Building a Convolutional Neural Network to Predict Inner Speech

Austin Drake, Nele Felicitas Werner[a]

[a]*Bar-Ilan University, , Ramat Gan, Israel*

Abstract

Brain-computer interfaces (BCIs) facilitate direct communication between humans and computers. While most BCIs use motor imagery (MI) and event-related potential (ERP) paradigms, this study extends these by focusing on the "inner speech" paradigm using electroencephalography (EEG) signals. Inner speech is the internalised process of thinking in words by imagining one's own voice. Using EEG data from 10 participants, our goal was to develop a 2-D Convolutional Neural Network (CNN) that could predict inner thoughts across 4 different directions (up, down, left, right). Another CNN based on the EEGNet architecture and a Support Vector Machine (SVM) were used for comparison. None of the models exceeded chance accuracy (above 25%), highlighting the current challenges in accurately classifying inner speech from neural measures. The results and model descriptions presented in this study provide insights for future research in this evolving field. The whole analysis was conducted on Kaggle, a notebook with the models used is publicly available under: `https://www.kaggle.com/code/austindrake/inner-speech-decoding/input`.

Keywords: EEG, BCI, Machine Learning, Inner Speech

1. Introduction

Brain-computer interfaces (BCIs) are a cutting edge technology that enable users to control or communicate with external devices through brain activity. Primarily developed for people with severe disabilities, BCIs provide a non-muscular communication channel by translating the user's intentions into digital commands for assistive applications such as speech synthesizers, wheelchairs and neural prostheses van den Berg et al. (2021).

Various BCIs use motor-imagery (MI) or P300 paradigms, which have shown success in the past. With visual stimulation P300 event-related potential enabled the user to spell words by selectively paying attention to the intended letters (Fazel-Rezai et al., 2012). However, most of these BCI paradigms are dependent on external stimuli and they lack a fluid and intuitive paradigm that allows natural communication without the need on external stimulation and training.

To address this, inner speech recognition is emerging as an alternative communication paradigm for BCIs as in general no external stimulus is needed. Inner speech is an internalized process in which the person thinks in means of language, generally associated with auditory imagery of an inner "voice" (Nieto et al., 2022). A study conducted by Nieto et al. (2022) collected EEG data from 10 subjects for the purpose of developing methods and techniques to allow for the directional control of units. The researchers specifically explored inner speech, contrasting it with verbal/pronounced speech and internal visualization to compare diverse neural representations.

Our goal was to develop a machine learning algorithm that could detect the direction condition of each individual trial, with a focus on the data from the Inner Speech condition. As this was an exploratory project, we built and trained multiple styles of algorithms to find the model that best predicts the direction of the trial. We tested two different convolutional neural networks (CNN) and one support vector machine (SVM). Given the complexity of EEG signals, CNNs ability to automatically extract spatial and temporal representations eliminates the need for time-consuming pre-processing and feature engineering, offering potential performance advantages over traditional machine learning methods (Tibrewal et al., 2021; van den Berg et al., 2021).

2. Methods

2.1. Data Collection

The dataset provided by Nieto et al. (2022) is publicly available on Kaggle [1]. Brain activity of 10 individuals was recorded using the standard BioSemi EEG cap with 128 channels (as depicted in Fig. 1) and an additional 8 external Electrooculography (EOG) and Electromyography (EMG) channels with a sampling rate of 1024 Hz. All participants were right-handed, healthy, native Spanish speakers, and had no prior BCI experience. The subjects were set up with a monitor that would display directional stimuli. The stimuli for all three conditions were the same: a focus point would appear on screen, after 0.5 seconds a triangle pointing to one of four directions (up, down, left, or right) would appear, and after a further 0.5 seconds the focus point would return, at which time the subject would begin the task. Depending on whether the task was pronounced speech, inner speech, or visualization, the subject would either repeatedly speak the directional word (pronounced

[1] `https://www.kaggle.com/datasets/truthisneverlinear/inner-speech-recognition`

speech), imagine speaking the word (inner speech), or imagine moving the focus point in the indicated direction visually (visual condition). One trial lasted about two seconds. Further details on stimuli and data collection methods can be found the original article by Nieto et al. (2022).

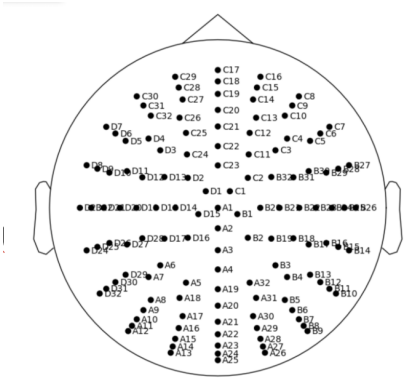

Figure 1: Location of channels (BioSemi 128 EEG Cap)

2.2. *Data Preprocessing*

For our analysis, we used the pre-processed data from the 10 participants. For preprocessing, the data was re-referenced (average of EXG1 and EXG2 located on the subjects' earlobes), downsampled to 256 Hz, high and low pass filters were applied (0.5 Hz - 100 Hz), the data was epoched (2 s per trial), and an Independent Component Analysis (ICA) to remove artifacts from the signal was performed. We used an utility script provided by the researchers to access and process the data files. After reviewing the details of the data, we decided that our first goal should be to predict the Direction label of the stimulus for a single subject. After extracting the Inner Speech trials, our input data for a single participant was structured such that there were 240 inner speech trials (60 for each stimulus (up,down,left,right), 128 EEG channels and 512 sample points per trial (2 seconds per trial, at fs = 256 Hz). Python utility scripts were publicly available via a GitHub link in the article, providing a set of functions for loading and modifying the data.

2.3. *Preliminary CNN Architecture*

A common machine learning algorithm used in EEG analysis and other time-series data is the convolutional neural network (CNN). Our initial model parameters were developed from a paper exploring epilepsy diagnosis using EEG recordings (Thomas et al., 2018). As shown in Figure 2, our model contained two convolutional layers, each with 32 kernels of length 7. Both layers were followed by maxpooling. This then connected to a fully connected layer with 3000 nodes, and a softmax output. All three layers included a ReLU activation function to allow for non-linearity.

After training our initial model, we found that it was consistently overfitting the training dataset, achieving perfect accuracy on the training samples while retaining near-chance accuracy of around 25% in the test dataset. Based on this, we determined that a simpler model with regularization was needed.

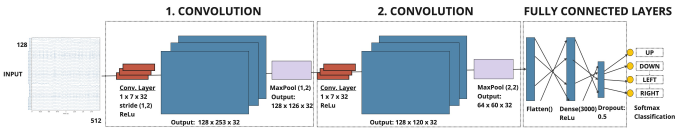

Figure 2: CNN Structure based on the paper by Thomas et al. (2018).

We experimented with fewer fully connected (FC) nodes, wider kernels and longer strides of varying scale. We also added weight decay to our Adam optimizer and Dropout (set to 0.5) to our FC layer to prevent inter-node dependency. It was observed that wider kernels inherently reduce dimensionality with zero padding, or 'valid' padding, as the output of this layer is proportional to the difference between kernel length and signal length. Additionally, we hypothesized that the patterns that distinguish different inner words might be more subtle than those used for diagnosing epilepsy, and further hypothesized that deeper layers might have an advantage in this regard. Therefore, the wide FC layer was replaced with two narrower layers. The kernels and layers determined to provide the greatest chance of success were implemented in our final CNN, outlined in Section 2.6.

2.4. *EEGNet Architecture*

The second model we implemented was EEGNet, a compact convolutional neural network (CNN) developed by Lawhern et al. (2018). EEGNet is specifically designed for different EEG paradigms within BCIs. The architecture of EEGNet, shown in Figure 6, consists of three main convolutional layers. The model first starts with a temporal convolutional layer consisting of eight kernels, each with a length of 64. The temporal nature of this layer is due to the kernel channel-depth of one, allowing the time-series data to be summarized for each channel separately. This is followed by a spatial depth-wise convolution with a channel-depth of 128 channels, a temporal dimension of one, and a depth multiplier of 2. This produces two consolidations of all channels across the 8 prior layer outputs. The last layer consists of a separable convolution with 16 kernels beginning with a spatial dimension of 1x16 and followed by a unit vector of 1x1. Spatial and separable convolution were followed by average pooling and a spatial dropout of 0.25. This processed information is then passed through a fully connected layer using softmax as an output function. The structure was derived from the original EEGNet implementation that is available on GitHub (vlawhern et al., 2022).

2.5. *Training and Evaluation*

To counteract overfitting across both CNN and EEGNet models, early stopping was integrated into the training process. Monitoring the validation loss, the training stopped if within three consecutive epochs the validation loss did not decrease. Additionally, a learning rate scheduler was defined, which dynamically adjusted the learning rate by a factor of 0.1 if the validation loss did not decrease for three consecutive epochs, with a minimum learning rate set to 0.00001.

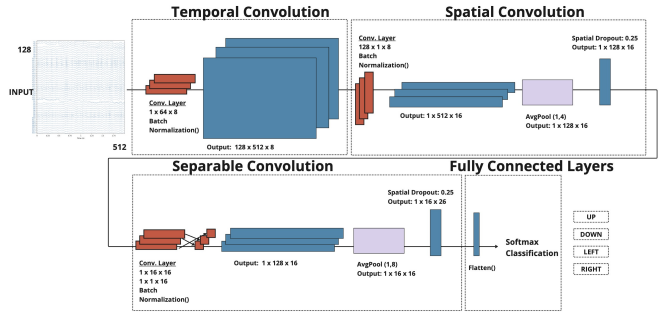

Figure 3: EEGNet structure based on the paper by Lawhern et al. (2018)

Another method to reduce overfitting is to expand the number of recordings. This can be done artificially with the use of data augmentation. While different methods exist, the method selected for this project was to add randomized values, or 'noise', to the time-series data. As can be seen in Fig. 4, this generated a second training dataset that contained signals similar to the original training data, without repeating identical recordings that could worsen the overfitting.

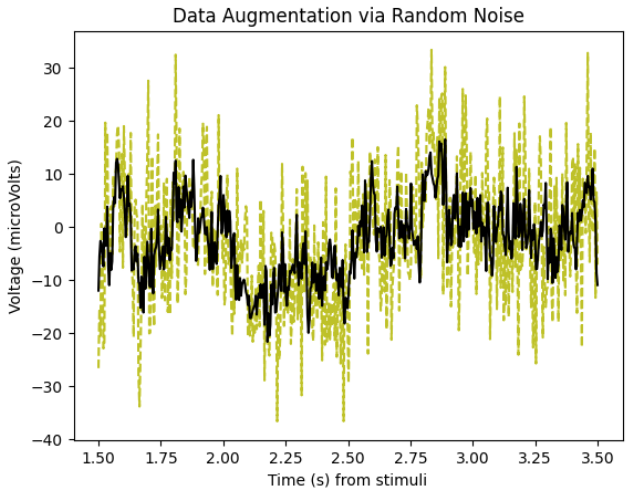

Figure 4: Data augmentation by adding noise to the original signal. Noisy signal where spikes are added is green and dotted. Original signal in black.

During training a stratified k-fold cross-validation approach was used. 5 folds were created, and the stratified method ensured an even distribution of all four classes between the training and validation sets. In general, the models were tested on each subject individually. For completeness we implemented one multi-subject model based on the CNN structure across all 10 subjects. A larger 13-wide kernel was used in the first convolutional layer with the hope that it would account for the additional complexity of the subject group. However, after discussing the challenges of variability across different subjects with our supervisor, we stopped our attempts to build a multi-subject model to focus our efforts on single subject models. Additionally, separate models were trained on the visual speech condition as well as inner and visual condition combined to increase the data set and to observe if correlated thought patterns across the two conditions could improve our models' performances. The pronounced speech condition was not included in the analysis since a higher noise ratio was expected. Accuracy was used as performance metrics across all models.

2.6. *Final CNN Model*

After comparing the results from the preliminary CNN and EEGNet models, we developed a new CNN model that aimed to build upon the principals of each model, specifically by incorporating the spatial depth-wise convolutional layer presented in the EEGNet paper (Lawhern et al., 2018). The first convolutional layer of the model has 32 kernels with a width of 31, stride 1 and a padding setting of 'valid'. We found that, for a 256Hz recording, a 31-sample window corresponded to the wavelength of an 8.3 Hz signal, allowing for better detection of low-frequency patterns. This is followed by a 1x2 max pooling layer, reducing the number of samples per recording to 241. The next layer is a depth-wise convolutional layer, which applies a spatial 128 channel-depth kernel across each of the 32 prior layer's outputs separately. A depth multiplier of 1 is used to reduce model complexity. This provides a spatial comparison across the 128 channels as mentioned in Section 2.4. The output of this depth-wise layer feeds into two FC layers, the first with 200 nodes and the second with 50 nodes. Dropout set to 0.5 is applied to both FC layers. A final softmax was applied to the output. For the convolutional and FC layers, a ReLU activation function was included. The performance of this model is compared with the EEGNet in Results Section (3).

2.7. *FFT and SVM*

An alternative to CNNs commonly used in EEG data is the support vector machine (SVM). SVMs tend to perform better on small datasets than neural networks, and take less time to train. However, data with significant dimensions can suffer in SVMs due to the sparsity of information, leading to long training times and reduced accuracy. In order to reduce the dimensions, a fast fourier transform (FFT) was applied to the data. This converts the time-series data of an EEG recording into an equivalent representation as a series of frequencies of varying magnitude. This was followed by averaging across channels and converting the resulting frequency magnitudes to decibels. Finally, a window from 4Hz to 40Hz was sliced as these frequencies encompass the frequencies ranges for theta brainwaves (4-8Hz), alpha brainwaves (8-12Hz), beta brainwaves (12-35Hz), and low gamma brainwaves (above 35Hz) (Abhang et al., 2016).

After the FFT and subsequent preprocessing, an SVM was trained over the dataset. The parameters of the SVM were optimized by using a grid search. The search found that the best regularization parameter C was 0.001, and that the best kernel was a polynomial kernel of degree 3. Cross validation was also used in training this model, however 8 folds were generated rather than the 5 folds used previously.

3. Results

3.1. CNN and EEGNet

In order to determine the effectiveness of our final CNN model, we compared the results of our final CNN model with those of the EEGNet model, as the EEGNet is a published and recognized model for EEG analysis. Our analysis has no preference for a particular outcome nor error, so general accuracy across the four classes was used to evaluate the models' performance. The mean test accuracy across all subjects for the final CNN model was 23% with a standard deviation of ±0.06, and for the EEGNet 24% with a standard deviation of ±0.03. Table 2 summarizes the results of the cross validation. Considering the odds of 1 in 4 evenly distributed outcomes is 25%, we conclude that both models fail entirely to predict the directionality of the trial. While the final CNN for Subject 8 has an accuracy of 32% which is higher than the mean by 1.5 standard deviations, this is unlikely to indicate any meaningful pattern recognition within the model. The test size for each subject was around 24 trials. At this scale, 32% accuracy is equivalent to one or two more correct predictions than the average. Comparing the training and validation losses of all subjects across the two models, we again see both models fail to improve. While the EEGNet models demonstrated a more consistent downward trend in training loss than the final CNN, the EEGNet models tended to start with higher initial loss as well (Fig. 5 and Fig. 6). The training accuracy is for both CNN and EEGNet above 0.3 (CNN: 0.30 ± 0.03 and EEGNet: 0.33 ± 0.05), however the validation accuracy is for both below chance (CNN: 0.24 ± 0.1, EEGNet: 0.19 ± 0.08). Further details can be found in the Appendix A. Overall, it is clear that both the final CNN models and EEGNet models showed equally low predictive capacity for all 10 subjects.

Training the model on the visual condition and visual and inner combined lead to similar results. Further details can be found in Appendix A.

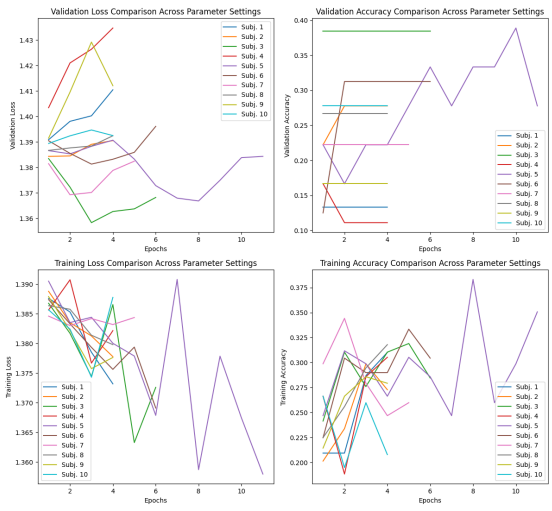

Figure 5: CNN performance on training and validation data for inner speech condition. Accuracy and loss.

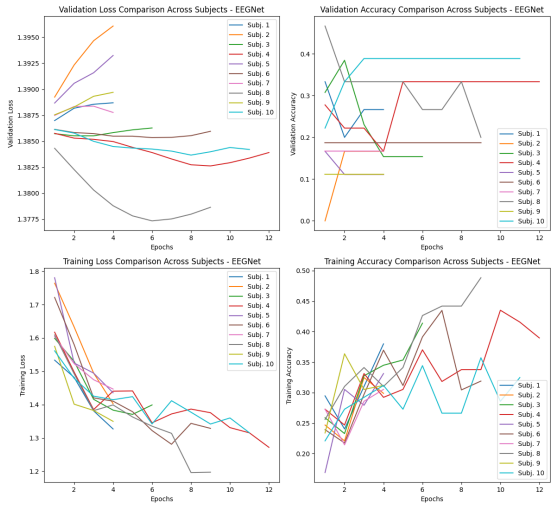

Figure 6: EEGNet performance on training and validation data for inner speech condition. Accuracy and loss.

	mean ± std			
Subject	Training Loss	Training Acc.	Val. Loss	Val. Acc.
CNN	1.38 ± 0.01	0.30 ± 0.03	1.40 ± 0.01	0.24 ± 0.1
EEGNet	1.41 ± 0.08	0.33 ± 0.05	1.39 ± 0.01	0.19 ± 0.08

Table 1: Mean ± std of training and validation loss and accuracy of both models

Test Accuracy		
Subject	CNN	EEGNet
1	0.30	0.25
2	0.25	0.25
3	0.13	0.25
4	0.14	0.27
5	0.20	0.21
6	0.23	0.18
7	0.19	0.27
8	0.32	0.22
9	0.25	0.29
10	0.29	0.21
mean ± std	**0.23 ± 0.06**	**0.24 ± 0.03**

Table 2: Test Accuracy of the CNN and EEGNet model for each participant for the Inner Speech condition.

3.2. Multi-Subject Model

Training the Multi-Subject Model across all participants did not reveal significant differences compared to the Individual Subject CNN models. A graphic on the training accuracy and loss can be found in the Appendix A. The accuracy of the test set was 0.25, meaning it was exactly at chance.

3.3. SVM on FFT Data

Table 4 summarizes the results on the SVM performance on the frequency extracted data using FFT for inner, visual and combined inner and visual conditions. After applying grid search the model with the optimal parameters was used to predict the data on the test set. Overall, the highest accuracy across subjects is reported for the inner speech condition with a slightly above chance accuracy of 0.266 ± 0.06 (visual mean accuracy: 0.23 ± 0.02, combined: 0.215 ± 0.02). The highest

performance can be observed for Subject 1 in the Inner Speech condition with 0.35. For the visual and the combined data set the highest accuracies were below 0.3 (visual: Subject 6 and 7 with 0.27, combined: Subject 2 with 0.25).

Subj.	Precision	Recall	F1-Score	Acc.
1	0.36	0.38	0.35	0.35
2	0.31	0.29	0.29	0.29
3	0.17	0.21	0.18	0.17
4	0.29	0.28	0.27	0.27
5	0.36	0.33	0.33	0.33
6	0.24	0.25	0.24	0.27
7	0.29	0.31	0.30	0.31
8	0.17	0.15	0.15	0.15
9	0.26	0.27	0.25	0.27
10	0.29	0.27	0.24	0.25
mean±std	**0.27±0.06**	**0.27±0.06**	**0.26±0.06**	**0.27±0.06**

Table 3: Performance of the SVM model with best parameters (after grid search) for each participant on the Inner Condition data.

Test Accuracy			
Subject	Inner	Visual	Inner and Visual
1	0.35	0.25	0.21
2	0.29	0.21	0.25
3	0.17	0.23	0.19
4	0.27	0.25	0.24
5	0.33	0.19	0.21
6	0.27	0.27	0.22
7	0.31	0.27	0.21
8	0.15	0.25	0.15
9	0.27	0.21	0.18
10	0.25	0.19	0.23
mean ± std	**0.266 ± 0.06**	**0.23 ± 0.02**	**0.215 ± 0.02**

Table 4: Test Accuracy of the SVM model with best parameters (after grid search) for each participant on the inner, visual and both conditions data.

4. Discussion

This project implemented two different convolutional neural networks to predict inner speech using an EEG dataset from Nieto et al. (2022): A simple CNN, originally inspired by Thomas et al. (2018) and then fine-tuned, and an EEGNet model, a compact CNN model developed by Lawhern et al. (2018) for various EEG-based classification tasks. After training versions of each model for all subjects individually across the Inner Speech condition data, all models failed to predict the Direction label of a trial at accuracies significantly greater than random chance. This was repeated with the Visualization subset and a combination of the two subsets, however the results remained the same. Multiple modifications and training changes were implemented such as stratified k-fold cross validation and model regularization. Still, all models showed accuracies that could be considered equivalent to random chance. A simpler SVM was also implemented in the hopes of better utilizing the limited dataset. When the SVM was trained on the frequencies extracted by FFT, a slightly higher accuracy was reported. However, most models still showed no significant predictive capabilities over the data.

While some studies have demonstrated the success of CNN models in BCI classification tasks, such as epilepsy classification, achieving an accuracy of 83.86% (Thomas et al., 2018) (for a review of other CNN applications in BCI, see Craik et al. (2019)), the inner speech dataset used in our study presents unique challenges. Previous work by van den Berg et al. (2021) using a CNN based on the EEGNet architecture on the same dataset achieved an accuracy of 29.67%. Unfortunately, our attempts to replicate their results were unsuccessful.

One of the greatest limiting factors that this project faced was the amount of data available. Across each subject, the expected number of Inner Speech condition trials was 240 samples. A dataset of this size has significant difficulty training complex models such as neural networks, as the networks are capable of learning perfect accuracy across the training dataset rather than learning underlying patterns in the data that would otherwise allow a model to effectively analyze and predict other trials. Although attempts were made to augment the data and cut down the model complexity, all models showed signs of overfitting with little to no improvement when tested on validation or test trials.

One suggestion that the original authors of EEGNet provide is the use of transfer learning to initialize weights for new models on individual subjects. However, a total dataset of 10 subjects is itself a small dataset for pretraining a universal EEGNet model for the purposes of transfer learning; the multisubject models themselves already struggled to avoid rapid overfitting to the data. Therefore, no transfer learning was applied and instead all EEGNet models were trained with random weight initialization. If an EEGNet model trained on a larger dataset of inner speech and/or directional thought conditions becomes available in the future, it would be advisable to apply transfer learning in this case.

Another potential limitation lies in the inherent structure of the data. EEG is known to have sparse spatial resolution. Visual inspection of the raw EEG data revealed no identifiable pattern during the inner speech trial (see Appendix A for the raw EEG signal across channels and in Nieto et al. (2022) Figure 5). In contrast, other BCIs paradigms using EEG data have successfully classified patterns such as the P300 component (Li et al., 2020), epilepsy (Thomas et al., 2018) and schizophrenia (Khare et al., 2021) with high accuracy using CNN models, which also were clearly visible by visual inspection. The generally low signal quality is also evident in the topoplot for a single trial provided in the Appendix A. Improving data quality for BCIs, particularly in the context of inner speech, requires further investigation.

One note about training models on EEG data that may be of interest to the reader is the consideration of units. In our earliest models, our training loss and accuracy failed to improve after 20 epochs, regardless of changes to the learning parameters such as batch size and learning rate. However, once our signals were converted from volts to micro-volts as done in the original analysis by the experimenters as well as other published EEG classifiers (Bayram et al., 2013), the models achieved im-

provement in training loss and accuracy. A similar phenomenon occurred when training the SVM model, and a change to decibel representation of the magnitudes produced better training results. This demonstrated the importance of understanding the scale of data before attempting to integrate it into machine learning.

A point of consideration with the selected paradigm is the choice of specific words (related to movement direction) and visual cues used for the inner speech task. The semantic information of these words may have elicited brain activity associated with functions beyond the intended task, such as motor imagery and visual processing. Evidence from a previous study by Nalborczyk et al. (2020); van den Berg et al. (2021) using surface electromyography (EMG) signals suggests that inner speech production is not solely controlled by the abstraction of linguistic representations, but may involve the simulation of the motor system, depending on the nature of the task.

Previous neuroscience research has provided compelling evidence indicating that inner speech activates specific brain regions linked to language comprehension and production. These areas encompass the temporal, frontal, and sensorimotor regions, with a notable prevalence in the left hemisphere Stephane et al. (2021); van den Berg et al. (2021). While we have considered all channels, further analysis could have used principal component analysis (PCA) to reduce dimensionality by including only the most informative channels in the analysis. To confirm the aforementioned point, one could check whether channels located in the left hemisphere in the temporal and frontal regions in particular were included. As higher accuracy was reported for SVM on FFT data, PCA could also be applied to the frequency information after FFT and not only to the preprocessed data.

5. Summary and Conclusion

The goal of this project was to explore the potential of classifying inner speech from EEG signals using a machine learning approach. Training a CNN model on a EEG dataset, we aimed to classify four directional categories within the paradigm of inner speech. The results showed that our models were not able to predict these categories with an accuracy greater than chance guessing. Future works should employ larger datasets or alternatively use data augmentation techniques that could improve robustness of an EEG classifier.

References

Abhang, P.A., Gawali, B.W., Mehrotra, S.C., 2016. Chapter 2 - technological basics of eeg recording and operation of apparatus, in: Abhang, P.A., Gawali, B.W., Mehrotra, S.C. (Eds.), Introduction to EEG- and Speech-Based Emotion Recognition. Academic Press, pp. 19–50. URL: https://www.sciencedirect.com/science/article/pii/B9780128044902000026, doi:https://doi.org/10.1016/B978-0-12-804490-2.00002-6.

Bayram, K., Kizrak, M.A., Bolat, B., 2013. Classification of eeg signals by using support vector machines. doi:10.1109/INISTA.2013.6577636.

van den Berg, B., van Donkelaar, S., Alimardani, M., 2021. Inner speech classification using eeg signals: A deep learning approach, in: 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), IEEE. pp. 1–4.

Craik, A., He, Y., Contreras-Vidal, J.L., 2019. Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering 16, 031001.

Fazel-Rezai, R., Allison, B.Z., Guger, C., Sellers, E.W., Kleih, S.C., Kübler, A., 2012. P300 brain computer interface: current challenges and emerging trends. Frontiers in neuroengineering , 14.

Khare, S.K., Bajaj, V., Acharya, U.R., 2021. Spwvd-cnn for automated detection of schizophrenia patients using eeg signals. IEEE Transactions on Instrumentation and Measurement 70, 1–9.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of Neural Engineering 15, 056013. URL: http://dx.doi.org/10.1088/1741-2552/aace8c, doi:10.1088/1741-2552/aace8c.

Li, F., Li, X., Wang, F., Zhang, D., Xia, Y., He, F., 2020. A novel p300 classification algorithm based on a principal component analysis-convolutional neural network. Applied Sciences 10. URL: https://www.mdpi.com/2076-3417/10/4/1546, doi:10.3390/app10041546.

Nalborczyk, L., Grandchamp, R., Koster, E.H., Perrone-Bertolotti, M., Lœvenbruck, H., 2020. Can we decode phonetic features in inner speech using surface electromyography? PloS one 15, e0233282.

Nieto, N., Peterson, V., Rufiner, H.L., Kamienkowski, J.E., Spies, R., 2022. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. Scientific Data 9, 52.

Stephane, M., Dzemidzic, M., Yoon, G., 2021. Keeping the inner voice inside the head, a pilot fmri study. Brain and Behavior 11, e02042.

Thomas, J., Comoretto, L., Jin, J., Dauwels, J., Cash, S.S., Westover, M.B., 2018. Eeg classification via convolutional neural network-based interictal epileptiform event detection, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3148–3151. doi:10.1109/EMBC.2018.8512930.

Tibrewal, N., Leeuwis, N., Alimardani, M., 2021. The promise of deep learning for bcis: Classification of motor imagery eeg using convolutional neural network. bioRxiv , 2021–06.

vlawhern, robintibor, Gramfort, A., 2022. vlawhern/arl-eegmodels. URL: https://github.com/vlawhern/arl-eegmodels.

Appendix A. Appendix

CNN				
Subject	Training Loss	Training Acc.	Val. Loss	Val. Acc.
1	1.36	0.28	1.38	0.5
2	1.38	0.29	1.39	0.30
3	1.38	0.27	1.39	0.28
4	1.37	0.35	1.41	0.15
5	1.38	0.29	1.39	0.15
6	1.38	0.28	1.39	0.28
7	1.40	0.29	1.40	0.30
8	1.38	0.31	1.41	0.18
9	1.36	0.30	1.41	0.10
10	1.38	0.37	1.39	0.20
mean ± std	**1.38 ± 0.01**	**0.30 ± 0.03**	**1.40 ± 0.01**	**0.24 ± 0.11**

Table A.5: Training and validation loss and accuracy of the CNN model across all subjects.

EEGNet				
Subject	Training Loss	Training Acc.	Val. Loss	Val. Acc.
1	1.48	0.29	1.39	0.25
2	1.29	0.37	1.39	0.10
3	1.35	0.36	1.40	0.0
4	1.49	0.25	1.39	0.25
5	1.38	0.31	1.40	0.15
6	1.35	0.41	1.38	0.27
7	1.37	0.31	1.38	0.25
8	1.47	0.31	1.39	0.25
9	1.38	0.37	1.39	0.20
10	1.54	0.27	1.39	0.15
mean ± std	**1.41 ± 0.08**	**0.33 ± 0.05**	**1.39 ± 0.01**	**0.19 ± 0.08**

Table A.6: Training and validation loss and accuracy of the EEGNet model across all subjects.

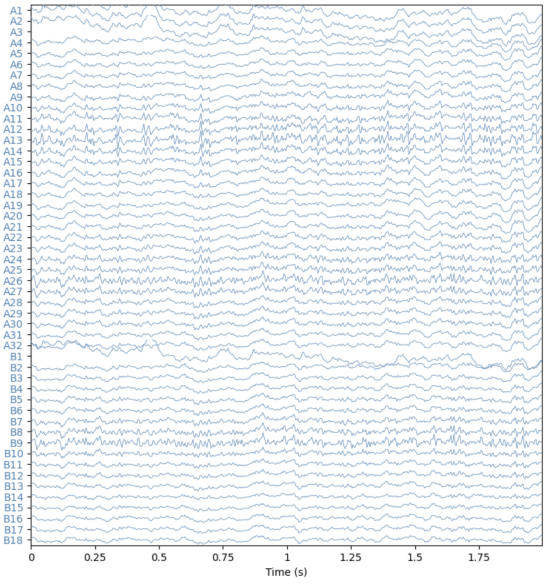

Figure A.7: Raw EEG signal. One trial of one subject. First 50 channels.

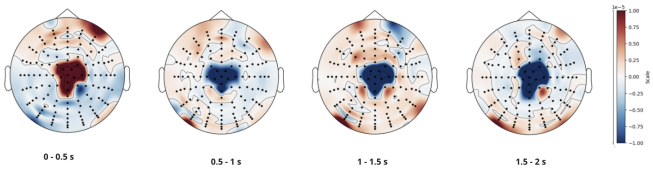

Figure A.8: Topoplot for one inner speech trial of one subject

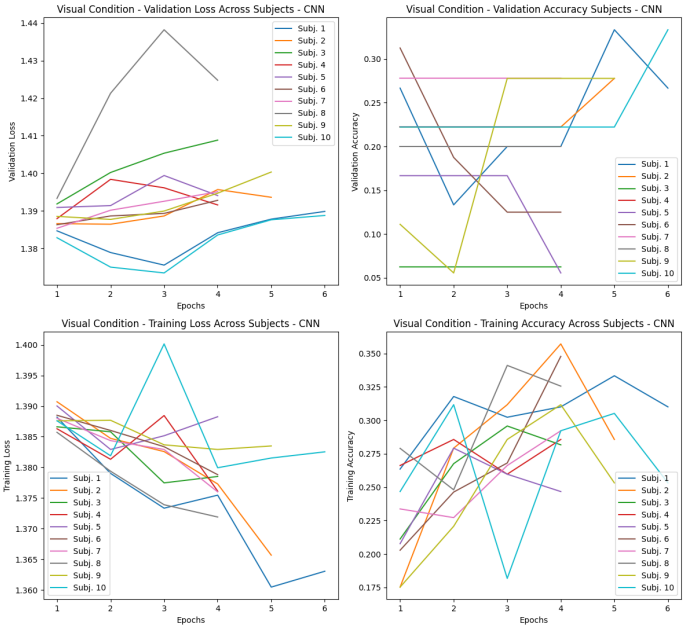

Figure A.9: CNN performance on training and validation data for visual condition. Accuracy and loss.

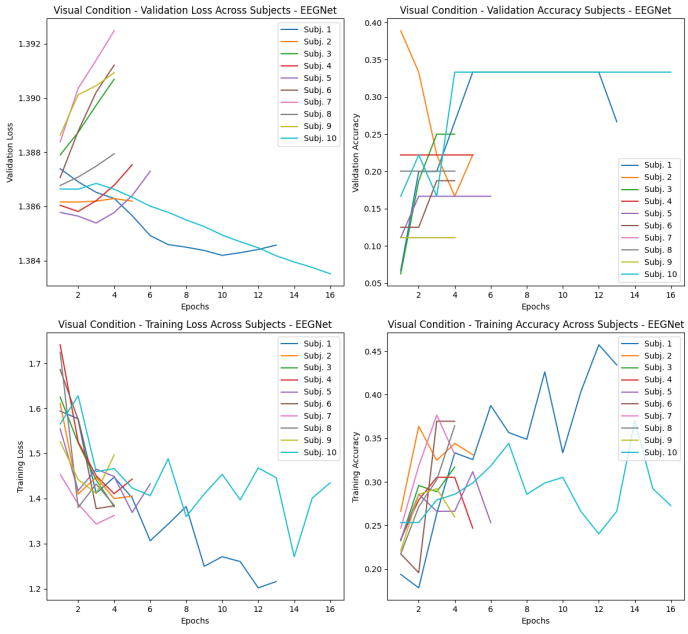

Figure A.10: EEGNet performance on training and validation data for visual condition. Accuracy and loss.

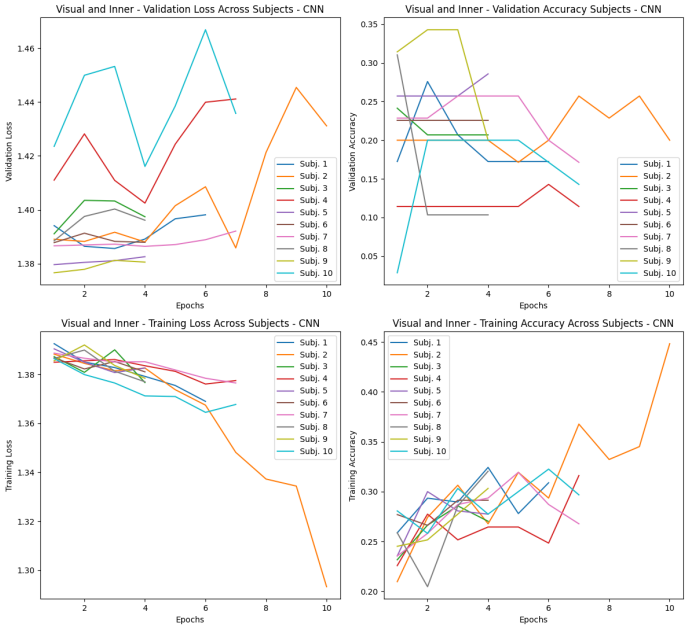

Figure A.11: CNN performance on training and validation data for visual and inner speech condition. Accuracy and loss.

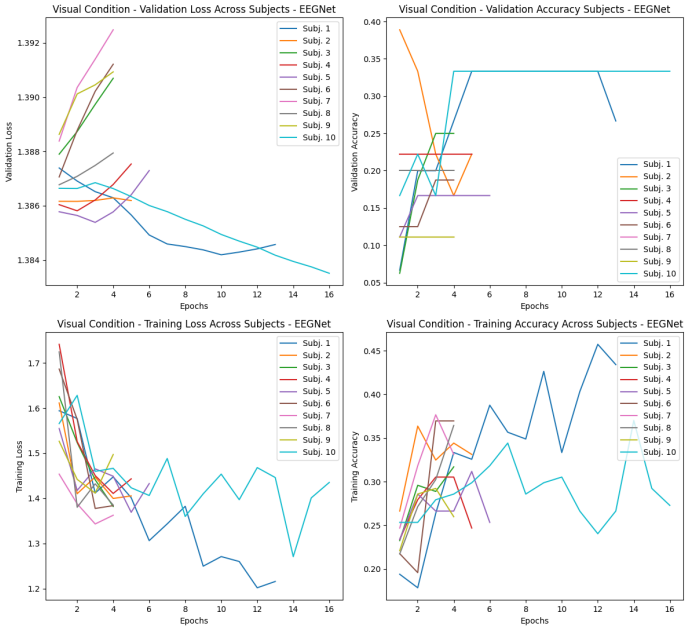

Figure A.12: EEGNet performance on training and validation data for visual and inner speech condition. Accuracy and loss.