Inner Speech Recognition

Madi Turgunov
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
madi.turgunov@nu.edu.kz

Yernur Aubakirov
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
yernur.aubakirov@nu.edu.kz

Dias Kuatbekov
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
dias.kuatbekov@nu.edu.kz

Alua Shamshiden
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
alua.shamshiden@nu.edu.kz

***Abstract*—Brain-Computer Interfaces serve as direct gateway for communication of humans with computers. A traditionally unexplored paradigm of BCI is Inner Speech, which deserves more attention and research. This project work intends to discover how traditional Machine Learning (ML) and Deep Learning (DL) algorithms perform on the task of Inner Speech classification. We report that classifying Inner Speech (imagination of one's own voice) is indeed possible via providing models that provide classification accuracy above chance.**

We find that traditional classification algorithms do not provide sufficient accuracy when it comes to analyzing Inner Speech. However, some more complex models are able to showcase a degree of reliability. We also showcase how our work compares to existing literature, and propose reasons as to how to improve the results.

***Index Terms*—BCI, Inner Speech classification, EEGNet**

I. Introduction

Brain-Computer Interfaces (BCIs) represent systems that enable users to control external devices or communicate through brain activity. Designed primarily for individuals with disabilities, BCIs offer a non-invasive communication channel by translating neural signals into digital commands for assistive applications like speech synthesizers, wheelchairs, or neural prostheses.

While several BCI paradigms have been developed in the past, each with its unique benefits, they often present drawbacks that limit their practical applicability. For example, motor imagery BCIs require extensive training, P300 and SSVEP paradigms rely on continuous attention to external stimuli, and letter-by-letter spelling or movement imagination demands significant mental effort [1]. Additionally, most BCI systems suffer from low Information Transfer Rates (ITR) and require individual calibration.

To address these challenges, inner speech recognition has emerged as a promising alternative communication paradigm for BCIs. Inner speech, the internalized process of speech without articulation, engages brain regions associated with language comprehension and production [2]. By monitoring these brain areas, it is theoretically possible to develop a BCI that classifies neural representations of imagined words.

While previous studies have explored inner speech classification using invasive methods like electrocorticography (ECoG), relatively little research exists on inner speech classification using non-invasive EEG signals [3]. Given the importance of non-invasiveness and accessibility in BCI applications, our study focuses on classifying inner speech from EEG data.

Our study aims to investigate the feasibility of classifying inner speech using EEG data. Leveraging a recently published dataset "Thinking out loud" containing EEG recordings during inner speech of four imagined words, we employ a convolutional neural network (CNN), LDA and SVM models for our multi-class classification task [2]. CNNs, a subclass of deep learning architectures, excel at learning spatial and temporal representations in EEG data, eliminating the need for time-consuming preprocessing and feature engineering. Additionally, we utilize Linear Discriminant Analysis (LDA) to find optimal feature combinations for clear class separation and Support Vector Machines (SVMs) to identify hyperplanes for effective class distinction in high-dimensional spaces. This multi-model approach aims to enhance classification accuracy and contribute to the development of user-friendly Brain-Computer Interface (BCI) systems. Furthermore, we conduct a comparative analysis of our results with existing literature to validate the effectiveness of our proposed methodology.

II. Methods

A. Dataset

1) Data Collection Tasks: Inner speech processes were examined in the dataset to be employed in the potential BCI applications. Participants completed three type of data recording tasks: inner speech, pronounced speech, and visualization [2].

Inner Speech: Participants silently imagined giving a word-based command to a computer, focusing on the internal voice without articulation(neither hand nor lips). Pronounced Speech: Participants articulated the word-based commands aloud to differentiate motor activity between pronounced and inner speech. Visualization: Participants mentally moved a circle on-screen in response to directional cues. This aimed to isolate activity related to visual and spatial aspects that might overlap with inner speech. The dataset design facilitates the localization of neural activation sources and network connections involved in inner speech. The pronounced speech condition helps isolate motor activity, differentiating it from inner speech processes. The visualization task helps identify activity related to visual and spatial thinking that might be present during inner speech [2].

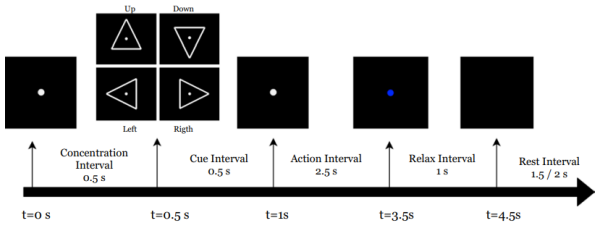

Fig. 1. Trial workflow

Set	Subjects	Samples	Channels	Trials
Train	10	512	128	**160**
Test	10	512	128	**40**
Overall	10	512	128	**200**

TABLE I
SET SPLIT DISTRIBUTION

Set	”Up”	”Down”	”Left”	”Right”
Train	40	40	40	40
Test	10	10	10	10
Overall	50	50	50	50

TABLE II
TEST LABELS DISTRIBUTION

2) Data Collection: The paper includes Electroencephalography (EEG), Electrooculography (EOG), and Electromyography (EMG) data were collected via a BioSemi ActiveTwo system with 128 active EEG channels and 8 external EOG/EMG channels [2]. An appropriately sized EEG head cap with prefixed electrode positions was used, impedance below 40 Ohm was reached via ”signagel” conductive gel.

3) Dataset Pre-Processing: The dataset already included a certain amount of pre-processing, with several filters, such as a 0.5-100Hz zero-phase band-pass filter, a 50Hz Notch filter to remove the powerline noise. The final data from the dataset was presented at a sampling rate of 254Hz, and the data was cut to a length of 4.5s. An Independent Component Analysis (ICA) was already applied to the dataset to isolate and remove signal artifacts (e.g., eye blinks, muscle movement) from the EEG channels. Finally, mouth movement during inner speech or visualization trials is detected by applying a simple threshold- based method on EXG7 and EXG8 (mouth area electrodes) [2]. Trials exceeding the movement threshold are excluded in order to ensure the focus purely on internal thought processes without the inference of motor functions. The final resultant data to be experimented on included only the 128 electrode EEG readings, since the 8 external channels were used in order to remove blink, gaze and mouth movement.

B. Data Processing

The data processing step was by far the longest and most tedious step in this experiment. Whilst the dataset used already had pre-processing applied to it [2], more pre-processing was applied on top in order to attempt to improve accuracy further. Many different band-pass filters, PCA and other transformations were applied on the data, however, none seemed to improve the performance of any of the models. In the end, the transformations that did improve performance was derived, which was then used throughout the experiment. The ”useful” segment of time from 0.5s to 3.5s was used when processing the data. The time slot falls directly on the cue and action interval (represented on Figure 1). On top of the already pre-applied 0.5-100Hz band-pass filter and ICA, an additional 8-30Hz band-pass filter was also applied. The reasoning for that is that the 8-30Hz frequency range combines the μ-rhythm (8-13Hz), associated with motor imagery, and the β-rhythm (13-30Hz), which is associated with language processing, active thinking and cognitive control [4].

However, this, too, yielded poor results. A Hilbert transform was applied afterwards, which, surprisingly, improved performance [5].

All 10 subjects' data was filtered with this process. For each of the four possible word (Up, Down, Left and Right) a class value of *0, 1, 2 and 3*, respectively, were assigned. A typical processed EEG signal for each class is shown in Figure 2.

The dataset was split, with 20% of the dataset being put into the *test set*, and 80% of the dataset being put into the *train set*.

With there being **10** subjects, each subject having **200** trials, each trial having **128** channels, with each channel having **512** recordings (samples), the training and test set were split according to the distributions being visible on Table I.

Out of the 200 trials, there were **50** trials of each label, making this dataset perfectly balanced. When splitting into train and test datasets, the values were stratified, thus, being perfectly split according to table II.

C. CSP-LDA

The first attempt at classifying the data was done with the help of Linear Discriminant Analysis (LDA). In order to better detect useful features, a Common Spatial Pattern (CSP) algorithm was applied in order to extract **4** spatial filters to apply to the data [6]. While traditionally, the CSP algorithm is used on binary classifiers to maximize the variance of one class, while minimizing the variance of the other class, the MNE library in python features a different implementation, which utilizes joint approximate diagonalization (JAD), which is equivalent to Independent Component Analysis [7]. The exact number of spatial filters was found using trial and error, with the optimal amount of filters being 4 to 8, 4 filters showed the best performance.

The CSP algorithm learned the spacial filters using the training dataset, and was then applied to both the train and test datasets, in order to keep the two datasets separate. After this, LDA was used, with shrinking applied using the Ledoit-Wolf method, which is used by default in the SciPy library, which is a way to estimate the shrinkage constant without cross-validation [8]. However, this step does not affect the result in

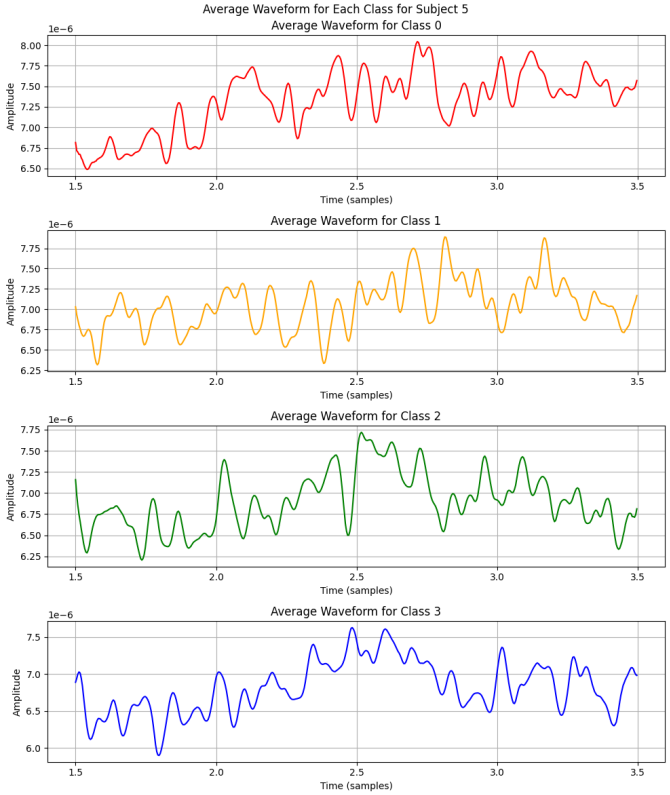

Fig. 2. Average Waveform for different classes for Subject 5

a noticeable way, since the dimensionality of the date (128) is a lot less than the number of samples (512).

Each subject was individually trained on, and results evaluated separately, then averaged.

D. CSP-SVM

We followed by applying Support Vector Machines, with the hope of improving on our results from LDA. Data preprocessing pipepine was the same as for CSP-LDA. After splitting the dataset, we transform our train and test splits with csp filters obtained from train split. Grid search for hyperparameters revealed the best performance while using 4 CSP filters and sigmoid kernel, which, in theory should effectively leverage the non-linearities present in the dataset.

In the same way as CSP-LDA, each subject was individually trained on, and results evaluated separately, then averaged.

E. EEGNet

EEGNet is a CNN-based architecture specifically tailored for classification tasks in BCI. EEGNet's convolution layers are capable of learning both spatial and temporal features on their own. This means it can learn from almost raw EEG data with minimal pre-processing. EEGNet's architecture consists of 2 blocks. The first block comprises of 2D convolutions and DepthWise convolutions. Conv2D layers are responsible for extracting the most discriminative temporal information from EEG timeline. DepthWiseConv2D applies convolution

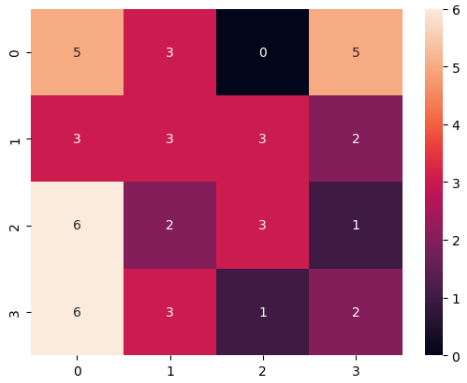

Fig. 3. LDA Confusion Matrix for Subject 10

across different EEG channels, and responsible for extracting temporal information. The dimensions of our EEGNet's Conv2D layer is (1, 512), and 512 signifies the kernel length of twice the sampling rate (256 Hz). The dimensions of DepthWiseConv2D filters are (128, 1), since our dataset contains 128 channels.

The second block does separable convolutions followed by classification in FC layer. The goal of using separable convolutions is to summarize information while decreasing the number of parameters of the model. The classification layer consists of a single dense layer followed by softmax activation function.

Our EEGNet was trained using cross validation with 4 folds, and the best performing model weights were selected across folds.

III. Results

A. CSP-LDA

The LDA classifier was not able to effectively separate the four classes, apart from a limited amount of subjects. Since there have been 10 subjects with different trained models, not all 10 confusion matrices can be shown. Instead, a collection of F1 and Accuracy of each subject's specification, and an example confusion matrix for two subjects is shown. The mean accuracy for all the subjects was 24%, and the mean F1 score was 23%, which is less than the 25% baseline accuracy by picking randomly. However, on some subjects, like subjects 1, 2, 3 and 10, the accuracy is slightly above that (Figure 5). The confusion matrices for subjects 2 (Figure 4) and 10 (Figure 3) are presented.

B. SVM

The SVM classifier was also a mixed bag. It was more performant on some, and less performant on other subjects. The detailed results are presented in the overall results (Table III) as well as Figure 6. The confusion matrices for some of the classes are also provided in Figures 7-8

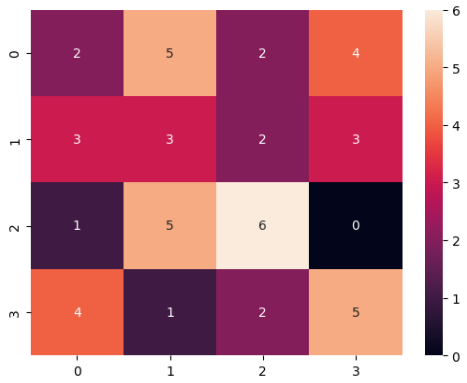

Fig. 4. LDA Confusion Matrix for Subject 2

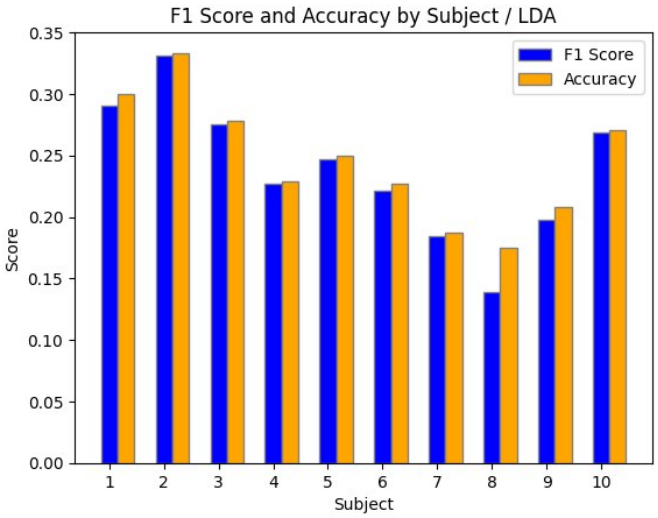

Fig. 5. F1 and Accuracy for each Subject w/LDA

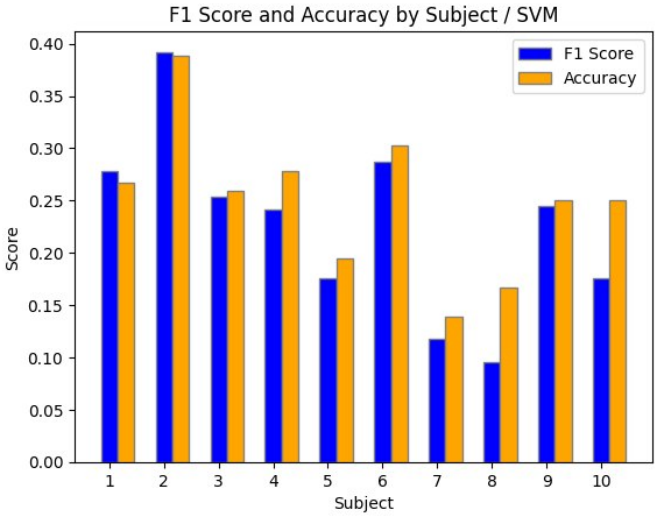

Fig. 6. F1 and Accuracy for each Subject w/SVM

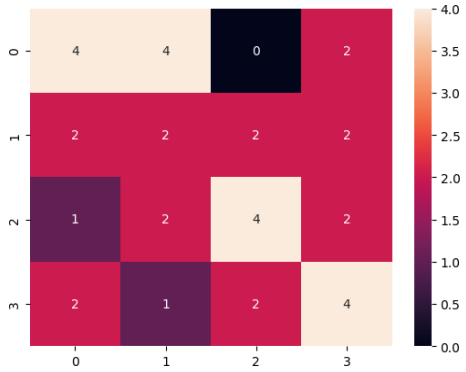

Fig. 7. SVM Confusion Matrix for Subject 2

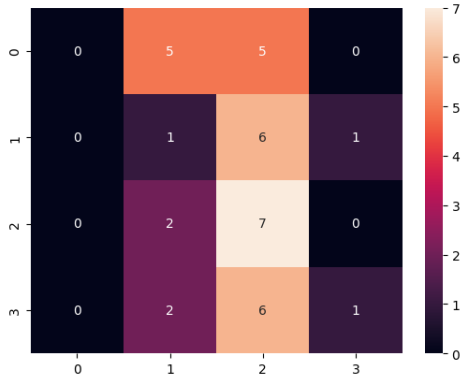

Fig. 8. SVM Confusion Matrix for Subject 10

C. EEGNet

EEGNet was the only model showing somewhat reasonable performance. 11 clearly shows that EEGNet was able to classify all subjects with accuracy higher than if classes were classified by random choice. Although the results are by no means very accurate, it can be considered success given the complexity of the task. Here we present EEGNet performance for all subjects in Figure 11, and confusion matrix for some subjects in Figure 10 and Figure 9. The results signify that classification of inner speech is possible.

Subj. No.	LDA		SVM		EEGNET	
	F1	Acc	F1	Acc	F1	Acc
1	0.29	0.30	0.28	0.27	0.33	0.32
2	0.33	0.33	0.39	0.39	0.36	0.38
3	0.27	0.28	0.25	0.26	0.47	0.47
4	0.23	0.23	0.24	0.28	0.29	0.29
5	0.25	0.25	0.18	0.19	0.30	0.31
6	0.22	0.23	0.29	0.30	0.29	0.30
7	0.18	0.19	0.12	0.14	0.41	0.42
8	0.14	0.18	0.10	0.17	0.47	0.48
9	0.20	0.21	0.25	0.25	0.31	0.31
10	0.27	0.27	0.18	0.25	0.34	0.33

TABLE III
TOTAL RESULTS FROM LDA, SVM AND EEGNET

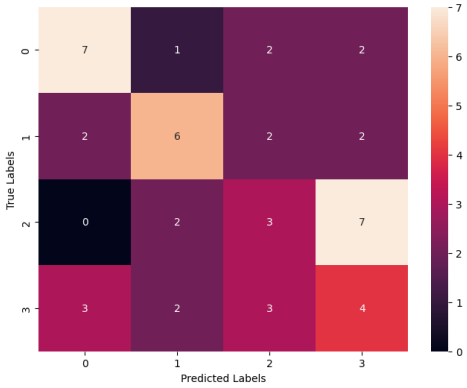

Fig. 9. EEGNet Confusion Matrix for Subject 2

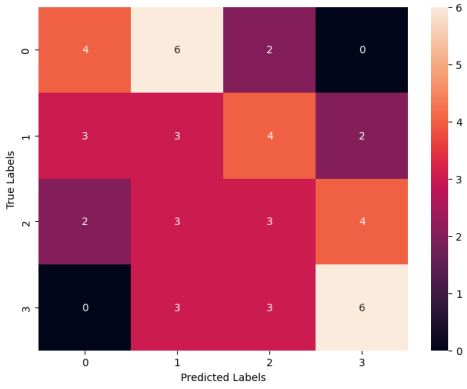

Fig. 10. EEGNet Confusion Matrix for Subject 10

IV. Discussion

Comparing the results from all three models (Figure 12), we get the image that only our EEGNet implementation was able to achieve a somewhat reliable classification result. Both the LDA and SVM approaches have proven to be, on average, worse than picking randomly out of the four labels, however, this varies from subject to subject. This dataset is one of the most detailed datasets available publicly on Inner Speech, however, the research around it has been lacking in volume. The dataset authors themselves did not publish any sort of classification methods to be used with it, and very little research has been done on that. However, whilst our results may look disappointing at first, it is important to consider that the dataset itself is quite complex. We were able to find only one research paper focusing on the dataset, and their average accuracy of 29.67%, with the average F1-score of 29.61% falls below the EEGNet implementation's average accuracy and F1-score of 35.79% and 36%, respectively.

One way to improve the overall performance of our experiment in the future could be the elimination of unnecessary channels, as 128 electrodes increases the dimensionality and the complexity of data for a task that only focuses on Inner Speech. Improving the CSP filtering and using other filters might prove useful as well, although experimentation on these has led us to nowhere throughout our experiment.

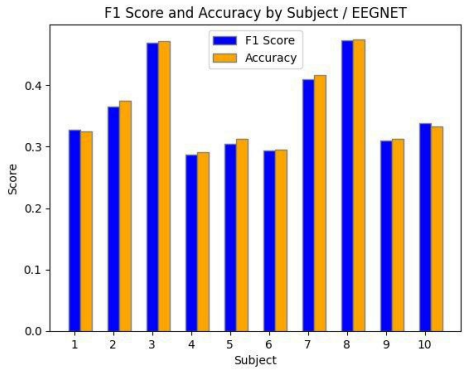

Fig. 11. F1 and Accuracy for each Subject w/EEGNet

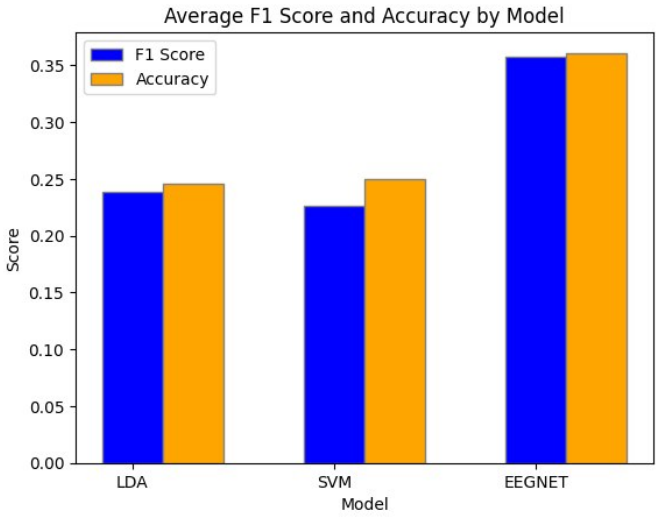

Fig. 12. Average Scores for All Models

References

[1] L. M. McCane, S. M. Heckman, D. J. McFarland, G. Townsend, J. N. Mak, E. W. Sellers, D. Zeitlin, L. M. Tenteromano, J. R. Wolpaw, and T. M. Vaughan, "P300-based brain-computer interface (bci) event-related potentials (erps): People with amyotrophic lateral sclerosis (als) vs. age-matched controls," *Clinical Neurophysiology*, vol. 126, no. 11, pp. 2124–2131, 2015.

[2] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, "Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition," *Scientific Data*, vol. 9, p. 52, Feb 2022.

[3] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLOS Biology*, vol. 10, pp. 1–13, 01 2012.

[4] T. Saltuklaroglu, A. Bowers, A. W. Harkrider, D. Casenhiser, K. J. Reilly, D. E. Jenson, and D. Thornton, "Eeg mu rhythms: Rich sources of sensorimotor information in speech processing," *Brain and Language*, vol. 187, pp. 41–61, 2018.

[5] Y.-W. Liu, *Hilbert Transform and Applications*. 04 2012.

[6] A. Jiang, J. Shang, X. Liu, Y. Tang, H. K. Kwan, and Y. Zhu, "Efficient CSP algorithm with Spatio-Temporal filtering for motor imagery classification," *IEEE Trans Neural Syst Rehabil Eng*, vol. 28, pp. 1006–1016, Mar. 2020.

[7] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1991–2000, 2008.

[8] O. Ledoit and M. Wolf, “Honey, i shrunk the sample covariance matrix,” *The Journal of Portfolio Management*, vol. 30, no. 4, pp. 110–119, 2004.