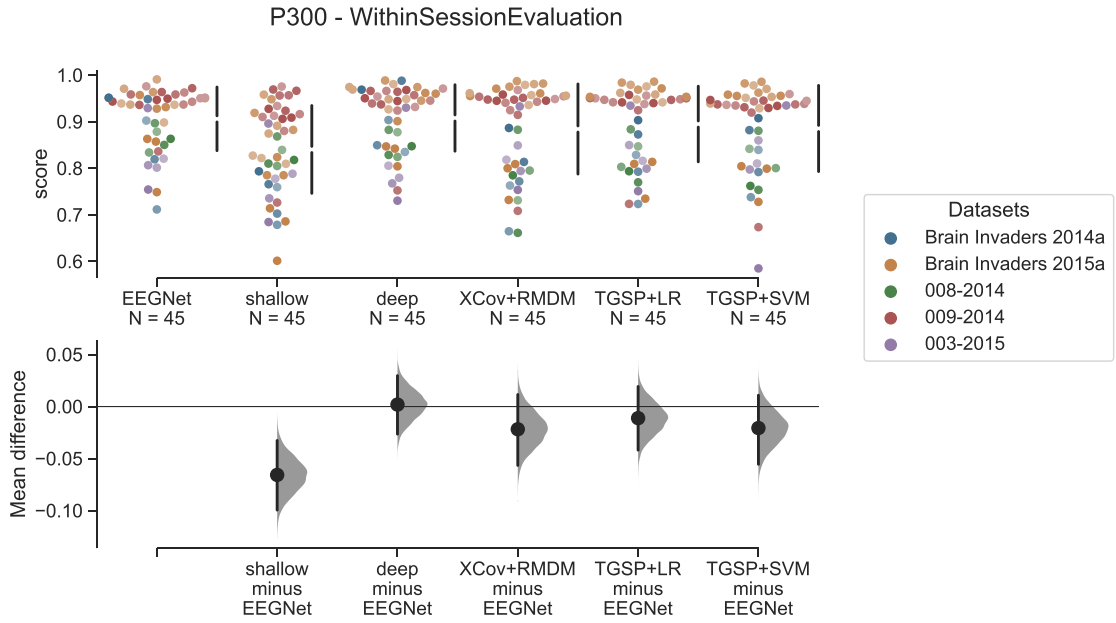

Figure A4. Gardner–Altman estimation plot for P300 within-session evaluation. Each point represents an individual classification score. The shadings represent individual sessions within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

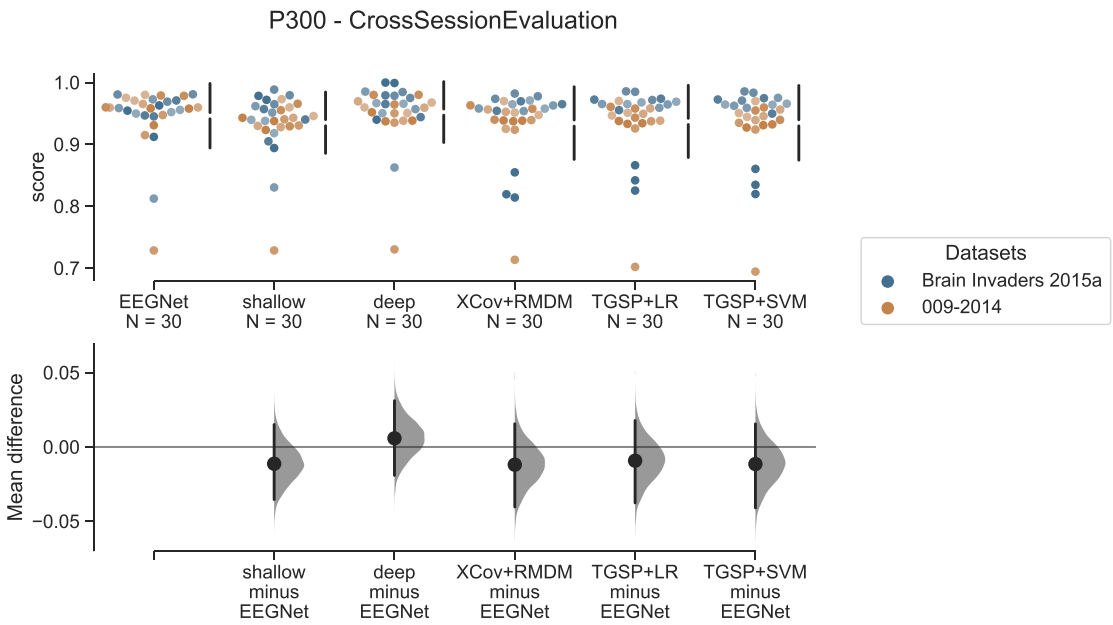

Figure A5. Gardner–Altman estimation plot for P300 cross-session evaluation. Each point represents an individual classification score. The shadings represent individual sessions within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

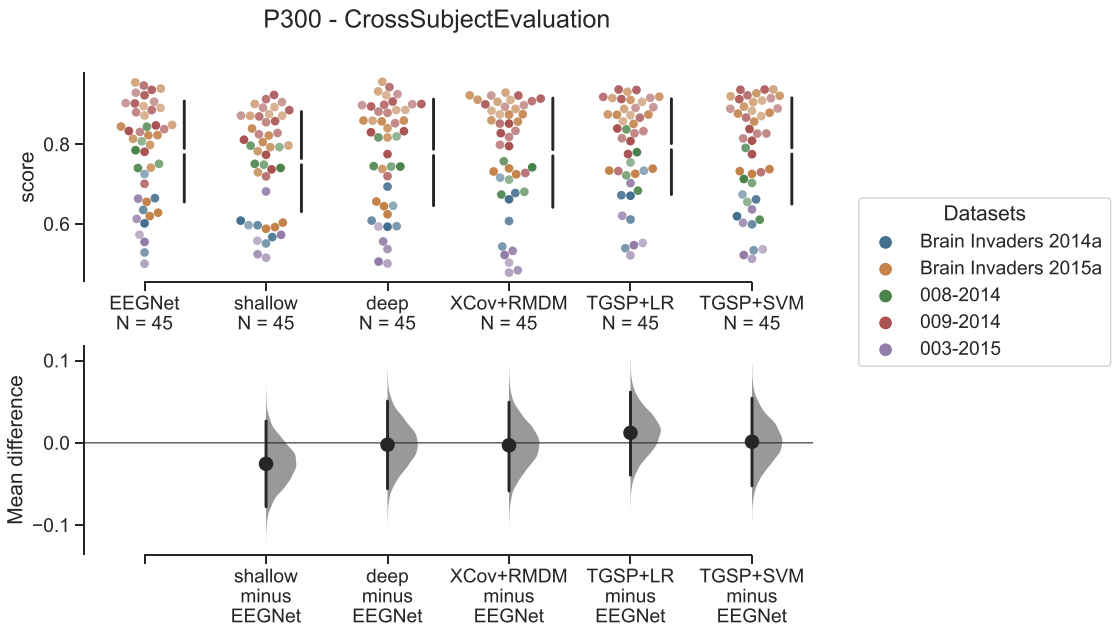

Figure A6. Gardner–Altman estimation plot for P300 cross-subject evaluation. Each point represents an individual classification score. The shadings represent individual subjects within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

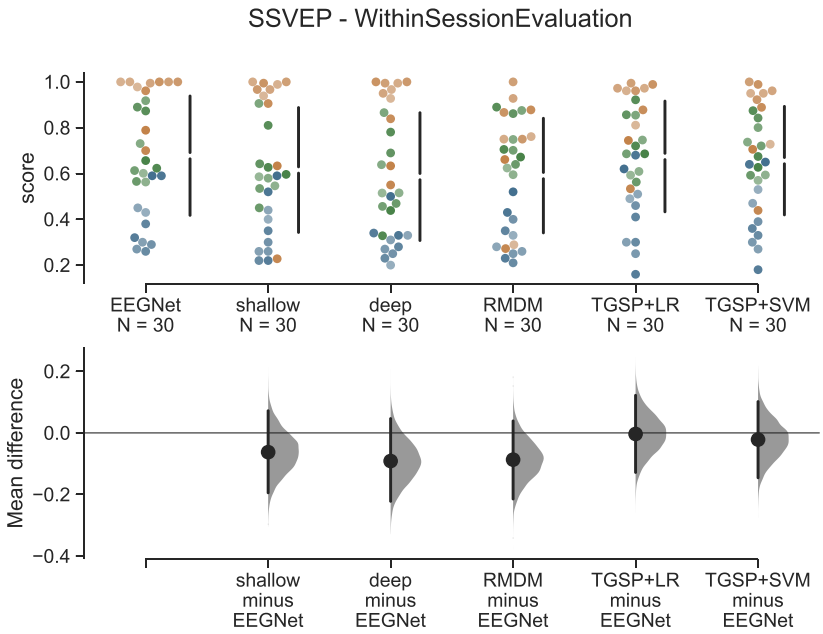

Figure A7. Gardner–Altman estimation plot for SSVEP within-session evaluation. Each point represents an individual classification score. The shadings represent individual sessions within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

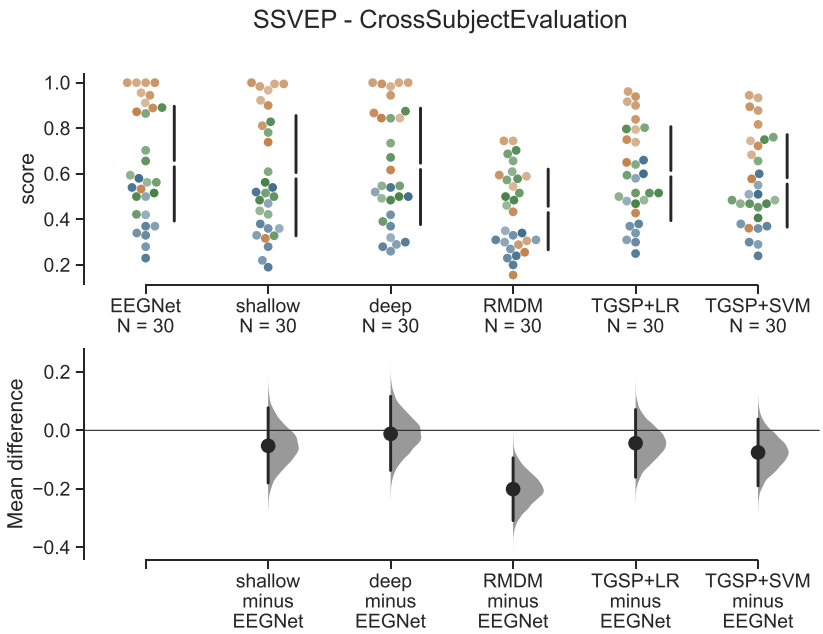

Figure A8. Gardner–Altman estimation plot for SSVEP cross-subject evaluation. Each point represents an individual classification score. The shadings represent individual subjects within a dataset. The lower axis shows the bootstrapped 95% confidence interval of the mean difference between the EEGNet and the other pipelines.

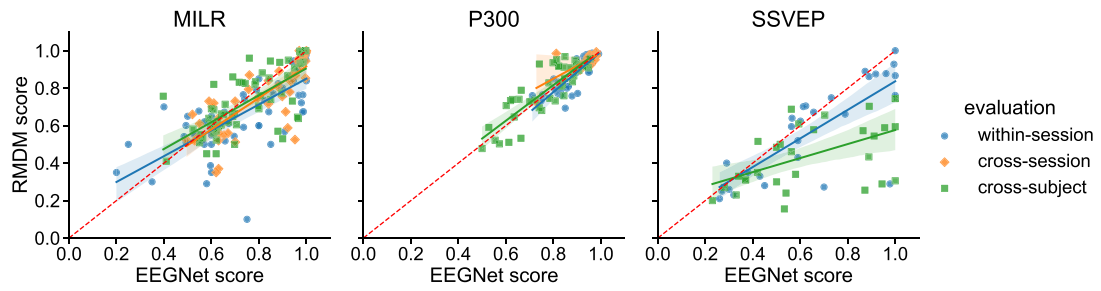

Figure A9. Comparative analysis of classification performances between the EEGNet and the ShallowConvNet. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

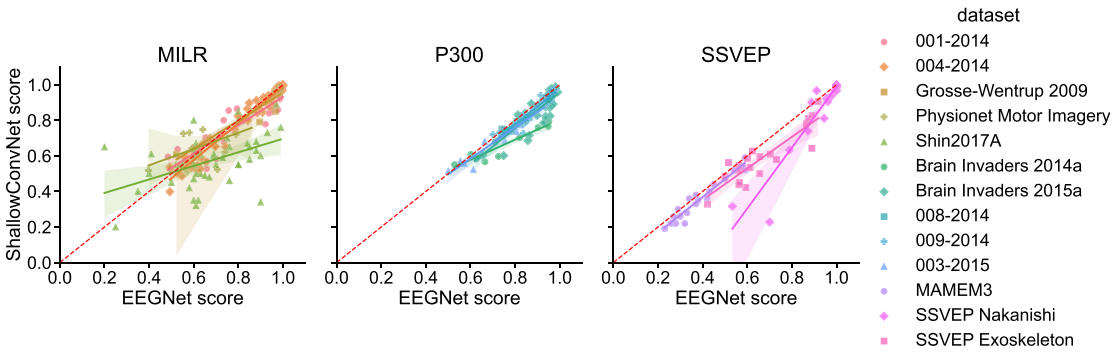

Figure A10. Comparative analysis of classification performances between the EEGNet and the ShallowConvNet. Each point represents an individual classification score. The distinct colors indicate the dataset.

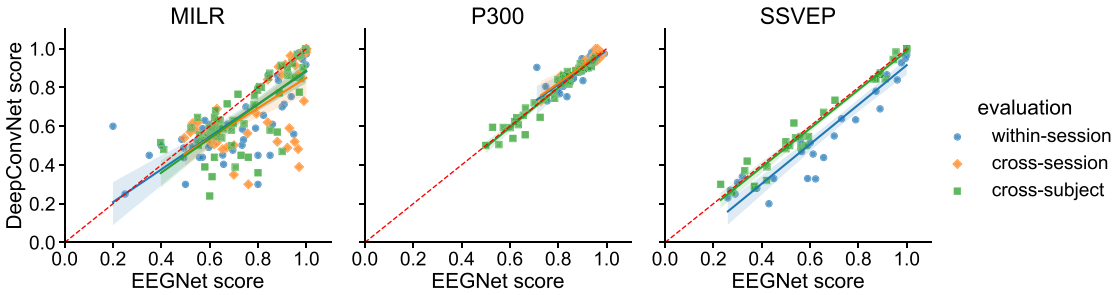

Figure A11. Comparative analysis of classification performances between the EEGNet and the DeepConvNet. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

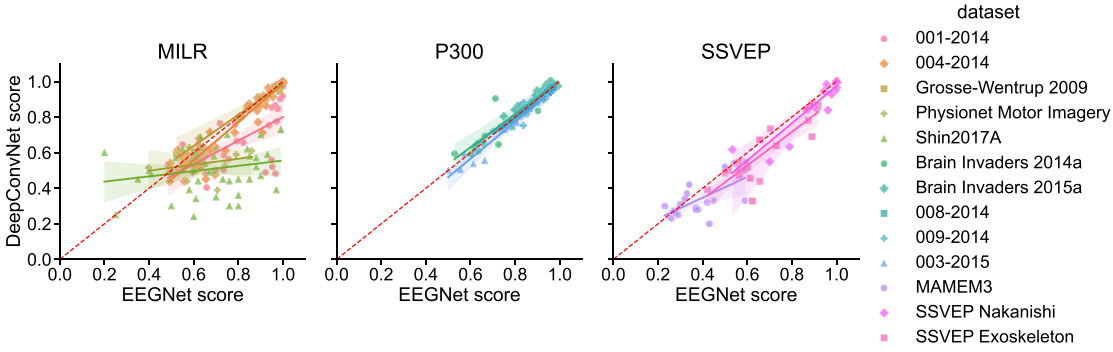

Figure A12. Comparative analysis of classification performances between the EEGNet and the DeepConvNet. Each point represents an individual classification score. The distinct colors indicate the dataset.

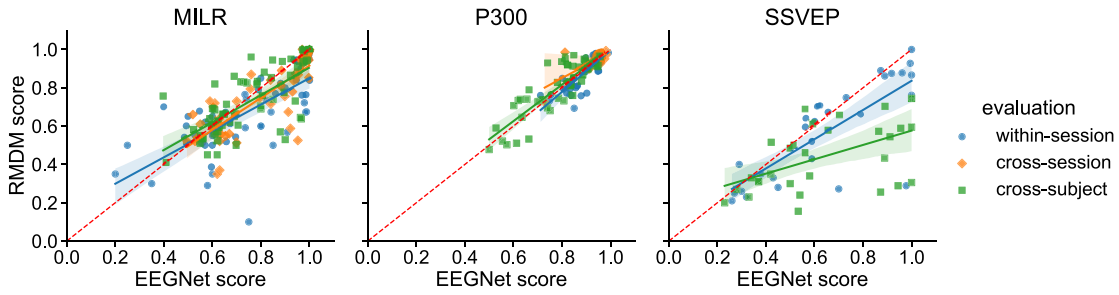

Figure A13. Comparative analysis of classification performances between the EEGNet and the Riemannian Minimum Distance to Mean (RMDM) approach. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

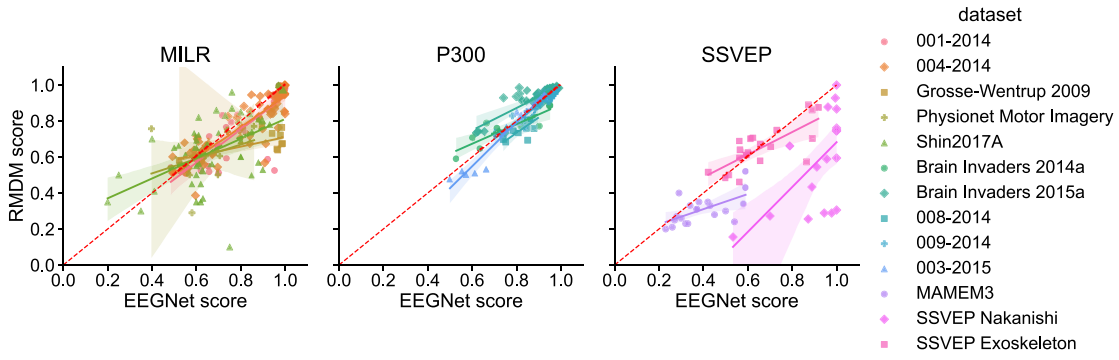

Figure A14. Comparative analysis of classification performances between the EEGNet and the Riemannian Minimum Distance to Mean (RMDM) approach. Each point represents an individual classification score. The distinct colors indicate the dataset.

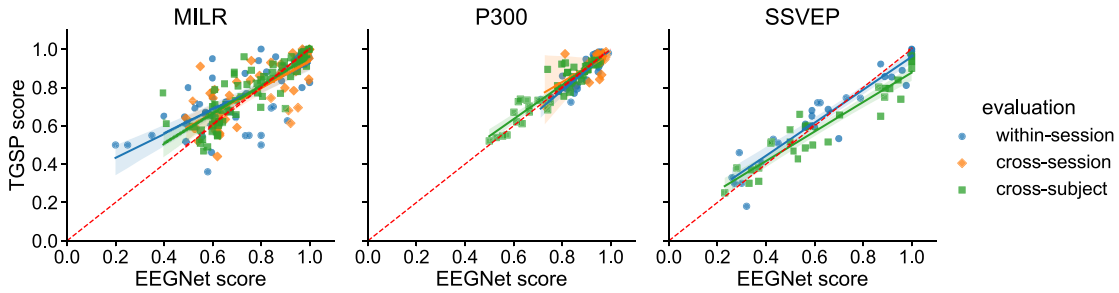

Figure A15. Comparative analysis of classification performances between the EEGNet and the respectively best-performing Riemannian tangent space mapping (TGSP) approach. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

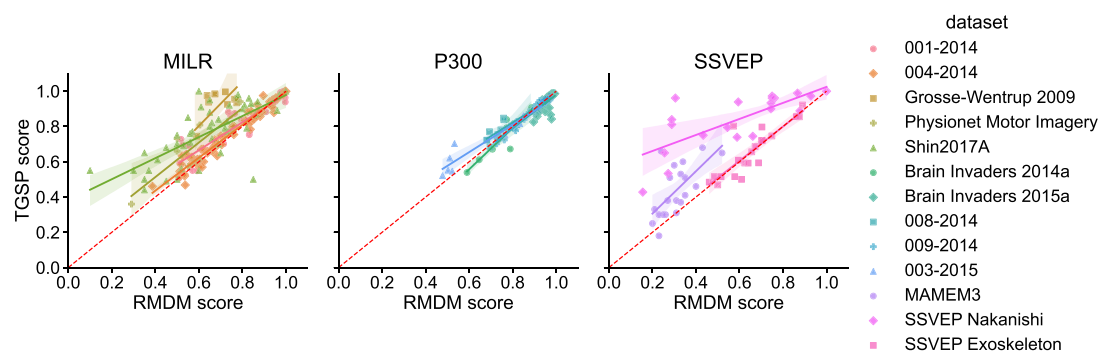

Figure A16. Comparative analysis of classification performances between the EEGNet and the respectively best-performing Riemannian tangent space mapping (TGSP) approach. Each point represents an individual classification score. The distinct colors indicate the dataset.

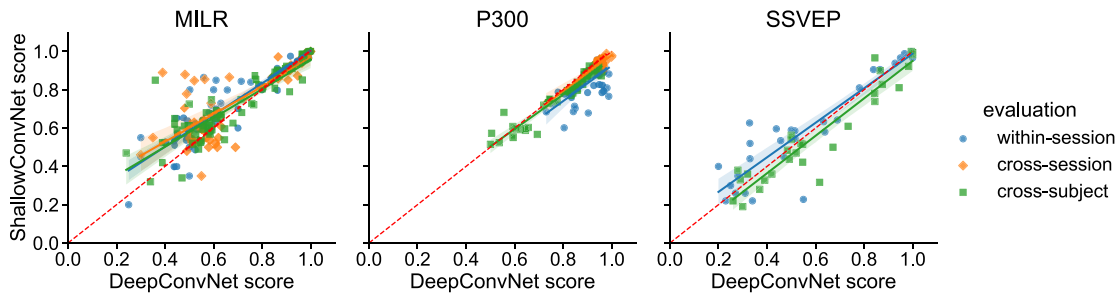

Figure A17. Comparative analysis of classification performances between the DeepConvNet and the ShallowConvNet. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

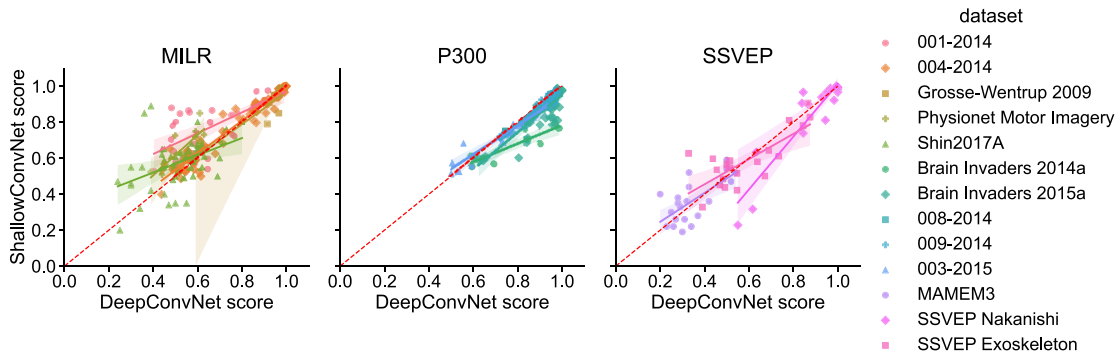

Figure A18. Comparative analysis of classification performances between the DeepConvNet and the ShallowConvNet. Each point represents an individual classification score. The distinct colors indicate the dataset.

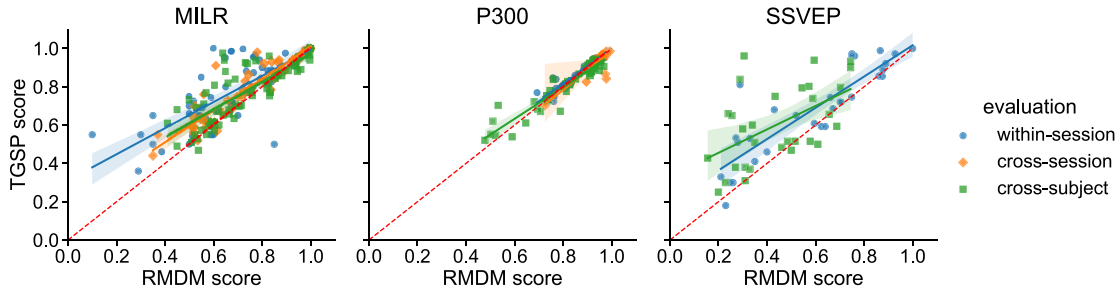

Figure A19. Comparative analysis of classification performances between the Riemannian Minimum Distance to Mean (RMDM) and the respectively best-performing Riemannian TGSP approach. Each point represents an individual classification score. The distinct colors indicate the evaluation method.

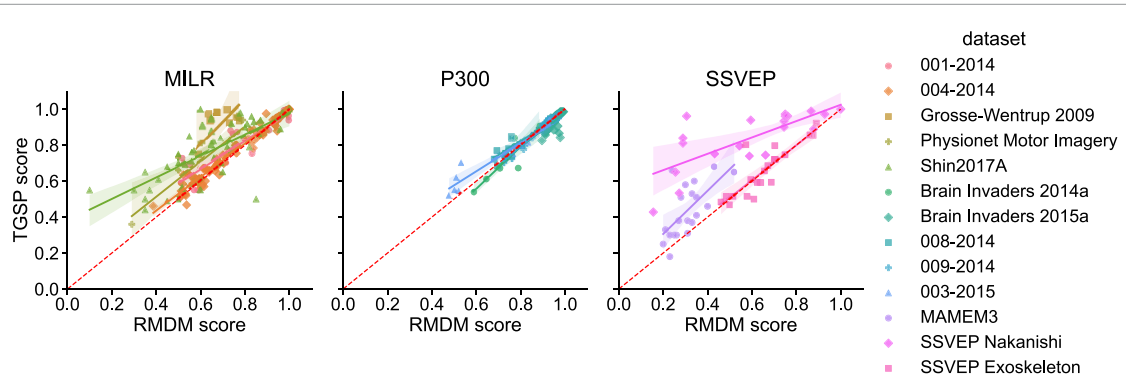

Figure A20. Comparative analysis of classification performances between the Riemannian Minimum Distance to Mean (RMDM) and the respectively best-performing Riemannian TGSP approach. Each point represents an individual classification score. The distinct colors indicate the dataset.

ORCID iDs

Manuel Eder https://orcid.org/0009-0005-2799-082X
Jiachen Xu https://orcid.org/0000-0002-9985-7447
Moritz Grosse-Wentrup https://orcid.org/0000-0001-9787-2291

References

[1] Vidal J J 1977 Real-time detection of brain events in EEG *Proc. IEEE* **65** 633–41
[2] Bozinovski S, Sestakov M and Bozinovska L 1988 Using EEG alpha rhythm to control a mobile robot *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* vol 3 pp 1515–6
[3] Hochberg L R, Serruya M D, Friehs G M, Mukand J A, Saleh M, Caplan A H, Branner A, Chen D, Penn R D and Donoghue J P 2006 Neuronal ensemble control of prosthetic devices by a human with tetraplegia *Nature* **442** 164–71
[4] Gomez-Rodriguez M, Peters J, Hill J, Schölkopf B, Gharabaghi A and Grosse-Wentrup M 2011 Closing the sensorimotor loop: haptic feedback facilitates decoding of motor imagery *J. Neural Eng.* **8** 036005
[5] Anumanchipalli G K, Chartier J and Chang E F 2019 Speech synthesis from neural decoding of spoken sentences *Nature* **568** 493–8
[6] Abiri R, Borhani S, Sellers E W, Jiang Y and Zhao X 2019 A comprehensive review of EEG-based brain–computer interface paradigms *J. Neural Eng.* **16** 011001
[7] Pfurtscheller G and Neuper C 2001 Motor imagery and direct brain-computer communication *Proc. IEEE* **89** 1123–34
[8] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A and Yger F 2018 A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update *J. Neural Eng.* **15** 031005
[9] Jayaram V and Barachant A 2018 MOABB: trustworthy algorithm benchmarking for BCIs *J. Neural Eng.* **15** 066011
[10] Barachant A, Bonnet S, Congedo M and Jutten C 2010 Riemannian geometry applied to bci classification *Int. conf. on latent variable analysis and signal separation* (https://doi.org/10.1007/978-3-642-15995-4_78)
[11] Schirrmeister R T, Springenberg J T, Fiederer L D F, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for eeg decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[12] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2016 EEGNet: a compact convolutional network for EEG-based brain-computer interfaces *J. Neural Eng.* **15** 5
[13] Huggins J E *et al* 2022 Workshops of the eighth international brain–computer interface meeting: bcis: the next frontier *Brain-Computer Interfaces* **9** 69–101
[14] Congedo M, Barachant A and Andreev A 2013 A new generation of brain-computer interface based on riemannian geometry (arXiv:1310.8115)
[15] Barachant A 2014 *Meg Decoding Using Riemannian Geometry and Unsupervised Classification* (Grenoble University)
[16] Chevallier S, Kalunga E, Barthélemy Q and Yger F 2018 Riemannian classification for SSVEP-based BCI: Offline versus online implementations *Brain–Computer Interfaces Handbook* (CRC Press) pp 371–96
[17] Barachant A and Chevallier S 2022 MOABB: tutorials (available at: https://github.com/NeuroTechX/moabb)
[18] Vernon J L 2023 Army Research Laboratory (ARL) EEGModels Project (available at:https://github.com/vlawhern/arl-eegmodels)
[19] pyRiemann Contributors 2022 pyRiemann: biosignals classification with Riemannian geometry (available at:https://github.com/ pyRiemann/pyRiemann)
[20] Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
[21] Joses H, Tumkaya T, Aryal S, Choi H and Claridge-Chang A 2019 Moving beyond p values: data analysis with estimation graphics *Nat. Methods* **16** 565–6
[22] Joses H, Tumkaya T, Aryal S, Choi H and Claridge-Chang A 2022 DABEST: data analysis with bootstrap-coupled ESTimation (available at:https://github.com/ACCLAB/DABEST-python)
[23] Barachant A, Bonnet S, Congedo M and Jutten C 2013 Classification of covariance matrices using a riemannian-based kernel for bci applications *Neurocomputing* **112** 172–8
[24] Tangermann M *et al* 2012 Review of the BCI Competition IV *Front. Neurosci.* **6** 55
[25] Leeb R, Lee F, Keinrath C, Scherer R, Bischof H and Pfurtscheller G 2007 Brain–Computer Communication: motivation, aim and impact of exploring a virtual apartment *IEEE Trans. Neural Syst. Rehabil. Eng.* **15** 473–82
[26] Schalk G, McFarland D J, Hinterberger T, Birbaumer N and Wolpaw J R 2004 BCI2000: a general-purpose brain-computer interface (BCI) system *IEEE Trans. Biomed. Eng.* **51** 1034–43
[27] Shin J, von Lühmann A, Blankertz B, Kim D-W, Jeong J, Hwang H-J and Müller K-R 2017 Open access dataset for EEG+NIRS single-trial classification *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 1735–45
[28] Grosse-Wentrup M, Liefhold C, Gramann K and Buss M 2009 Beamforming in noninvasive brain–computer interfaces *IEEE Trans. Biomed. Eng.* **56** 1209–19
[29] Riccio A, Simione L, Schettini F, Pizzimenti A, Inghilleri M, Olivetti Belardinelli M, Mattia D and Cincotti F 2013 Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis *Front. Hum. Neurosci.* **7** 732
[30] Aricó P, Aloise F, Schettini F, Salinari S, Mattia D and Cincotti F 2014 Influence of P300 latency jitter on event related potential-based brain–computer interface performance *J. Neural Eng.* **11** 035008
[31] Guger C, Daban S, Sellers E, Holzner C, Krausz G, Carabalona R, Gramatica F and Edlinger G 2009 How many people are able to control a P300-based brain–computer interface (BCI)? *Neurosci. Lett.* **462** 94–98
[32] Korczowski L, Ostaschenko E, Andreev A, Cattan G, Luiz Coelho Rodrigues P, Gautheret V and Congedo M 2019 Brain invaders calibration-less P300-based BCI using dry EEG electrodes dataset (bi2014a) *Research Report* (GIPSA-lab) (https://doi.org/10.5281/zenodo.3266223)
[33] Korczowski L, Cederhout M, Andreev A, Cattan G, Luiz Coelho Rodrigues P, Gautheret V and Congedo M 2019 Brain invaders calibration-less P300-based BCI with modulation of flash duration dataset (bi2015a) *Research Report* (GIPSA-lab) (https://doi.org/10.5281/zenodo.3266930)
[34] Kalunga E K, Chevallier S, Barthélemy Q, Djouani K, Monacelli E and Hamam Y 2016 Online SSVEP-based BCI using Riemannian geometry *Neurocomputing* **191** 55–68
[35] Nakanishi M, Wang Y, Wang Y-T and Jung T-P 2015 A Comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials *PLoS One* **10** e0140703
[36] Oikonomou V P, Liaros G, Georgiadis K, Chatzilari E, Adam K, Nikolopoulos S and Kompatsiaris I 2016 Comparative evaluation of state-of-the-art algorithms for SSVEP-based BCIs (arXiv: 1602.00904)
[37] Wei X *et al* 2022 2021 BEETL Competition: advancing transfer learning for subject independence and heterogenous EEG data sets *Proc. NeurIPS 2021 Competitions and Demonstrations Track* (PMLR) pp 205–19
[38] Zanini P, Congedo M, Jutten C, Said S and Berthoumieu Y 2018 Transfer Learning: a riemannian geometry framework

with applications to brain–computer interfaces *IEEE Trans. Biomed. Eng.* **65** 1107–16
[39] Huang Z and Van Gool L 2017 A riemannian network for spd matrix learning *Association for the Advancement of Artificial Intelligence (AAAI)* (https://doi.org/10.48550/arXiv.1608.04233)
[40] Wilson D, Tibor Schirrmeister R, Alexander Wilhelm Gemein L and Ball T 2023 Deep riemannian networks for eeg decoding (https://doi.org/10.48550/arXiv.2212.10426)
[41] Kobler R, Hirayama J-I, Zhao Q and Kawanabe M 2022 Spd domain-specific batch normalization to crack interpretable unsupervised domain adaptation in eeg *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc) pp 6219–35
[42] Dickhaus T, Sannelli C, Müller K-R, Curio G and Blankertz B 2009 Predicting bci performance to study bci illiteracy *BMC Neuroscience* **10** 84
[43] Krol L R, Haselager P and Zander T O 2020 Cognitive and affective probing: a tutorial and review of active learning for neuroadaptive technology *J. Neural Eng.* **17** 012001