Mindreading using EEG data: Decoding visually perceived and imagined numbers in the brain

Pjotr Straathof, s4339630,
p.j.straathof@student.rug.nl

Abstract: Recent advances in brain-computer interfaces (BCIs) have shown great potential in improving the lives of people with limited capabilities such as paraplegics or people suffering from locked-in syndrome (LIS). This study was done to further the development of decoding electroencephalogram (EEG) data by researching the difference in accuracy of a Recurrent Neural Network (RNN) trained on two different datasets. One dataset consisted of EEG data of participants imagining a number, whereas the other consisted of EEG data of participants visually perceiving a number. By comparing the results, I aimed get a decisive difference that helps pave the way for more efficient EEG decoding by creating a better understanding of it. I used the consumer-grade MUSE 2 headband, as it had been proven to achieve high results with only four electrodes and could be used at home by anyone. After training, the classifiers were evaluated using their accuracy, but neither received an accuracy score significantly above chance level. Therefore, the results of this study remained inconclusive. Further research should be done using the EPOC headset with 15 electrodes and a Convolutional Neural Network (CNN), as these have been proven to achieve better results.

1 Introduction

The field of Brain-computer Interfaces (BCIs) has seen a lot of advancements in recent years (Janapati et al. (2022)). These advancements range from the decoding of brain signals into motor movements, to the usage of BCIs in video games (Kerous et al. (2018)).
These advancements have a broad range of impact. In fields such as in the treatment of stroke survivors, it's not only being used as a tool for rehabilitation, but also as a tool to completely bypass certain damaged parts of the brain to regain control of certain motor functions (Robinson et al. (2021)). This technology has already shown great possibilities in bettering the quality of life of people with cognitive impairments, and its development is still ongoing.
Non-invasive techniques such as electroencephalograms (EEG) have become popular with researchers in this field due to its simpleness, mobility and temporal resolution. In recent year, new EEG devices have come out that are not only usable to researchers, but can also be used at home. One such consumer-grade device is the MUSE 2 headband. This headband is small in size, relatively cheap and can be set up within a matter of minutes, without having to know any in-depth knowledge of EEG signal processing. Further development using such devices may prove fruitful in the future, as regular consumers could be able to buy and use these at home. Aside from that, they also make further development and research more easily accessible (Krigolson et al. (2017)). Making use of its portability and relative ease of use, researchers can set up faster and more efficient experiments while having to give up little in terms of quality of data.

In order to add to this ongoing research, I decided that I wanted to investigate the classification of mental imagery using EEG data. Using EEG data to decode brain activity has already proven to be highly successful for Motor Imagery (MI) tasks. In their paper, Tarahi et al. (2024) investigate the usage of a method that employs a Convolutional Neural Network (CNN) to declassify imagined motor movements. The results showed an average classification accuracy of 87,3% and 86,29% on the BCI

Competition IV 2a and BCI Competition IV 2b datasets respectively.

Both datasets make use of EEG data, however, this data was captured using a full EEG set. This means that the BCI competition IV 2a dataset contained 22 EEG channels and three EOG channels. Such full EEG sets are currently too expensive and complicated to be used outside of a lab. However, Garcia-Moreno et al. (2020) proved that high accuracy in MI tasks can also be achieved using the MUSE headband. They collected data from four participants, after which they used a combination of a CNN and an LSTM in order to decode their movements based on the EEG data. Their results captured a validation accuracy of 96.5%.

While classifiers are scoring high in accuracy for MI tasks, they are still in their infancy when it comes to tasks such as decoding imagined speech. A recent study by Lee & Lee (2022) explored the use of a Deep Neural Network (DNN) for classifying the words 'in' and 'cooperate' from the ASU imagined speech dataset. With an average accuracy of 71.8 ± 8.6%, it achieved state-of-the-art results.

Another study that works with imagined speech was done by Nguyen et al. (2017). This study tried to classify between three and two imagined words. The results reached a maximum of 70% and 95% respectively.

While these tasks are celebrated as successful, when contrasting them against the MI classification tasks, it becomes clear that there's a great difference in capabilities. In order to further further develop the decoding of visual imagery, more research and understanding is necessitated.

Further research into visual imagery comes from Alazrai et al. (2020). They conducted 4 different experiments where participants were shown and then had to imagine objects of four different categories: nature (fruits and animals), decimal digits, the English alphabet (in capital letters), and arrow shapes (arrows with different colors and orientations). They used the Choi-Williams time-frequency distribution to analyze up to 15 EEG channels in the joint time-frequency domain. Their method proved successful, as the average decoding accuracies went up to 96.67% for the nature category, 93.64% for the decimal digits, 88.95% for the alphabet category, and 92.68% for the arrow shapes.

They then followed this up with the same experimental setup, but focused on a better approach of decoding the visually imagined digits and letters (Alazrai et al. (2022)). The processing of the EEG data was experimented upon. Here, they used the Choi-Williams time-frequency distribution again, this time employing 16 EEG channels. The second phase consisted of a novel deep learning (DL) approach in order to classify the data accordingly. Using this approach, they outperformed all previous studies by achieving an average standard deviation accuracy of 95.47 ± 2.3%.

Both studies were done using an EPOC headset with 15 channels. However, with an eye on personal consumer usage for future applications, it is quite expensive. Therefore, it would be beneficial to explore imagery using a headband that's cheaper and easier to use, much like the MUSE 2 headband. This was done by Mahapatra & Bhuyan (2023). In their study, they made use of the MindBigData open-access database. Using the MUSE data of this dataset and a discrete wavelet transformation (DWT), they employed a multilayer bidirectional long short-term memory (LSTM) recurrent neural network (RNN). This resulted in an accuracy of 96.18%. However, this dataset consists of one participant, whose data has been recorded for two years. This might have resulted in the classifier being successful for this specific person's EEG data, but it might be different for others. Therefore, it's worth finding out if this result can be reproduced over a variety of participants.

These studies focus mainly on the visual imagery, however, they don't seem to explore the difference in visual perception and visual imagery. Dijkstra et al. (2019) explain in their paper that there is a lot of overlap between imagery and perception in high-level visual areas in the brain. They generate similar neural representations for the same content in occipital, parietal, and frontal brain areas. That is not to say there is no difference, as they found that early bottom-up processing that occurs in perception is absent during imagery.

By exploring this difference and further the development of BCIs, I aim to train a classifier on two different datasets and study if a classifier is better at classifying imagined or visually perceived digits. In their paper, Koenig-Robert & Pearson (2021) hypothesize that the difference between imagining and visually perceiving boils down to imagery just being a weaker version of perception. Therefore, I hypothesize that the classifier trained on the visually perceived data will outperform the classifier that's trained on the imagery data.

2 Method

In order to study the difference in decoding imagery and visual stimuli, I was going to need two different datasets. Both datasets were limited to six digits going up from one to six. Whereas one dataset contained EEG data from imagery numbers, the other dataset contained EEG data of visually perceived numbers. Both datasets were recorded equally over multiple participants to prevent creating participant dependent classifiers, after which the data was processed following the steps Mahaptra et al. (2023) laid out.

2.1 Participants

In order to create the datasets, an experiment was conducted using 20 people that volunteered as participants. This group of participants consisted of 9 males and 11 females. All of the participants signed a consent form before the start of the experiment, explaining what the experiment entailed and that the collected data would become anonymous. They were also asked to fill in any neurological condition that could influence the results, such as epilepsy, however all participants claimed to have none.

2.2 Experiment

In order to create a dataset of imagined and visually perceived numbers, there had to be a range of numbers that could be used. This needed to be limited due to the current capabilities in classification based on EEG data. The range of numbers from one to six was chosen in order to be able to share the collected data with colleagues that were conducting a similar experiment with dice.

The experiment itself was conducted in a quiet room where the participant had to sit in front of a laptop. First, the participants had to imagine a number within the given range for sixty consecutive trials. During this imagery phase, participants were first instructed to think of a number. They then had to keep thinking of this number for five seconds, during which a fixation dot was shown. After these five seconds, they had to press the number they had just thought of. When they had filled in the number, they were asked to think of a new number that was different than the one before.

The second phase consisted of the numbers being shown to participants. The participants were instructed to just focus on looking and reading the number. All numbers were shown ten different times in a randomized order. The numbers were shown for five seconds per trial, followed by a 2.5 second break to establish a new baseline. The length of the break was established after participants in the trial experiment indicated that their concentration was being severely impacted by the length of the second phase. An overview of both phases can be seen in Figure 1.

After the second phase, there was a break of five minutes. Following the break, the two phases were repeated, this time consisting of 30 trials each, resulting in 90 trials per condition in total.

In order to further minimize the effects that a lack of concentration might cause, occasional questions were asked. To do this, a researcher stayed in the room. Every ten trials, the participant would be shown a screen that told them the researcher was going to ask them a question. Neither the questions nor the answers were recorded, since the information was of no importance to the study. They were also made easy to answer, asking participants about things such as what they had had for breakfast that day. This was done so they wouldn't think too long about these questions afterwards, as this could potentially influence the results.

2.3 EEG data

The EEG data was collected using an InteraXon MUSE 2 headband at a sampling rate of 256 Hz for 4 channels (TP9, AF7, AF8 and TP10). The channels were located on the frontal and temporal lobe regions, as can be seen in Figure 2. The data was then recorded using Muselsl and Labrecorder. When technical problems interrupted the recordings, a researcher would reset the headband, after which the participant could resume where they left off. Using the timestamps on the start of the sessions, the labels could later be stitched together with the correct markers. If the timestamps could

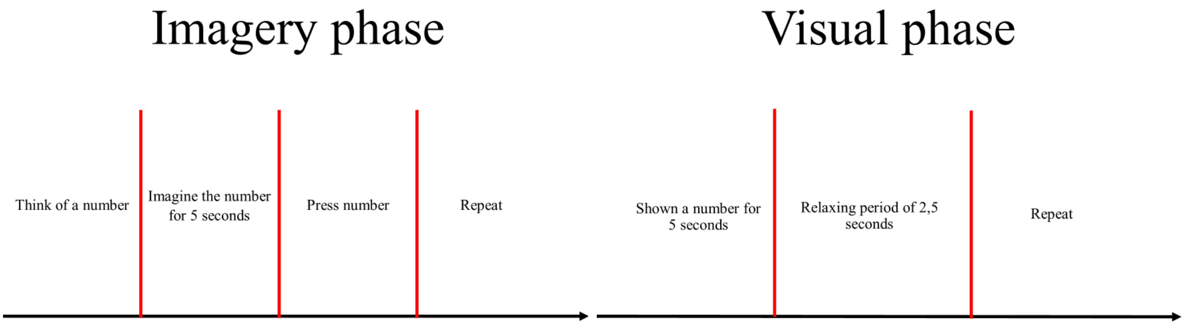
Imagery phase
Think of a number
Imagine the number for 5 seconds
Press number
Repeat
Visual phase
Shown a number for 5 seconds
Relaxing period of 2,5 seconds
Repeat

Figure 1: The experiment design of the two different phases

not be matched, that data was then discarded.

After aligning the data and checking its usability, a bandpass filter of 1-40 Hz was applied. The data was then inspected by looking at the power spectral density (PSD), in order to check for any abnormalities. If none were detected, epochs were created based on the event markers with a baseline of [0, -0.2] seconds. An example of a usable PSD is shown in Figure 3. The epochs were then saved, together with their labels, in a csv file so that they could be loaded into the classifier.

2.4 Classifier

After extracting all the relevant data, the epochs were loaded into one big dataset. Then, the datasets were checked for their lengths. Because the aim of the study is to find out whether a classifier is better at decoding visually perceived digits, both datasets were cut to be the same length. In order to check if the amount of different numbers was roughly equal, the spread was inspected as seen in Figure 4.

Although most numbers were represented roughly the same amount of times, the number '6' was shown a bit less. Because of the relatively small dataset, the decision was made to still use all the available data from the imagined numbers.

In order to obtain high results, using Python 3 and Keras, the preprocessing steps laid out in the paper by Mahapatra et al. (2023) were followed. First, the datasets were shuffled and the labels put in a separate dataset. Then, a DWT was applied in order to get rid of extra noise that could influence the results. The Daubechies-4 wavelet was used, as according to Mahapatra et al. (2023), this has been proven to be very effective in feature extraction for the classification of EEG signals. After that, the wavelet was decomposed by an order of 3 and the universal thresholding technique was applied.The formula for this threshold is as follows:

$$\lambda = \sigma\sqrt{2ln(N)} \tag{1}$$

Here, σ is the average variance of the noise whereas N is the signal length. σ can be calculated using the following equation:

$$\sigma = \frac{Median(|W_{1},_{K}|)}{0.6745} \tag{2}$$

In this equation, $W_{1},_{K}$ represents all the scale 1 coefficients.

Using the thresholded values, small components that were assumed to be noise were eliminated. This resulted in signals with a slight reduction of noise, as can be seen in Figure 5. The reconstructed data was then normalized using z-normalization and split into training-, validation- and test data.

Mahapatra et al. (2023) chose to create a bidirectional LSTM, in order to capture the temporal values of the EEG signals while also preventing both the vanishing and exploding gradient issues. They created three bidirectional layers and one dense layer. I followed this approach, but adjusted the parameters to fit it more appropriately to the data collected from my experiment. The first bidirectional layer had 440 units, the second one

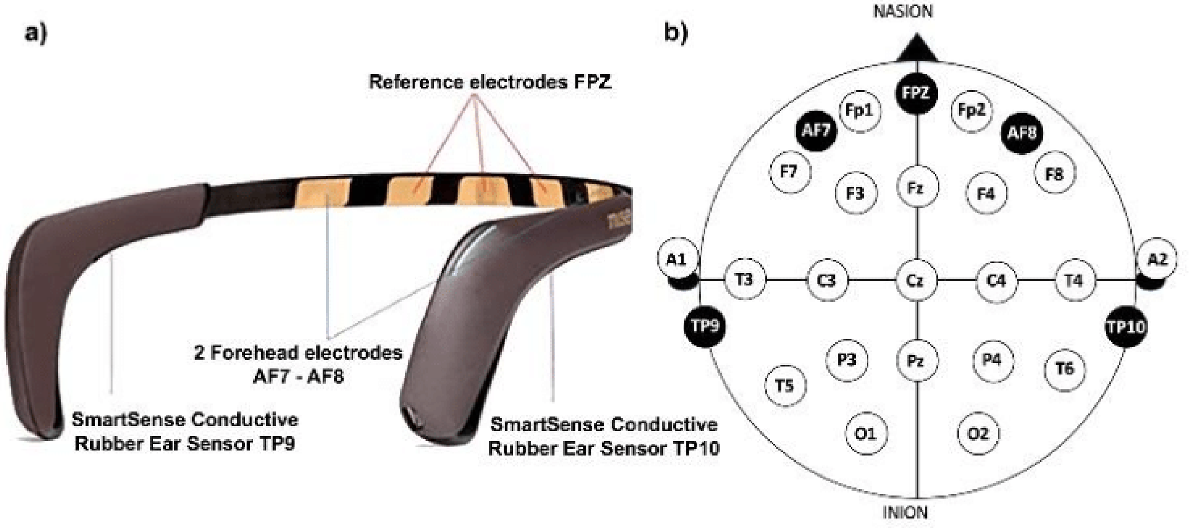

Figure 2: a) Overview of the MUSE headband by Mansi et al. (2021).

b) The relevant electrodes measured by the headband

220 and the third one 110, with a dense layer of ten units and a softmax activation function. Since the dataset used for this study contained a lot less input data, the size of the layers had to be adjusted. The dense layer was lowered to a size of six, and the layer size of the bidirectional LSTM layers were decided by a hyperparameter search using keras' Randomsearch. The first layer had a range of 200 to 500, the second layer had a range of 100 to 300, and the third layer had a range of 50 to 150.

I implemented an Adam optimizer with a cross-entropy loss function, and performed a hyperparameter search for the batch size and the learning rate as well, settling for a batch size of 32 and a learning rate of 0.002. The final hyperparameters can be seen in Table 1.

The data was split up into training-, validation and test data, with an 80-10-10 split respectively. The training- and validation data were used for the hyperparameter search, whereas the test data was used to evaluate the performance of the models. After training the models, their results were evaluated and compared. In order to compare the models however, there needed to be assurance that their performance isn't chance based. Using the theoretical approach derived in the paper by Müller-Putz et al. (2008), the model performances were assessed for their level of chance.

		Values	
	Classifier	Visual	Imagined
Bidirectional layer	Number of layers	3	3
Layer 1	Layer size	300	200
Layer 1	Dropout rate	0.3	0.4
Layer 2	Layer size	100	300
Layer 2	Dropout rate	0.3	0.5
Layer 3	Layer size	70	150
Layer 3	Dropout rate	0.3	0.5
Classification	Dense layer	6	
	Activation	Softmax	
Training	Optimization	Adam	
	Loss function	Cross-entropy	
	Batch size	32	
	Learning rate	0.002	

Table 1: Table specifying the hyperparametes

3 Results

Using the training and validation data in the hyperparameter search, the best model was chosen and evaluated on the test data using accuracy and the F1-score as metrics.

The model that was trained on the data of the imagery condition had an accuracy of 22.35 and an

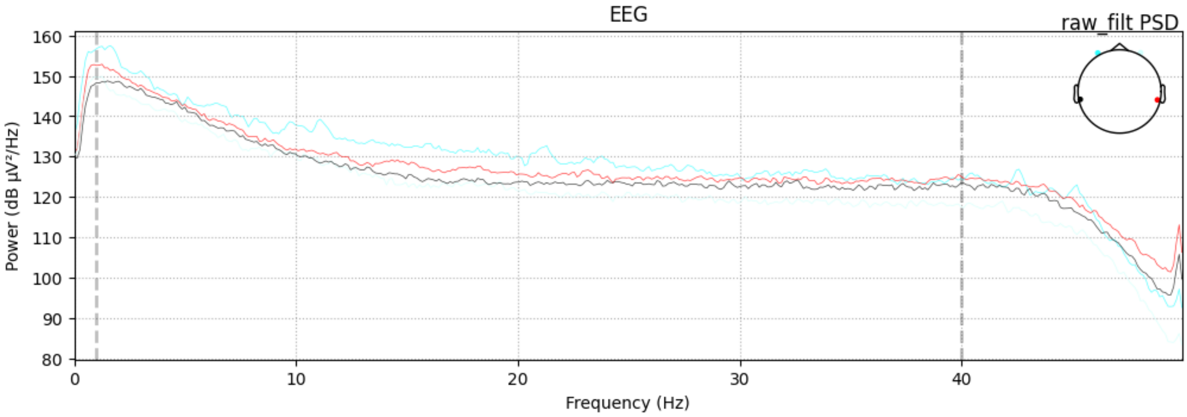

Figure 3: PSD of a session deemed good enough for usage

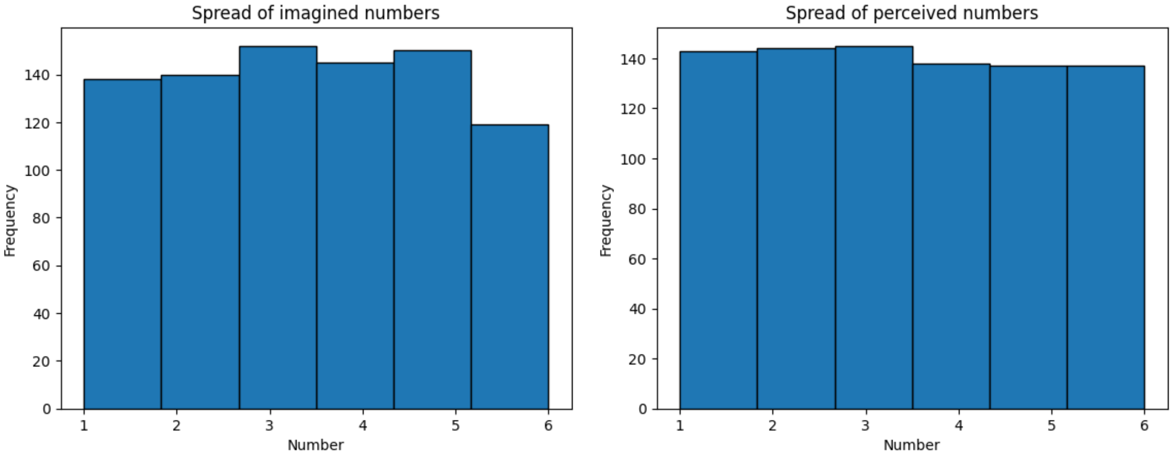

Figure 4: Spread of digits of visually perceived and imagined data

F1-score of 18.1%. The model that was trained on the data of the visually perceived numbers resulted in an accuracy of 22.35% and an F1-score of 20.0.
In order to test if the classification performance was above chance level, I used the method as proposed by Müller-Putz et al. (2008). They calculated the confidence interval using chance level and argued that if the score is above that interval, the accuracy can be seen as significantly above chance. This confidence interval is calculated as seen in Equation 3:

$$\rho \pm \sqrt{\frac{\rho(1-\rho)}{n+4}} * Z_{1-\frac{\alpha}{2}} \tag{3}$$

Where ρ is the standard chance level, n is the number of trials and $1 - \frac{\alpha}{2}$ pertains to the relevant quantile of the standard normal distribution. In this case, working with a chance level of $\frac{1}{6}$, ρ was 0.167, n was 85, and, assuming a significance threshold of $\alpha = 0.05$, the 95% confidence interval was computed.
Using these values, the theoretical limits of the confidence interval were 7.8% and 24.4%. This means that the accuracy scores of both classifiers aren't significantly above chance level. Since neither performed significantly above chance, it didn't make sense to perform any further statistical test on the results.
In order to further inspect the data, two confusion matrices were created, as seen in Figure 6. To make these more insightful, the accuracy and F1-scores

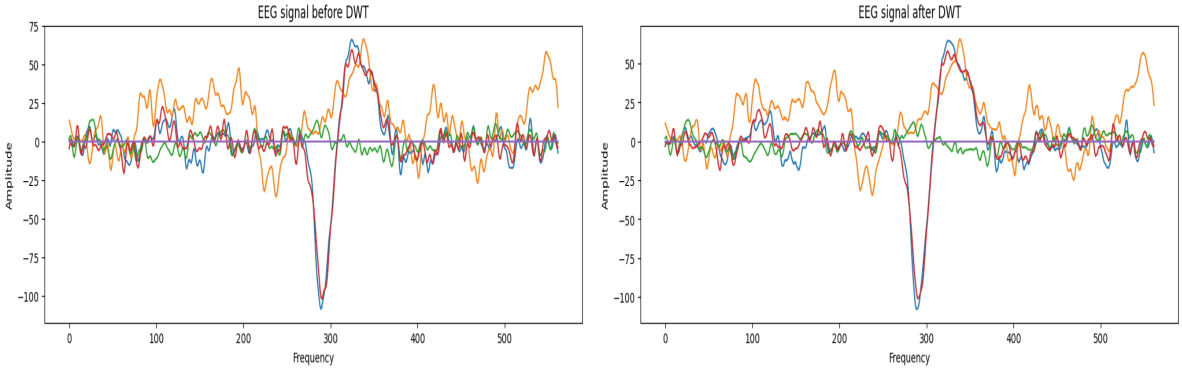

Figure 5: Signal before and after DWT. The signal is shown to be slightly more smooth after the DWT.

were calculated per class, as can be seen in Table 2. Here, it can be seen that the accuracy in predicting four, five or six for imagined digits was 0.0%. The only time the classifier trained on the visually perceived digits scored 0.0% was for the number four.

	Visual		Imagined	
Class	Accuracy	F1-score	Accuracy	F1-score
1	25.0%	0.33	57.14%	0.28
2	23.53%	0.24	31.58%	0.33
3	12.50%	0.17	27.78%	0.29
4	0.0%	0.0	0.0%	0.0
5	23.08%	0.18	0.0%	0.0
6	50.0%	0.33	0.0%	0.0

Table 2: Table showing the accuracy and F1-scores per class

4 Discussion

In this experiment, two classifiers were trained on either EEG data of imagined digits or visually perceived digits in order to find out whether they would perform better on one of the two datasets. In order to create the classifiers, the pipeline of Mahapatra et al. (2023) was followed. With the use of this pipeline I created a bidirectional LSTM model, after which a hyperparameter random search was done to get the best possible models. However, after evaluating the test data using the theory devised by Müller-Putz et al. (2008), both models failed in achieving an accuracy score that's significantly above chance.

With this study, I aimed to further explore the differences in imagery and perception. If there is a better understanding on the performance of certain classifiers on different data, future research can become more focused on the limitations still present in the field, producing better results. With the results of both classifiers, no conclusion can be drawn. Looking at the confusion matrices in Figure 6, it does become clear that the classifier didn't just pick one number in order to obtain the highest results. For some unknown reason, the classifier that was trained on the imagined digits never classified any data as ”five”. This is noteworthy, as it was the second most frequent class that was in the dataset, as seen in Figure 4. The digits of four, five and six weren't correctly predicted a single time by this classifier.

It's further worth noting that while the classifier trained on imagined digits seems to have predicted the first three classes most, whereas the classifier trained on the visually perceived data predicted mostly five and six. There's nothing conclusive to get from these results, as they can't be proved to be above chance level.

During this study however, the paper by Mahapatra et al. (2023) has gone under review, since other researchers can't seem to reproduce the results and the authors have become unresponsive. This could be one of the reasons why both classifiers haven't gotten any significant results above chance level.

This study had its own problems as well. The usage

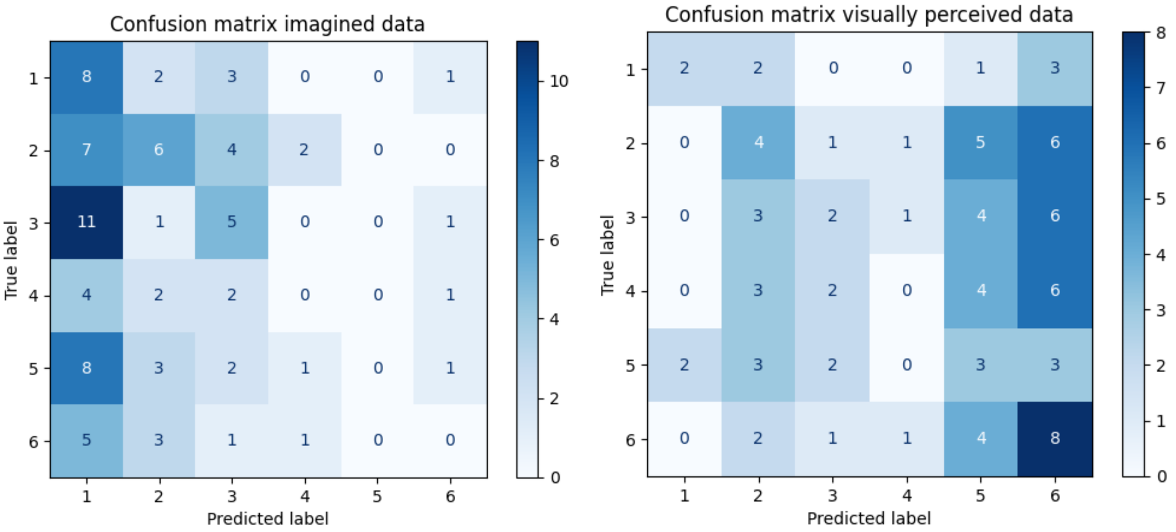

Figure 6: Confusion matrices of the predicted classes vs the true classes

of the MUSE 2 headband caused some issues. After fixing the time dependency issues, it turned out some timestamps made by the headband went back in time. This problem rendered almost half of the data useless. A bigger dataset might have helped improve the classifiers. The headband also disconnected at random moments, causing some data to be missed.

The MUSE 2 headband was chosen for its ease of use after I was convinced of its results by the paper of Mahapatra et al. (2023) However, looking at the issues that were had with the device and the paper being under review, future research might be better done with another type of headset such as Emotiv's EPOC X headset. Although this headset is a lot pricier and thus less available for the general public, it has more electrodes and covers other important areas such as the parietal cortex as well. It is however completely wireless and easy to set up, meaning that it could still be relevant in day-to-day use in the future.

Alazrai et al. (2022) also used the EPOC headset, with better results. They make use of a CNN, which has been proven to be effective in selecting temporal features. Following their pipeline using this headset, the research question could be repeated.

Other impairments to the study came from the experiment design. There's no way to know for certain that people thought of the number they said they thought of. It is assumed they participated in the way that they were supposed to, but, seeing as multiple participants later complained of boredom, they might have accidentally thought of other things.

The imagining might have also been impaired by the fact that they had to press a number on a keyboard after five seconds. Seeing as the experiment repeated itself constantly, participants would have known what they were going to do next, which means they might have been looking at the keyboard instead of just imagining the number. Future experiments could avoid this pitfall by using an eye tracking device and using different methods of getting to know what the participants were thinking.

The second part of the experiment could use a bit more interactivity however. Participants stated that even with the occasional question, their minds started to drift due to a lack of interactivity. The goal was to let the experiment progress without them having to perform any extra motor functions such as pressing the space bar, as that could potentially interfere with the results. However, participants stated that tiny things like that in the first part of the experiment helped them focus a bit more. This is also something to keep in mind for future research.

5 Conclusion

In this study, I set out to find whether a classifier would be more effective on visual or imagined data. Using Mahapatra's pipeline and training the classifiers over the EEG data of multiple participants, the classifiers never scored significantly above chance.

This may have been caused by a number of reasons, but impaired the goal of the study heavily. As both classifiers never performed significantly above chance level, there's no saying whether one could perform better than the other.

In order to further analyze the difference in visual and imagined classification, I've proposed using a different headset and pipeline. The experiment design for this study could also use some improvements, after feedback from multiple participants.

For now, the results of this study are inconclusive.

References

Alazrai, R., Abuhijleh, M., Ali, M. Z., & Daoud, M. I. (2022). A deep learning approach for decoding visually imagined digits and letters using time–frequency–spatial representation of eeg signals. *Expert Systems with Applications*, *203*, 117417.

Alazrai, R., Al-Saqqaf, A., Al-Hawari, F., Alwanni, H., & Daoud, M. I. (2020). A time-frequency distribution-based approach for decoding visually imagined objects using eeg signals. *IEEE Access*, *8*, 138955–138972.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in cognitive sciences*, *23*(5), 423–434.

Garcia-Moreno, F. M., Bermudez-Edo, M., Rodríguez-Fórtiz, M. J., & Garrido, J. L. (2020). A cnn-lstm deep learning classifier for motor imagery eeg detection using a low-invasive and low-cost bci headband. In *2020 16th international conference on intelligent environments (ie)* (pp. 84–91).

Janapati, R., Dalal, V., & Sengupta, R. (2022). Advances in experimental paradigms for eeg-bci. In *Proceedings of the 2nd international conference on recent trends in machine learning, iot, smart cities and applications: Icmisc 2021* (pp. 163–170).

Kerous, B., Skola, F., & Liarokapis, F. (2018). Eeg-based bci and video games: a progress report. *Virtual Reality*, *22*, 119–135.

Koenig-Robert, R., & Pearson, J. (2021). Why do imagery and perception look and feel so different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1817), 20190703.

Krigolson, O. E., Williams, C. C., Norton, A., Hassall, C. D., & Colino, F. L. (2017). Choosing muse: Validation of a low-cost, portable eeg system for erp research. *Frontiers in neuroscience*, *11*, 243179.

Lee, Y.-E., & Lee, S.-H. (2022). Eeg-transformer: Self-attention from transformer architecture for decoding eeg of imagined speech. In *2022 10th international winter conference on brain-computer interface (bci)* (pp. 1–4).

Mahapatra, N. C., & Bhuyan, P. (2023). Eeg-based classification of imagined digits using a recurrent neural network. *Journal of Neural Engineering*, *20*(2), 026040.

Mansi, S. A., Pigliautile, I., Porcaro, C., Pisello, A. L., & Arnesano, M. (2021, 12). Application of wearable eeg sensors for indoor thermal comfort measurements. *ACTA IMEKO*, *10*, 214. doi: 10.21014/acta$_{i}meko.v10i4$.1180

Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R., & Pfurtscheller, G. (2008). Better than random: a closer look on bci results. *International journal of bioelectromagnetism*, *10*(1), 52–55.

Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2017). Inferring imagined speech using eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering*, *15*(1), 016002.

Robinson, N., Mane, R., Chouhan, T., & Guan, C. (2021). Emerging trends in bci-robotics for motor control and rehabilitation. *Current Opinion in Biomedical Engineering*, *20*, 100354.

Tarahi, O., Hahmou, S., Moufassih, M., Agounad, S., & Azami, H. I. (2024). Decoding brain signals: A convolutional neural network approach for motor imagery classification. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 100451.