A Appendix (MAtt: A Manifold Attention Network for EEG Decoding)

A.1 Time complexity

The original time complexity of training the query, key, and value in Algorithm 1 is $O(p)$, where p is the number of time iterations. We simplify this procedure for better reducing the time complexity. If parallel processing is available in the computational environment, the time complexity poses significant influence on the efficiency of executing Algorithm 1. We first define the notations: Suppose $A \in \mathbb{R}^{m \times n_1}, B \in \mathbb{R}^{m \times n_2}$, then $Concat(A, B)$ $means$ $concatenating$ A and B. Given a sequence of SPD data $\{X_i\}_{i=1}^p$ as the input of the manifold attention module, for the query $Q = Concat(\{Q_i\}_{i=1}^p)$, key $K = Concat(\{K_i\}_{i=1}^p)$, and value $V = Concat(\{V_i\}_{i=1}^p)$, W_q, W_k, and W_v are parameters to determine Q, K, and V:

$$\begin{aligned} Q &= [Q_1 \quad Q_2 \quad \cdots \quad Q_p] \\ &= \begin{bmatrix} W_q X_1 W_q^T & W_q X_2 W_q^T & \cdots & W_q X_p W_q^T \end{bmatrix} \\ &= W_q Concat(X_1, \cdots, X_p) W_q^T \end{aligned}$$

same for K and V:

$$\begin{aligned} K &= W_k Concat(X_1, \cdots, X_p) W_k^T \\ V &= W_v Concat(X_1, \cdots, X_p) W_v^T \end{aligned}$$

Then we can reduce the time complexity of computing Q, K, and V, to $O(3)$, a constant complexity. Moreover, we have another perspective to reduce the time complexity from linear to a unit constant:

$$\begin{bmatrix} Q \\ K \\ V \end{bmatrix} = diag\left(\begin{bmatrix} W_q \\ W_k \\ W_v \end{bmatrix} Concat(X_1, \cdots, X_p) \begin{bmatrix} W_q^T & W_k^T & W_v^T \end{bmatrix} \right)$$

We use two cores in Intel(R) Xeon(R) W-2133 CPU to train the proposed model. Table 1 shows the average training time for the three datasets, BCIC-IV-2a, MAMEM-SSVEP-II, and BCI-ERN.

Table 1: A comparison of the mean training time (seconds) per iteration across models. The error denotes the standard deviation.

	BCIC-IV-2a	MAMEM-SSVEP-II	BCI-ERN
ShallowNet	0.58± 0.0503	0.11± 0.0165	2.20±0.3533
EEGNet	0.45± 0.0308	0.72± 0.0285	7.79±0.7059
SCCNet	0.06± 0.0070	0.20± 0.0274	0.43±0.3064
EEG-TCNet	0.36±0.0019	0.22±0.0136	0.33±0.0264
TCNet-Fusion	0.26± 0.0045	0.07±0.0012	0.20±0.0027
FBCNet	0.93± 0.0047	0.15± 0.0035	0.13±0.0017
MBEEGSE	0.72±0.0051	0.41±0.0012	2.24±0.0066
MAtt	0.96± 0.0843	2.26± 0.1598	0.52±0.0169

A.2 Affine invariant metric

The geodesic distance between two points P_1 and P_2 is defined by the infimum of length of all curves go through from P_1 to P_2 on the Riemannian manifold. Suppose a piecewise smooth curve $\gamma : [0, 1] \mapsto \mathbb{R}$ with $\gamma(0) = P_1$, $\gamma(1) = P_2$, the geodesic distance from P_1 to P_2 on $(\mathcal{M}, g)$ can be defined as:

$$\delta_g(P_1, P_2) = \inf\{Length(\gamma)\} = \inf\{\int_0^1 \|\gamma'(t)\|_g dt\}$$

Given a Riemannian metric (i.e. affine invariant metric) [1], we have *Riemannian geodesic* as follows:

$$\delta_R(P_1, P_2) = \|Log({P_1}^{-1} P_2)\|_F = \|Log({P_1}^{-1/2} P_2 {P_1}^{-1/2})\|_F = \left[\sum_{i=1}^{n} log^2 \lambda_i \right]^{\frac{1}{2}}$$

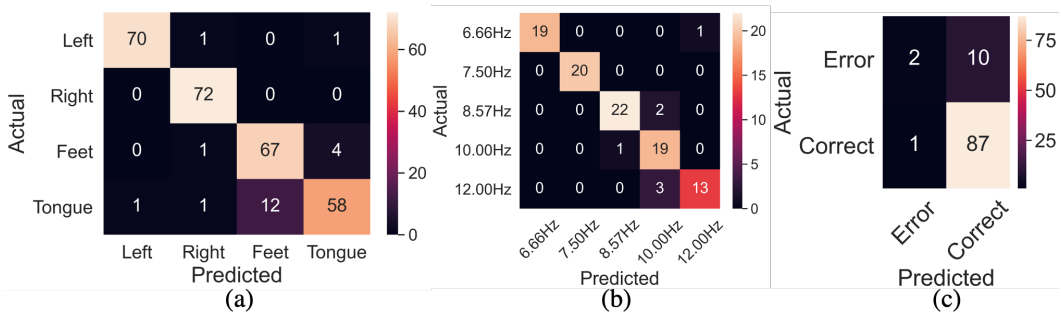


Figure 1: (a) Confusion matrix of S3 in the BCICIV2a dataset. 'Left' and 'Right' refer to 'Left hand' and 'Right hand' respectively. (b) Confusion matrix of S11 in the MAMEM dataset. (c) Confusion matrix of S7 in the BCI-ERN dataset.

The following is the definition of the Riemannian mean (denoted as $\mathcal{B}$). Suppose there are k SPD matrices on the SPD manifold, called $P_1, P_2, ..., P_k$:

$$\mathcal{B}(P_1, ...P_k) = \underset{P \in Sym^+(n)}{\arg\min} \sum_{l=1}^{k} \delta_R^2(P, P_l)$$

However, the solution to the above optimization problem doesn't have a closed-form solution. Thus, we should compute the final $\mathcal{B}$ in an iteration manner [2, 3] until conditions of convergence are satisfied. Due to the high computational complexity of computing Riemannian mean, we herein, alternatively, use the Log-Euclidean metric to measure the distance between two points on the manifold in our method.

A.3 Confusion matrices

Figure 1 depicts the confusion matrix of single-subject classification results on all three datasets. As shown in Figure 1 (a), the 'Left hand' and 'Right hand' are relatively classified correctly, while there are 12 MI EEG samples of 'Tongue' being misclassified as 'Feet'. According to the visualization of model interpretation, both 'Feet' and 'Tongue' are characterized by symmetric topographical distribution, which may cause the samples of these two classes to be misclassified. On the other hand, Figure 1 (b) presents the classification result of the MAMEM-SSVEP-II dataset. The accuracy of each class may be determined by its SNR, because the SNR of SSVEP is unevenly distributed across frequencies [4, 5, 6, 7]. Figure 1 (c) shows the confusion matrix of a single subject in the BCI-ERN dataset. We observe a biased classification where most trials were classified as 'correct' due to the imbalance of class samples within this dataset. [8].

A.4 Additional results of model interpretation

This part aims to uncover the characteristics learned from EEG signals. Figure 2 illustrates the gradient response of S3 when performing MI across the channel and the time domains. C4 and C3 channels located on the contralateral side of the brain are activated during the left/right hand MI. Strong responses of left hand MI almost occur across the whole trial, and the conspicuous responses for right hand MI arises at 1-2 seconds. For feet/tongue MI, the responses are strong at the CPz channel located in the midline of the motor cortex. A strong response of the tongue occurs at 0.8-1 seconds, and the strong response of the feet is at 0.9-1.5 seconds in the early experiment. The spatiotemporal distribution of the SSVEP signals is illustrated in Figure 3. Five heatmaps exhibit strong responses at the Oz channel over the visual cortex for all visual stimulation frequencies. In contrast, the distribution across the channel and time domains differs from the one of MI. However, the spatiotemporal distribution within all visual stimulation frequencies is analogous. Figure 4 and 5 depict the gradient response and the spatial topoplot of S7 in BCI-ERN dataset respectively. As shown in the two figures aforementioned, vivid activation in FCz channel located in the midline of the frontal region is elicited on both error and correct stimuli around 0.1 and 0.4 seconds, which is consistent with the observation in [9]. Moreover, Figure 5 exhibits the strong activation distributed

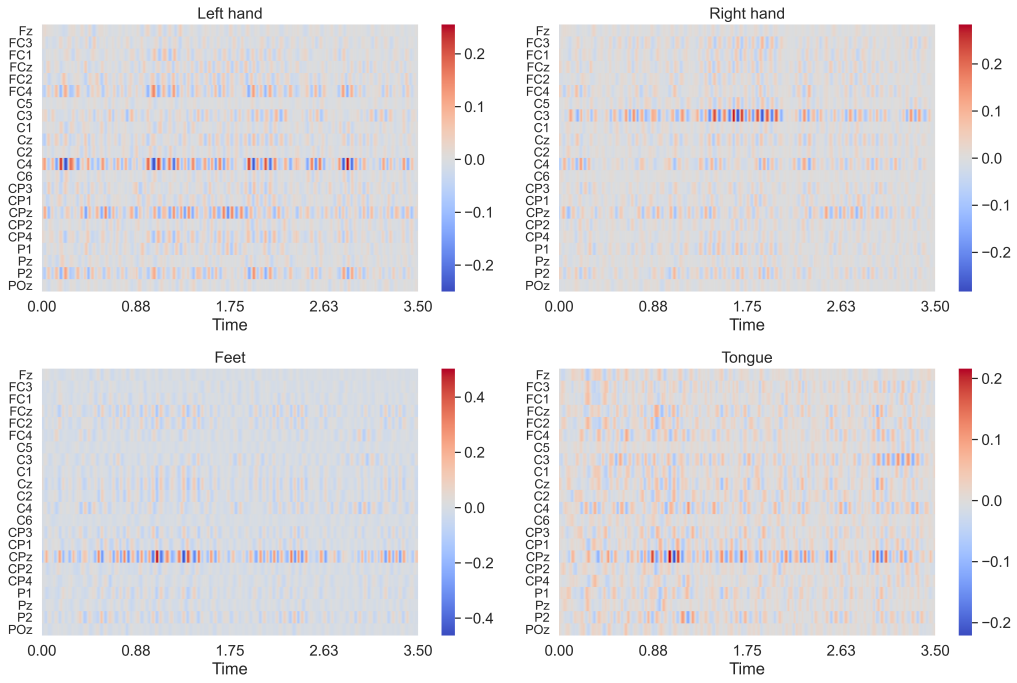


Figure 2: Heatmaps illustrate the gradient response across EEG channels (y-axis) and time (x-axis in seconds) from the visualization of the model S3 in the BCIC-IV-2a dataset for the four motor-imagery classes (left hand, right hand, feet, and tongue). Red/blue pixels indicate strong positive/negative gradient response on the input 22-channel MI EEG data. The discernible gradient responses indicate strong importance of specific EEG channel locations corresponding to the four classes, as observed at C4 (over right motor cortex) for the left hand, C3 (over left motor cortex) for the right hand, CPz (over motor cortex) for the feet and the tongue motor imagery.

over the frontal-central area on the scalp and moderate activation around the occipital region in both classes. In summary, Figure 2, 3, and 4 present different characteristics between MI, SSVEP, and ERN EEG signals learnt by our model.

Regardless of the stimulation frequency, all SSVEP signals present strong activations at the Oz channel (see Figure 6) located visual cortex. The discernible patterns of stimulation frequencies are shown in Figure 7, where we observe strong responses at the fundamental and harmonic frequencies corresponding to the visual stimuli. These results match the traits of SSVEP signals that oscillatory brain activity arises from the visual cortex and resonates with the flickering visual stimulus [10].

Figure 8 indicates the brain activity at each epoch when S3 performs four types of MI. The tendency of all spatial topoplots of the left/right hand MI shows the gradient response activation occurs in the right/left cerebral hemisphere respectively, which matches the model interpretability in the previous part. When it comes to the feet/tongue MI, responses above the midline of the motor cortex are vivid. Moreover, the topoplots across epochs present different but analogous brain activities during the whole trial for all MI classes. Figure 9 presents the spatial distributions across epochs for all types of visual stimulatioi of the MAMEM-SSVEP-II dataset. All epochs present analogous spatial topoplots with strong activation at the Oz channel, and the duration of the activation at the Oz channel lasts until to the end of the trial. In addition, the major gradient responses on each epoch over the scalp are similar for each visual stimulation frequency. In a nutshell, the proposed MAtt can reveal subtle differences underlying similar spatial distributions of each epoch topoplot for five frequencies that are utilized to decode the SSVEP-EEG signals. Our results justify the efficiency and capability of the proposed MAtt in capturing the elusive non-stationarity in the dynamical brain.

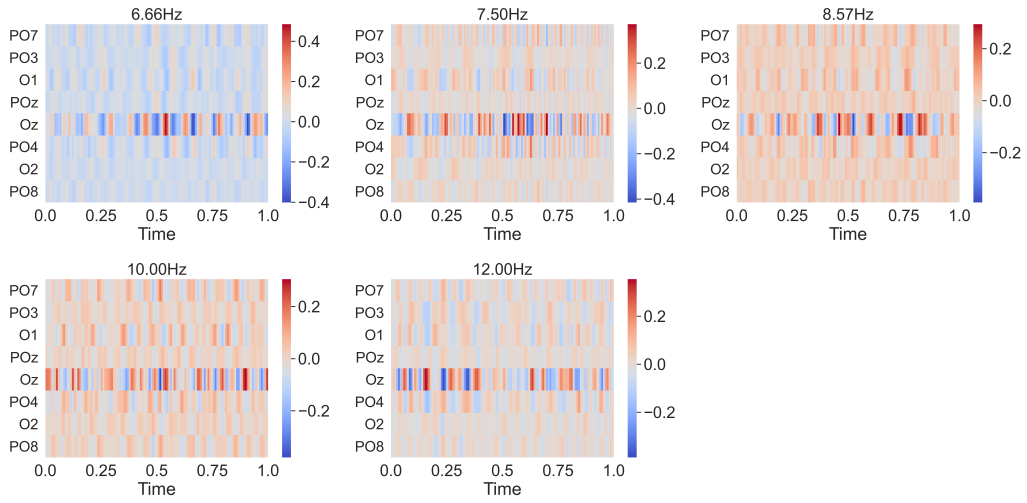


Figure 3: Heatmaps illustrate the gradient response across EEG channels (y-axis) and time (x-axis, in seconds) from the visualization of the model S11 in the MAMEM-SSVEP-II dataset for the five frequency classes (6.66, 7.50, 8.57, 10, and 12 Hz). Red/blue pixels indicate strong positive/negative gradient response on the input 8-channel SSVEP EEG data. The discernible gradient responses indicate strong importance of channel Oz over the visual cortex for all stimulation frequencies.

A.5 Network parameters

The different parameter setups are adopted for different types of EEG tasks. Herein we modified the suggestion in [11] for time-asynchronous SSVEP datasets. The kernel size of the first temporal convolution block is enlarged to (1, 125) and the corresponding number of temporal filters is 100. The number of separable filters is 10, and the number of spatial filters to learn per temporal filter is 8 for this application. Analogous setup for ShallowConvNet, 125 filters with kernel size (1, 40), or about 0.3 seconds for the first temporal convolution block, and 15 spatial filters with kernel size (8, 1), where 8 corresponds to the input number of electrodes per EEG input, in the second layer. For SCCNet, 125 spatial filters are adopted in the spatial convolution block, and the corresponding kernel size is (8, 1). 15 temporal filters with kernel size (1, 36).

A.6 Future work

We still need to investigate the extension of the presented MAtt including the choice of the Riemannian metrics on SPD manifold since the different choices of Riemannian metrics on SPD manifold adopted in the manifold attention module may lead to distinct evolution of the presented MAtt. Meanwhile, although the experimental results justify the robustness of MAtt applied in three different types of EEG datasets (including time-synchronous and time-asynchronous EEG data), we will further validate our proposed method on other EEG datasets to assess its generalizability. As other neuromonitoring modalities such as MEG (magnetoencephalogram), ECoG (electrocorticography), LFP (local field potential), and fNIRS (functional near-infrared spectroscopy) are also multi-channel time series that represents brain activity, we will extend our exploration to test the capability of MAtt on decoding non-EEG neural signals. Last but not least, the neuroscientific insight associated with the attention score in our model requires further investigation including designing new experiments to explore the deeper relevance of this model interpretation.

A.7 Statistical result

Table 2, 3, and 4 depict the multiple-comparison significance testing results with Wilcoxon signed rank test based on Bonferroni correction for MI, SSVEP, and ERN datasets respectively. The aim of the multiple comparison tests (MCT) is to reduce the chance of type I error in this section. Among a variety of correction methods, to rigorously scrutinize the significance of proposed MAtt, herein

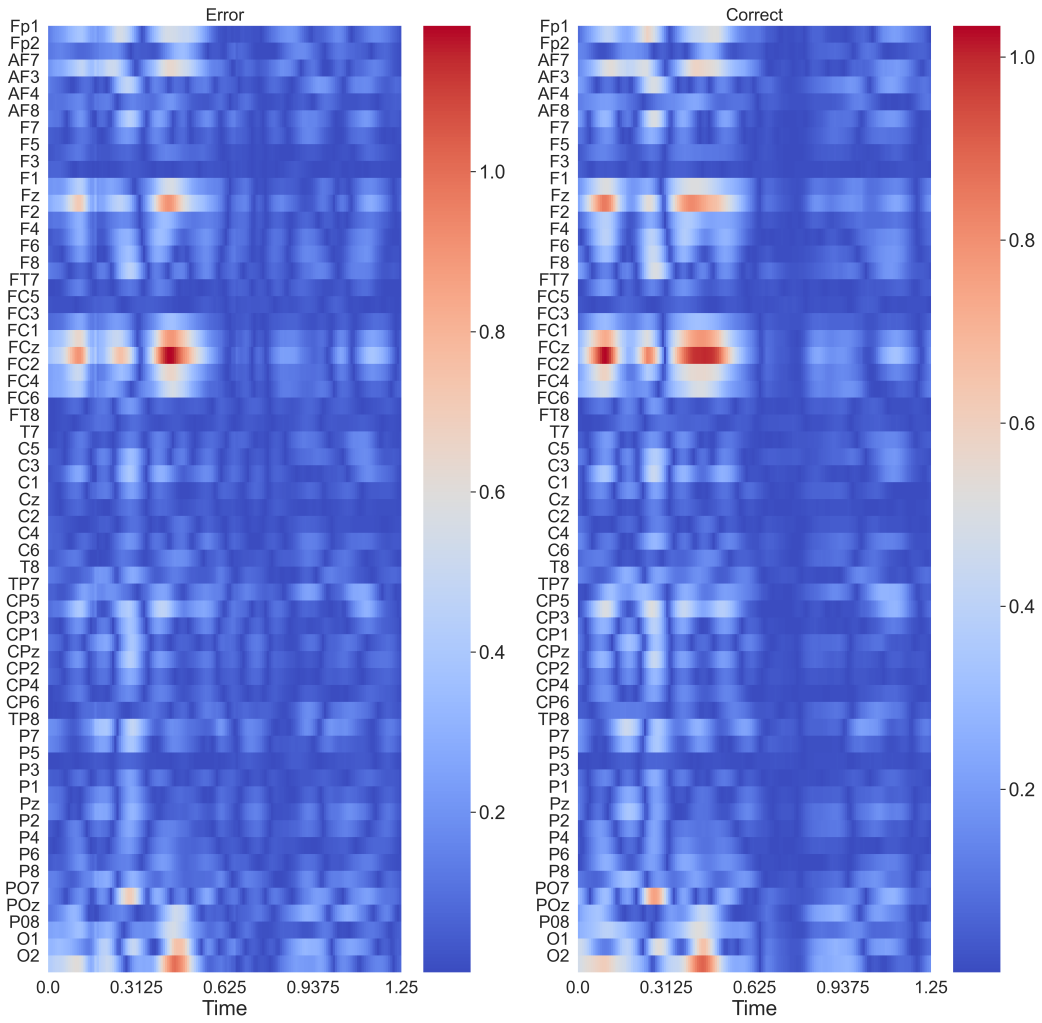


Figure 4: Heatmaps of the gradient response across EEG channels (y-axis) and time (x-axis, in seconds) from the visualization of the model S7 in the BCI-ERN dataset for the classes of 'error' and 'correct' feedback given by the BCI speller. Red/blue pixels indicate strong positive/negative gradient response on an input EEG segment. Consistent gradient response is observed for both classes at FCz around 0.1 and 0.4 second, which is highly consonant with the ERP waveform discrepancy between error/correct stimuli [9].

Table 2: P-value matrix for multiple comparison test (Wilcoxon signed rank test based on Bonfferoni correction) on MI dataset. * statistical significance at reliability levels of 95%.

	EEGNet	EEG-TCNet	FBCNet	MAtt	MBEEGSE	SCCNet	ShallowConvNet
EEG-TCNet	0.11	-	-	-	-	-	-
FBCNet	0.11	0.77	-	-	-	-	-
mAtt	0.11	0.11	0.77	-	-	-	-
MBEEGSE	0.11	1.00	0.55	0.33	-	-	-
SCCNet	0.11	1.00	1.00	0.33	0.33	-	-
ShallowConvNet	1.00	1.00	0.22	0.11	1.00	0.11	-
TCN-Fusion	1.00	0.55	0.11	0.11	1.00	0.22	1.00

Bonferroni correction is adopted since it is insensitive to moderate differences [12]. As shown in table 2, the p-value matrix of the MI dataset demonstrates the smallest p-value between the MAtt against all baseline models are EEG-TCNet, EEGNet, TCNet-fusion, and ShallowConvNet. The first

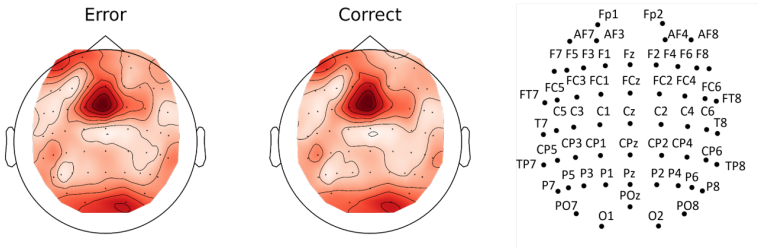


Figure 5: Spatial topomaps for the mean absolute gradient response across time from the visualization of the model S7 in the BCI-ERN dataset for two classes (error and correct). Dark red marks the brains region presenting strong gradient activation at channel FCz over the frontal region for all stimulation frequencies.

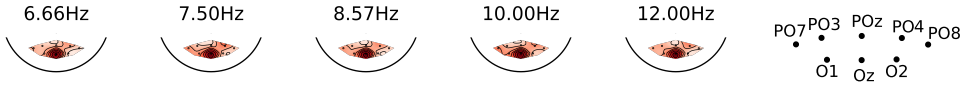


Figure 6: Spatial topomaps for the mean absolute gradient response across time from the visualization of the model S11 in the MAMEM-SSVEP-II dataset for five frequency classes (6.66, 7.50, 8.57, 10, and 12 Hz). Dark red marks the brains region presenting strong gradient activation at channel Oz over the visual cortex for all stimulation frequencies.

Table 3: P-value matrix for multiple comparison test (Wilcoxon signed rank test based on Bonfferoni correction) on SSVEP dataset. * statistical significance at reliability levels of 95%.

	EEGNet	EEG-TCNet	FBCNet	MAtt	MBEEGSE	SCCNet	ShallowConvNet
EEG-TCNet	1.00	-	-	-	-	-	-
FBCNet	1.00	1.00	-	-	-	-	-
mAtt	0.14	0.03*	0.38	-	-	-	-
MBEEGSE	1.00	1.00	1.00	0.14	-	-	-
SCCNet	0.19	0.08	0.52	1.00	0.52	-	-
ShallowConvNet	1.00	1.00	1.00	0.08	1.00	0.19	-
TCN-Fusion	0.14	0.68	1.00	0.06	0.14	0.06	0.38

three models are based on the temporal-causal-convolution-based DL models. We infer the three models may contribute insignificantly to the MI-EEG decoding task. Other attention-based (such as MBEEGSE) and self-defined temporal feature exploration method (FBCNet) has a little higher p-values against MAtt on the other hand. Table 3 illustrates the p-value matrix for the second SSVEP dataset. The p-value in the cell that corresponds to MAtt and EEG-TCNet denotes the significant

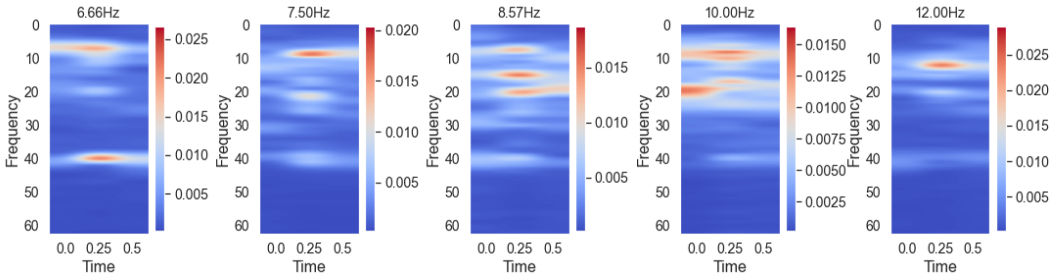


Figure 7: Time-frequency spectrograms from the visualization of the model S11 in the MAMEM-SSVEP-II dataset for the five frequency classes (6.66, 7.50, 8.57, 10, and 12 Hz). Strong response of SSVEP is marked by dark red at specific frequency bands and time intervals. Increased response of SSVEP is found at the fundamental and harmonic frequency corresponding to each stimulation.

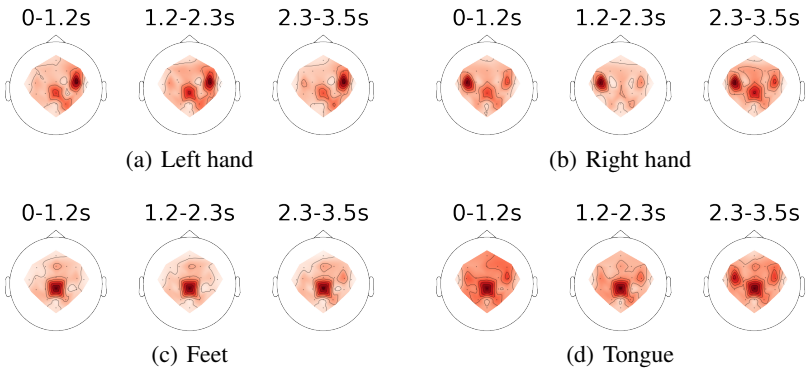


Figure 8: Spatial topoplots for each epoch of S3 on the BCIC-IV-2a dataset for four MI tasks(left hand, right hand, feet, and tongue). Strong mean absolute gradient activation of MI is marked by dark red in specific brain regions over the scalp.

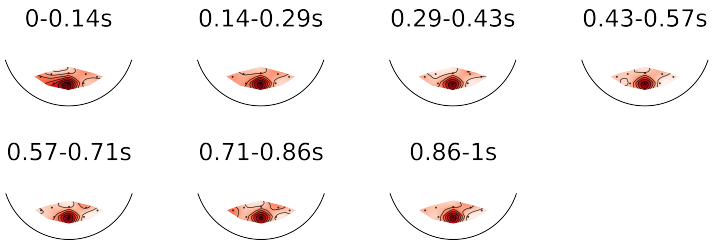
Table 4: P-value matrix for multiple comparison test (Wilcoxon signed rank test based on Bonfferoni correction) on ERN dataset. * statistical significance at reliability levels of 95%.

	EEGNet	EEG-TCNet	FBCNet	MAtt	MBEEGSE	SCCNet	ShallowConvNet
EEG-TCNet	1.00	-	-	-	-	-	-
FBCNet	0.05	0.01*	-	-	-	-	-
mAtt	1.00	1.00	0.01*	-	-	-	-
MBEEGSE	1.00	1.00	0.03*	1.00	-	-	-
SCCNet	1.00	1.00	0.06	0.43	1.00	-	-
ShallowConvNet	1.00	1.00	0.07	1.00	1.00	1.00	-
TCN-Fusion	0.81	0.43	0.31	1.00	1.00	1.00	1.00

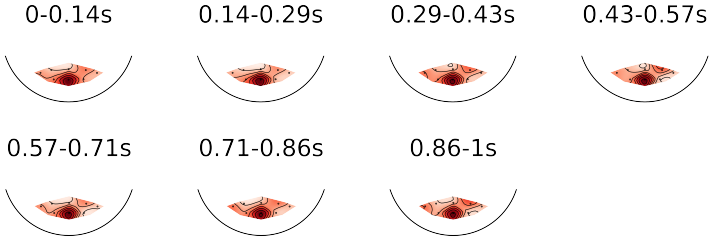
difference between the proposed MAtt and EEG-TCNet. By contrast, MCT is insensitive to detect the difference between MAtt and SCCNet. For ERN dataset, the p-value matrix is shown in table 4. The table exhibits the corresponding p-value between all models with each other. Although the EEG-TCNet ourperofrms the proposed MAtt slightly on the ERN decoding, the difference is statistically significant according to the p-value. Furthermore, among all baseline models, the p-values between the MAtt and FBCNet suggest that MAtt has a stronger capacity than FBCNet in decoding the ERN dataset. In summary, the p-value matrices above illustrate the significance of performance comparison among all models.

A.8 Limitation

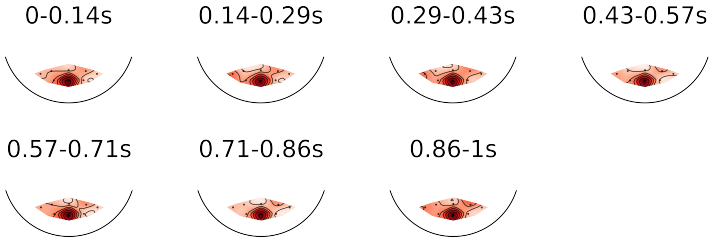
In our framework, vacuum permittivity ϵ is added on all main diagonal elements of covariance $cx_i x_i^T$ to ensure the rigor of SPD matrix. But the operation may cause the repeated singular value ϵ in S_i. Therefore, we proposed possible solutions for this issue: 1) Let $m < n$ when dividing the embeddings into several time segments, reducing the possibility of getting low-rank S_i; 2) Let ϵ be randomly drawn from a specific distribution, such as $Uniform\ (1e-8, 1e-4)$ to solve this issue, which is also a practicable solution; 3) Use the derivative of a low-rank matrix [13] to cope with this issue.



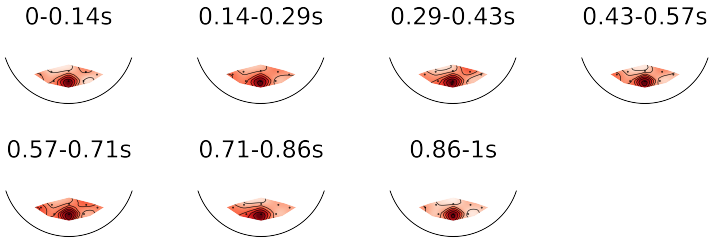
(a) 6.66 Hz



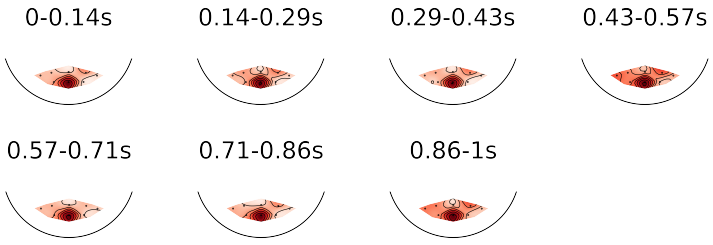
(b) 7.50 Hz



(c) 8.57 Hz



(d) 10.0 Hz



(e) 12.0 Hz

Figure 9: Spatial topoplots over the occipital region on the scalp for each epoch of S11 on the MAMEM-SSVEP-II dataset for five visual stimulation frequencies(6.66, 7.50, 8.57, 10, and 12 Hz). Strong mean absolute gradient activation of SSVEP is marked by dark red in specific brain regions over visual cortex.

References

[1] Frédéric Barbaresco. Innovative tools for radar signal processing based on Cartan's geometry of SPD matrices & information geometry. In *2008 IEEE Radar Conference*, pages 1–6. IEEE, 2008.

[2] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to BCI classification. In *International conference on latent variable analysis and signal separation*, pages 629–636. Springer, 2010.

[3] Rudrasis Chakraborty, Jose Bouza, Jonathan Manton, and Baba C Vemuri. Manifoldnet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] Wang Yijun, Wang Ruiping, Gao Xiaorong, and Gao Shangkai. Brain-computer interface based on the high-frequency steady-state visual evoked potential. In *Proceedings. 2005 First International Conference on Neural Interface and Control, 2005.*, pages 37–39. IEEE, 2005.

[5] Toshihisa Tanaka, Cheng Zhang, and Hiroshi Higashi. SSVEP frequency detection methods considering background EEG. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 1138–1143. IEEE, 2012.

[6] Chun-Shu Wei, Yuan-Pin Lin, Yijun Wang, Yu-Te Wang, and Tzyy-Ping Jung. Detection of steady-state visual-evoked potential using differential canonical correlation analysis. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 57–60. IEEE, 2013.

[7] Chun-Shu Wei, Toshiaki Koike-Akino, and Ye Wang. Spatial component-wise convolutional network (SCCNet) for motor-imagery EEG classification. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 328–331. IEEE, 2019.

[8] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during P300-based spelling. *Advances in Human-Computer Interaction*, 2012, 2012.

[9] Greg Hajcak. What we've learned from mistakes: Insights from error-related brain activity. *Current Directions in Psychological Science*, 21(2):101–106, 2012.

[10] Christoph 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*, 137(3):346–353, 2001.

[11] Nicholas Waytowich, Vernon J Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M Vettel. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of neural engineering*, 15(6):066031, 2018.

[12] Stephen Olejnik, Jianmin Li, Suchada Supattathum, and Carl J Huberty. Multiple testing and statistical power with modified Bonferroni procedures. *Journal of educational and behavioral statistics*, 22(4):389–406, 1997.

[13] James Townsend. Differentiating the singular value decomposition. Technical report, Technical Report 2016, https://j-towns. github. io/papers/svd-derivative . . . , 2016.