# Combining EEGNet with SPDNet towards an end-to-end architecture for imagined speech decoding

Georgios Rousis
Dept. of Electrical & Computer Engineering
Aristotle University of Thessaloniki
Email: geroussis@gmail.com
Corresponding Author

Fotis P. Kalaganis
Information Technologies Institute
CERTH
Email: fkalaganis@iti.gr

Spiros Nikolopoulos
Information Technologies Institute
CERTH
Email: nikolopo@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute
CERTH
Email: ikom@iti.gr

Panagiotis C. Petrantonakis
Dept. of Electrical & Computer Engineering
Aristotle University of Thessaloniki
Email: ppetrant@ece.auth.gr

***Abstract*—Imagined speech is the mental task where individuals internally simulate the articulation of a prompt without actual vocalization. Recently, it gained widespread attention due to its simplicity and intuitiveness as a Brain-Computer Interface (BCI) paradigm. Hence, the decoding of imagined speech from brain signals emerges as a pivotal challenge addressed with various signal processing and machine learning techniques documented in the literature. The most commonly employed neuroimaging method is Electroencephalography (EEG) because of its non-invasive nature, low cost and high temporal resolution. Recent attempts of deciphering imagined speech from EEG signals deploy Convolutional Neural Network (CNN) architectures such as shallow Conv Net, deep Conv Net and EEGNet while others use Cross-Covariance (CCV) matrices as an alternative form of signal representation. Our novel architecture combines EEGNet with CCV matrices, extracting discriminative features from the latter with the use of bilinear transformations as proposed in the SPDNet architecture. Our method is validated on two publicly available datasets and exhibits on par with State-of-the-Art performance, while substantially surpassing EEGNet performance on both datasets.**

## I. Introduction

Imagined speech (also known as inner, silent or covert speech) is a form of thinking in terms of sound - one imagines of articulating a prompt without actually moving the articulators. Regarded as a fundamental aspect of conscious life, the act of engaging in inner dialogue is a commonplace practice among us, human beings. This internal discourse serves the purpose of rehearsal, prepare our expressions in anticipation of an impending speech or interview.

The process of capturing imagined speech through brain signals and translating them into words or phrases has been a far-fetched yet intriguing scientific and technological ambition for decades. Recent advances in technology involve deploying imagined speech as a simple and intuitive paradigm in the context of Brain-Computer Interface (BCI) applications [1]. This particular mental task stands out as an advantageous option as it directly conveys the user's intention and thus suggests a natural way of controlling external devices. Thus, a brain signals decoding paradigm based on imagined speech is particularly suitable for systems that restore basic communication to individuals who are not capable of conventional speech articulation due to an accident or a disorder of the Central Nervous System (CNS). For this purpose, Electroencephalography (EEG) signals are usually preferred to record brain activity due to their non-invasive nature and low cost. Despite its low spatial resolution compared to other neuroimaging methods such as fMRI, EEG can capture brain activities that take place within a time frame of a few milliseconds [2].

Preliminary attempts of decoding imagined speech from EEG signals comprised extraction of statistical features (such as mean, variance, skewness and kurtosis) [3], [4] or wavelet transform coefficients and classification [5], [6], [7], [8], [9], [10] with conventional machine learning algorithms such as Support Vector Machines (SVM), Deep Belief Networks (DBN) and Extreme Learning Machines (ELM). In another approach, instead of working with raw EEG data, researchers used Cross Covariance matrices (CCV) encoding statistical correlation between EEG channels [11], [12], [13]. A basic characteristic of CCV matrices, namely being symmetric positive definite, allows for alternative processing directions utilizing basic manifold properties that originate from Riemannian geometry. Recent approaches following this pathway have achieved remarkable performance on similar BCI applications [14], [15], [13], [16], [17]. These results motivated the integration of Riemannian geometry with deep learning techniques,

This work is part of project BINGO, implemented in the framework of H.F.R.I. call "Basic research Financing (Horizontal support of all Sciences)" under the National Recovery and Resilience Plan "Greece 2.0" funded by the European Union – NextGenerationEU (H.F.R.I. Project Number: 15986).

one of the most prominent example of which is the SPDNet [18]. More recent studies focused on implementing different deep learning methods, mostly Convolutional Neural Network (CNN) architectures, that demonstrated very promising results, at least when examined on other BCI paradigms. Some of the widely used architectures are the shallow ConvNet, the deep ConvNet [19] and the EEGNet [20].

Our approach utilizes some basic concepts, stemming from both CNNs and CCVs, in an effort to combine the best of the two worlds towards building a novel, end-to-end, deep learning architecture. Specifically, the first part of the introduced architecture consists of a convolutional layer of temporal filters as implemented in EEGNet. The output feature maps (i.e., EEG signals filtered in the temporal domain) correspond to selected frequency bands where the most significant brain activation occured. The second part involves converting the resulting maps into CCV matrices. Each matrix is then subjected to multiple linear transformations such that the output matrices also lie in Riemannian manifolds (potentially of varying dimensions) in accordance with SPDNet architecture. Ultimately, the Log-Euclidean metric is computed for each matrix and the information is transferred to a fully connected layer for the purpose of classification. As stated earlier, the combination of the above mentioned parts constitute an end-to-end trainable network. In essence, the proposed architecture calculates CCV matrices that can capture the brain connectivity structure that underpins the imagined speech paradigm at various frequency bands. The motivation for employing this particular architecture for the task at hand is related to the ”dual stream model” according to which several brain regions are involved and interconnected during speech formulation and understanding [21].

The introduced method is validated on two publicly available datasets that revolve around the intuitive paradigm of inner speech decoding, namely the Kara One [4] and the 2020 BCI Competition [22]. The former contains multichannel EEG data from 11 distinct imagined prompts, 7 phonemes or syllables and 4 words, whereas the later consists of data from 5 imagined words. Two common EEG decoding approaches are also tested on the datasets, namely MFCC features with SVM classifiers and EEGNet architecture. Their performance is then compared with the proposed method. In both cases, the EEGNet-SPDNet architecture significantly outperformed the other two methods. For the first dataset, the network achieved an average test accuracy exceeding 24% on a classification task with 11 classes, a result comparable to the top performing models of the particular problem. In the second case, the introduced method exhibited performance similar to the competition's second place, reaching almost 67% average accuracy. Moreover, the substantial improvement over EEGNet performance illustrates the effectiveness of the SPDNet-based component within the proposed architecture, extracting features that more accurately represent the imagined prompts from the EEG signals.

## II. Datasets

### A. Kara One

The dataset consists of 14 participants, with an average age of 27, who were instructed to imagine pronouncing and consequently to speak aloud 7 phonemes or syllables: (/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/) and 4 words: ( pat, pot, knew, and gnaw) over the course of 30 to 40 minutes. The participants were seated in front of a computer monitor and a Microsoft Kinect camera and a research assistant placed an EEG cap on their heads. The data collected combine 3 modalities: EEG signals, face tracking and audio. A 64-channel Neuroscan Quick-cap was used, the electrode placement followed the 10-20 rule and the data were sampled at 1kHz.

Each trial consisted of 4 states. At first, there is a 5-second rest state where the participants were instructed to relax. Next, in the stimulus state, the prompt text appeared on the screen and its corresponding audio played from the speakers. A 5-second imagined speech state follows where the participants imagined pronouncing the prompt without moving their articulators and finally they spoke the prompt aloud. In this work, we only employed the EEG segments corresponding to imagined speech. The data from 11 out of 14 participants were utilized to maintain uniformity in the number of trials. For each participant, 132 trials were conducted, 12 for each prompt.

### B. 2020 BCI Competition dataset

In this case, 15 participants, aged between 20-30 years, were instructed to imagine pronouncing five words/phrases, namely: (“hello,” “help me,” “stop,” “thank you,” and “yes”). During the experiment, the subjects were seated in a comfortable chair in front of a 24-inch LCD monitor screen and were asked to solely focus on the given task without moving their articulators nor making any sound. For the recording, 64 EEG electrodes following a 10-20 international configuration were used.

An auditory cue of a randomly chosen prompt is introduced to the participants for 2 s, followed by the visual cue of a cross mark on the screen that lasted between 0.8-1.2 s. The subjects imagined pronouncing the given prompt as soon as the cross mark disappears. During this phase, the participants received no stimulus at all in order to avoid any unwanted brain activity. The cross mark presentation on the screen and the consequent imagined speech phase (2s) were repeated 4 times in a row for each random cue. Before proceeding to the next word/phrase, the subjects were given 3s to relax and clear their minds.

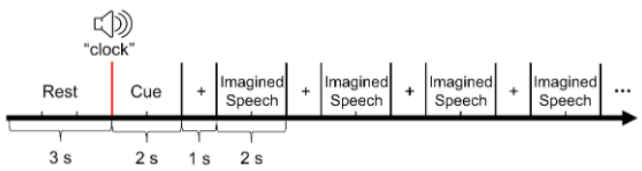


Fig. 1. Timeline of the experimental procedure. Image source:https://osf.io/ymvjz.

In total, 400 trials were conducted for each subject, 80 trials per prompt, out of which 300 are dedicated for training, 50 for validation and the remaining 50 for testing.

## III. Methodology

### A. Preprocessing

In the Kara One dataset, the raw EEG signals are filtered using a 3rd order high-pass Butterworth filter with cut-off frequency 1Hz. A wavelet ICA algorithm is then employed for the purpose of artifact removal and denoising. The EEG segments corresponding to the imagined speech stage are ultimately fed to the classification algorithms.

No preprocessing steps are followed in the second dataset since the 2020 BCI competition provided the signals in the form of labelled EEG trial segments.

### B. Network architecture

The proposed architecture is an end-to-end trainable network that combines the convolutional temporal filters as implemented in EEGNet with the linear Symmetric Positive Definite (SPD) matrix transformations utilized in SPDNet architecture as shown in Fig.2.

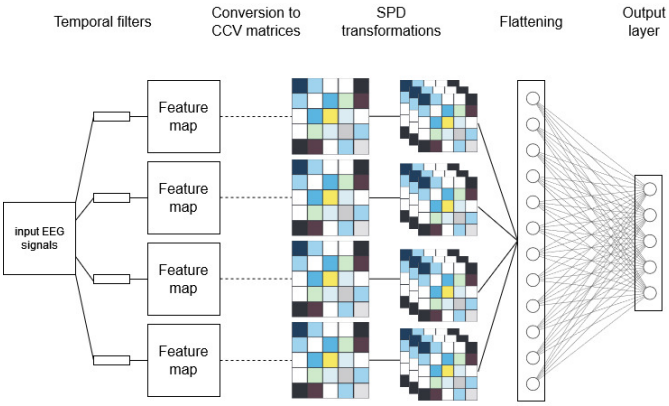


Fig. 2. EEGNet-SPDNet architecture.

Below we provide the key components of the network, tailored to imagined speech decoding.

- The temporal filters are convolutional 2D filters with shape $(1, \frac{f_s}{2})$, where $f_s$ is the sample frequency. The convolutional layer is followed by a batch normalization layer, a non-linearity ReLU, a mean pooling layer of size (1,4) and a dropout layer with probability 0.5. The filters output feature maps containing the EEG signal in different band-pass frequencies leaving the channel dimension unaffected.
- The different versions (feature maps) of the EEG signal are then converted to the corresponding covariance matrices.
- Each matrix is then subjected to multiple bilinear transformations that map SPD matrices to other SPD matrices of different dimension. If the input matrix is denoted as: $\mathbf{X} \in R^{N\times N}$, the output matrix as: $\mathbf{Y} \in R^{M\times M}$ and the transformation matrix as: $\mathbf{W} \in R^{M\times N}$ then the mapping is as follows:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}\mathbf{W}^T. \tag{1}$$

The trainable parameters in this part are the elements of the transformation matrix. The output matrix $Y$ is symmetric positive definite if the transformation matrix $W$ is full rank on the rows. Thus, an exclusive optimizing procedure take place here such that this essential matrix property is preserved. The transformation is followed by a non-linearity layer the function of which is the rectification of the eigenvalues. Given the diagonalization of $\mathbf{Y}$ as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ and the output matrix denoted as: $\mathbf{Z}$, the rectification is:

$$\mathbf{Z} = \mathbf{U}max(\varepsilon\mathbf{I}, \mathbf{\Sigma})\mathbf{U}^T \tag{2}$$

where $\varepsilon > 0$ is the rectification threshold. The $max(\varepsilon\mathbf{I}, \mathbf{\Sigma})$ is a diagonal matrix where each diagonal value (eigenvalue) of $\mathbf{\Sigma}$ is replaced by: $max(\varepsilon, e_i)$. Essentially, this process prevents eigenvalues from approaching zero. Finally, the matrices are mapped to Euclidean space via the implementation of Log-Euclidean metric and consequently to a linear output layer through flattening.

## IV. Classification Results

### A. Kara One

Fig.3 (top) presents the overall accuracy of the proposed architecture when tested with Kara One dataset as well as the accuracy scores obtained for each subject. Both training and validation are conducted on a personalized level, fitting a different network for each subject. It is apparent from the confusion matrix (Fig.3 - Top Right) that the network was able to adequately distinguish phoneme from word prompts but hardly disentangled phonetically similar words (pat/pot and knew/gnaw). The best accuracy is measured in one letter phonemes (/m/, /n/). The novel method implemented in this work exhibits superior performance compared to the EEGNet that barely exceeded random level (9.1%). When it comes to intra-subject multi-classification of words and phonemes, our approach outperforms handcrafted features (i.e. MFCC) combined with SVMs [23] by 4%. It achieved similar overall accuracy with inter-subject training and validation attempts that employed CNN architectures as well [24]. The overall performance of our approach as shown in Table I surpassed all but one competitive approach found in the literature regarding this particular classification task [8].

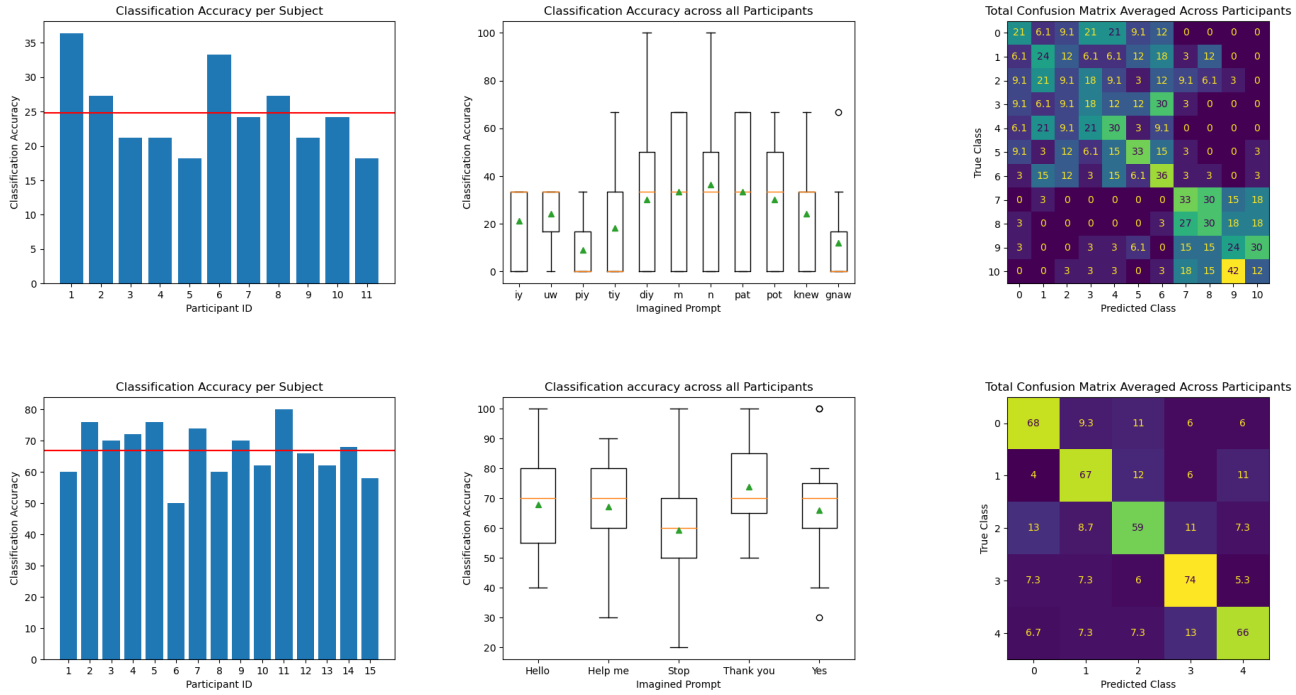


Fig. 3. The overall and subject-wise performance on Kara One dataset (top), 2020 BCI Competition dataset (bottom). Left: Classification accuracy across all participants for each imagined prompt. Middle: Classification accuracy per subject and overall (red line). Right: Total confusion matrix

TABLE I
MEAN ACCURACY FOR EACH CLASSIFICATION METHOD TESTED ON TWO DISTINCT DATASETS

| | **Mean Accuracy (%)** | |
|---|---|---|
| | **Kara One** | **2020 BCI Comp.** |
| **MFCC - SVM** | 17.84 | 37.86 |
| **EEGNet** | 14.05 | 50.26 |
| **EEGNet - SPDNet** | 24.79 | 66.93 |

### *B. 2020 BCI Competition*

Training and validation took place separately for each subject in this case as well. A major observation from the overall accuracies for each subject (Fig.3 - Bottom Left) is the subject variability: there are subjects with accuracy as high as 80% (S11) while others that do not exceed 60% (S6,S15). The complexity of the specific task (imagined speech) could be one factor contributing to this. The novel architecture employed significantly outperformed EEGNet in this case as well (Table I). As mentioned earlier, the attained performance here is on par with the top results of the competition, namely similar to the competitor approach with the second highest average accuracy [25].

## V. DISCUSSION

In this paper, we proposed a novel architecture that combines the temporal filters from EEGNet with the SPD matrix transformations as implemented in SPDNet into an end-to-end trainable network that extracts connectivity features from EEG signals. The method was validated on two distinct imagined speech datasets and substantially outperformed EEGNet, a widely used convolutional network architecture in various EEG decoding tasks. Moreover, the achieved performances are close to the State-of-the-Art on the employed datasets. This result implies that the aforementioned transformations are capable of capturing important information from the EEG signals, and thus it supports the hypothesis that correlation and connectivity of distinct brain regions play a significant role as far as the mental task of imagined speech is concerned. It may also suggest that covariance matrix is more suitable form of EEG signal representation for decoding and feature extraction purposes (than raw signal form) as it allows for mappings in Riemannian spaces that efficiently depict brain activity.

The broader subject of decoding imagined speech in a BCI paradigm comes with several challenges, manifestations of which are apparent in the above presented results as well. Although promising, the classification performance of EEG signals decoding is not yet sufficient in terms of developing robust BCI systems, while the imagined prompts chosen are relatively simple and short in number. Additionally, subject variability signifies that the particular mental task is often not adequately comprehensible from the participants. Future work

is hence needed to examine the full potential of imagined speech paradigm and EEG as a neuroimaging method in particular.

## REFERENCES


[1] C. Herff, A. de Pesters, D. Heger, P. Brunner, G. Schalk, and T. Schultz, *Towards Continuous Speech Recognition for BCI*. Cham: Springer International Publishing, 2017, pp. 21–29. [Online]. Available: https://doi.org/10.1007/978-3-319-57132-4_3.

[2] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press, Jan. 2014. [Online]. Available: https://doi.org/10.7551/mitpress/9609.001.0001

[3] B. Min, J. Kim, H.-J. Park, and B. Lee, "Vowel imagery decoding toward silent speech bci using extreme learning machine with electroencephalogram," *Biomedical Research International*, pp. 1–11, 2016.

[4] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 992–996, 2015.

[5] A. Jahangiri, J. M. Chau, D. R. Achanccaray, and F. Sepulveda, "Covert speech vs. motor imagery: a comparative study of class separability in identical environments," in *Annual International Conference of the IEEE Engineering Medical and Biology Society*. IEEE, 2018.

[6] S. F. Jahangiri A, "The relative contribution of high gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in eeg data," *Journal of Medical Systems*, vol. 43, no. 2, p. 20, 2019.

[7] S. F. Jahangiri A, Achanccaray D, "A novel eeg-based four-class linguistic bci,," in *Annual International Conference of the IEEE Engineering Medical and Biology Society*. IEEE, 2019, pp. 3050–3053.

[8] A. G. R. J. T. Panachakel and T. V. Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in *16th India Council International Conference*. IEEE, 2019, pp. 1–4.

[9] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Eeg classification of covert speech using regularized neural networks," in *ACM Transactions on Audio, Speech, and Language Processing*. IEEE, 2017.

[10] A. A. Torres-García, C. A. Reyes-Garcia, and L. Villaseñor-Pineda, "Toward a silent speech interface based on unspoken speech," in *BIOSIGNALS 2012 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, 2012, pp. 370–373.

[11] F. S. Saha P. and A.-M. M., "Deep learning the eeg manifold for phonological categorization from active thoughts," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2762–2766.

[12] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from eeg," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1. IEEE, 2020.

[13] F. P. Kalaganis, K. Georgiadis, V. P. Oikonomou, S. Nikolopoulos, N. A. Laskaris, and I. Kompatsiaris, "Exploiting approximate joint diagonalization for covariance estimation in imagined speech decoding," in *International Conference on Brain Informatics*. Springer, 2023, pp. 409–419.

[14] F. P. Kalaganis, N. A. Laskaris, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "A riemannian geometry approach to reduced and discriminative covariance estimation in brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 245–255, 2019.

[15] F. P. Kalaganis, N. A. Laskaris, V. P. Oikonomou, S. Nikopolopoulos, and I. Kompatsiaris, "Revisiting riemannian geometry-based eeg decoding through approximate joint diagonalization," *Journal of Neural Engineering*, vol. 19, no. 6, p. 066030, 2022.

[16] K. Georgiadis, F. P. Kalaganis, V. P. Oikonomou, S. Nikolopoulos, N. A. Laskaris, and I. Kompatsiaris, "Rneumark: A riemannian eeg analysis framework for neuromarketing," *Brain Informatics*, vol. 9, no. 1, p. 22, 2022.

[17] ——, "Harneshing the potential of eeg in neuromarketing with deep learning and riemannian geometry," in *International Conference on Brain Informatics*. Springer, 2023, pp. 21–32.

[18] Z. Huang and L. V. Gool, "A riemannian network for spd matrix learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[19] S. R. et al., "Deep learning with convolutional neural networks for eegdecoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[20] V. Lawhern, S. AJ, W. NR, G. SM, H. CP, and L. BJ., "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 2018.

[21] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, p. 393–402, 2007.

[22] B. C. Committee, "2020 international bci competition," 2020, https://osf.io/pq7vb/.

[23] C. Cooney, R. Folli, and D. Coyle, "Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from eeg," in *2018 29th Irish Signals and Systems Conference (ISSC)*, 2018, pp. 1–7.

[24] A.-L. Rusnac and O. Grigore, "Generalized brain computer interface system for eeg imaginary speech recognition," in *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, 2020, pp. 184–188.

[25] J. J. et al., "2020 international brain-computer interface competition: A review." *Frontiers in human neuroscience*, vol. 16, 2022.