Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Analysing Deep Learning Decoding Methods on Multiple ERP Paradigms

Hannes Bonasch

Course of Study:	M.Sc. Informatik
Examiner:	Jun.Prof. Dr. Benedikt Ehinger
Supervisor:	Jun.Prof. Dr. Benedikt Ehinger
Commenced:	August 5, 2021
Completed:	February 5, 2022

Abstract

Deep learning methods have successfully advanced many fields of research with their ability to learn complex features from data. While they have been used successfully in BCI research, their use for cognitive science, where the increased complexity of deep learning methods could reveal novel insights about how our brain functions, is just starting to be explored. In this thesis, we look at three established EEG decoding models, EEGNet, Shallow ConvNet, and Deep ConvNet, on the ERP CORE dataset, which includes seven different ERP components. We will look at how parameters like model architecture, sample count, and preprocessing affect decoding accuracies, compare subject accuracies across the different ERP paradigms, and look at how feature attribution can be used to explain the decisions of our networks, as well as gain new insights into cognitive processes. We conclude that that deep learning can be a valuable tool for cognitive science that needs further research to reach its full potential.

Contents

1 Introduction

In event-related potential (ERP) analysis, electroencephalography (EEG) signals are examined in order to find out information about how the brain reacts to certain stimuli. Researchers have identified many ERP components by eliciting them in specific experimental setups, called ERP paradigms. As an example, we can look at a face perception paradigm, where a subject is either shown a picture of a face or a picture of a car. When looking at the average response of all trials of the different conditions, as can be seen in Figure 2.1, there is an increased negativity originating from the visual cortex at around 170 milliseconds when the subject is looking at a face. This negative potential is known as the N170 component and has been used to show that for example face perception is partially automatic and also modulated by attention [Luc12].

A common way to find ERP components is to first split the EEG signal into time windows called epochs, where each epoch starts at a specific event and ends after a set amount of time. Averaging these epochs for different events will then reduce the random noise while keeping time-locked responses. This averaged signal is then traditionally analyzed by univariate methods in order to gain statistically founded insights into how the brain works.

With the success of machine learning in many fields, it has also arrived in neuroscience, often under the names multivariate pattern analysis (MVPA), classification, or decoding. Here a decoder is trained, often using supervised learning, and can then predict which event a new signal belongs to. One well known application of this would be a brain-computer-interface (BCI) where generally a paradigm that is reliable to decode, such as a P300 speller, is decoded in real time and used as an input to a computer. With their increased complexity, multivariate methods allow researchers to find previously unidentified patterns of brain activity, and therefore are now also a central part of cognitive science [HB18]. However, due to the change from the firm statistical framework of univariate analysis to prediction with multivariate analysis, the interpretability of the results can suffer [HB18].

Deep learning methods for decoding, which stack multiple layers that extract progressively higher-level features, exacerbates both the positive and negative aspects of multivariate methods. Without the need for feature engineering, these models can work without domain knowledge and could find complex representations that traditional

feature extraction might miss. In consequence, as they are hard to interpret, they are also considered black-box models. A lot of research into deep learning methods comes from BCI researchers, as the goal here is often to maximize decoding accuracy.

For cognitive science, even accuracies just above chance can reveal that there is information present that helps distinguish the different conditions. The adoption of deep learning models in cognitive science is held back by the need for large datasets and interpretability issues [TRP21]. However, with large public datasets, data augmentation, and advances in explainable artificial intelligence (XAI), deep learning models could become a valuable tool for cognitive science [TRP21].

The main idea behind this thesis was to take a dataset with multiple known ERP components, explore how deep learning decoding methods work, and how they can be used for ERP analysis. For this we took the ERP CORE [KFZ+21] dataset, which provides six different ERP paradigms eliciting seven different components. What makes this dataset interesting is that it was recorded from the same recording system and the same forty subjects, which allows for better comparisons between the ERP components, as they should have similar noise levels. For the deep learning methods we used EEGNet [LSW+18], as well as Shallow and Deep ConvNet [SSF+17], as these methods have been successfully used for multiple different EEG decoding tasks before [GSC+20] [NFM+19] [ZLLG21] [SJYX21].

After a short overview of the ERP CORE dataset, the thesis is divided into three parts. In the first part, we looked at the decoding accuracy of the different models and tasks, and how it is affected by changing the size or preprocessing of the dataset. In the second main part of the thesis, we analyzed the decoding accuracies of individual subjects. Here we unexpectedly found no significant correlation of the task accuracies between subjects, going against the idea of “BCI Illiteracy” [AN10] in subjects. For the last part we examined feature attribution methods, in our case specifically DeepLift [SGK17], and looked at how they can be used to interpret the deep learning models.

2 Dataset

The ERP Compendium of Open Resources and Experiments (ERP CORE) released by Kappenman et al. [KFZ+21] consists of optimized paradigms, experiment control scripts, processing pipelines, and a dataset for seven different ERP components. ERP CORE aims to help address a lack of standardization in ERP paradigms and analysis, by serving as a starting point for new ERP labs, novel paradigm development, or how it is used in this thesis, for testing different analysis methods.

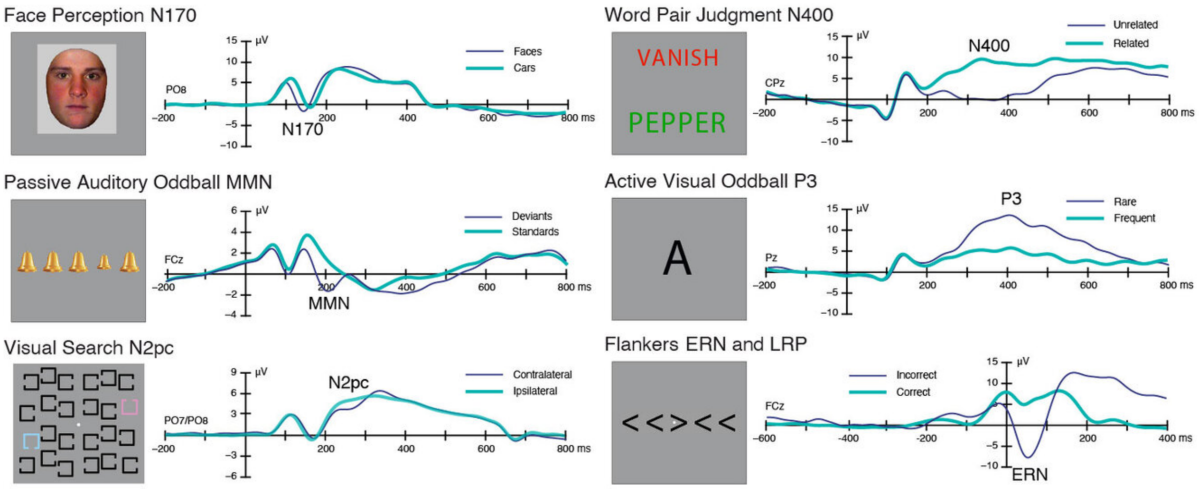

Figure 2.1: Sample image of a stimulus and averaged epochs of a relevant electrode for each condition of all the paradigms. Adapted from the ERP CORE paper [KFZ+21].

The six different ERP paradigms evoke seven different ERP components. The averaged epochs and a sample stimulus for the different conditions are shown in Figure 2.1. A visual discrimination between cars and faces isolates the N170 component. The mismatch negativity (MMN), is evoked using an auditory oddball paradigm. For the N2pc component, the different conditions are based on a visual search with the target either on the left or the right side of the visual field. The N400 component is elicited when a prime word that is usually followed by a semantically related word is followed by a surprising unrelated word. In a visual oddball paradigm, where the subjects are

Table 2.1: Number of samples per condition of all tasks, pooled over all 40 subjects after rejecting late or incorrect responses. For the N2pc and LRP tasks, the conditions left and right were labeled zero and one, respectively. For all other tasks, condition zero refers to the component being absent, and condition one to the component being present.

	Component						
Condition	N170	N400	P3	MMN	N2pc	ERN	LRP
0	2860	1894	6186	31265	5370	14230	7951
1	2921	1549	1338	7967	5456	1818	8097

shown five random letters, the P3 component occurs, when the designated target letter appears. A flanker paradigm, where the subjects press a button corresponding to the direction of the central arrow, is used to elicit both the error related negativity (ERN) and lateralized readiness potential (LRP) component. The LRP occurs as a negative potential contralateral to the response hand. The ERN occurs as a negative potential in incorrect responses. Each paradigm was recorded for around ten minutes [KFZ+21], leading to varying sample counts for the different tasks. In addition, the amount of samples per condition are also unbalanced for some of the tasks, as can be seen in Table 3.1.

The raw data of the ERP CORE dataset was loaded using MNE BIDS [ASB+19], and unless mentioned otherwise further processing was done using MNE-Python [GLL+13]. It was then downsampled from 1024Hz to 250Hz in order to reduce computational time. As in the ERP CORE processing pipeline, the raw data was then referenced to the P9 and P10 electrodes for all components except N170, which was referenced to the average of all electrodes. This was done to have a better comparison to the ERP CORE analysis, and while it did not have an impact on decoding accuracy, it did change the spatial feature attribution.

At this point, we introduce our three different levels of preprocessing: “light”, “medium” and “heavy”. While for the lightly preprocessed data no filter was used, for the medium and heavy preprocessing we applied the FIR filter as implemented in MNE between 0.5Hz and 40Hz. For the heavy preprocessing, we additionally used Automagic [PBL19], a MATLAB based EEG preprocessing toolbox, to remove artifacts with independent component analysis based on AMICA [PKM12] and the ICAlabel toolbox [PKM19], and then later used the AutoReject [JEB+17] library to automatically set a rejection threshold for epochs with remaining artifacts. The difference between the preprocessing levels for a single trial of the P3 task on the FP1 electrode can be seen in Figure 3.6.

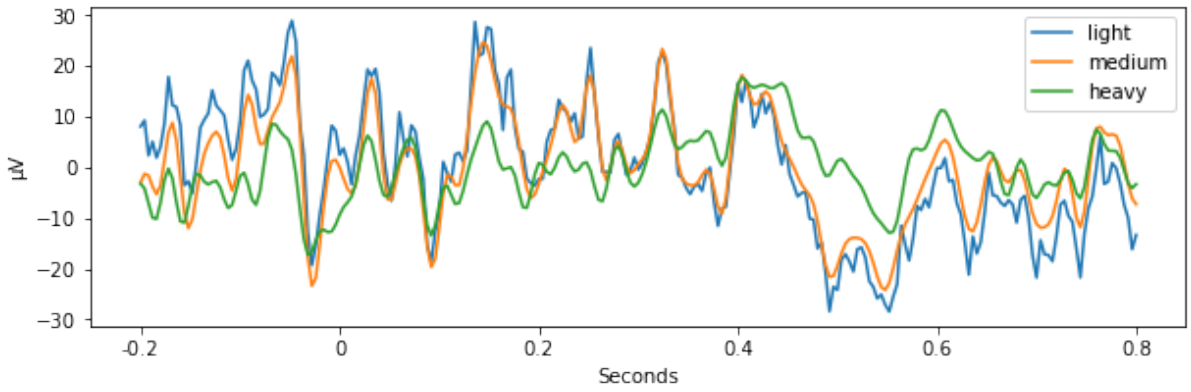

Figure 2.2: Single trial showing the difference between the light, medium and heavy preprocessing on the FP1 electrode of the P3 task.

After shifting the events of paradigms including visual cues by 26ms to account for the LCD delay, we cut the raw signal into one second long epochs at the appropriate events for all paradigms. If the button response to the stimuli was incorrect or not inside this time window, the epoch was removed, except for ERN, where this error response is of interest. Linear detrending was applied to remove the DC offset. Without removing the DC offset, the models did not converge, this was surprising and might warrant further investigation. The epochs were then stacked into a Pandas [McK+10] dataframe together with their task and subject information, ready to be used in PyTorch [PGM+19].

3 Benchmarks

3.1 Architectures and Training

As there was no comparison for the decoding accuracy of the ERP CORE dataset, we decided to use the EEGNet model by Lawhern et al. [LSW+18], and the Shallow ConvNet and Deep ConvNet by Schirrmeister et al. [SSF+17], as all three of these models have been used successfully for multiple different EEG decoding tasks [GSC+20] [NFM+19] [ZLLG21] [SJYX21].

3.1.1 Shallow and Deep ConvNet

The paper by Shirrmeister et al. [SSF+17] investigated how convolutional neural networks could be used for EEG decoding. They used the popular Filter Bank Common Spatial Pattern (FBCSP) [ACW+12] method as a baseline and as an inspiration for their Deep and Shallow ConvNets. Both ConvNets start with a temporal convolution in the first layer, followed by a spatial filter in the second layer, similar to the band-pass filter followed by the CSP spatial filter of FBCSP. For the Deep ConvNet these two first layers are followed by max pooling and then three standard convolution-max-pooling blocks. The Shallow ConvNet replaces the three convolution-max-pooling blocks with mean pooling followed by a single dense layer. The architecture was intentionally kept generic, in order to be used as a general purpose EEG decoder that does not rely on expert knowledge.

Their results showed that the ConvNets are competitive with accuracies reached by FBCSP, and input-perturbation revealed that the deep ConvNet did learn relevant band power features and plausible spatial distributions. They also outlined that while their ConvNets did not improve over FBCSP significantly, larger datasets, further improvements in deep learning methodology, and better feature attribution methods make deep learning a promising tool for EEG decoding.

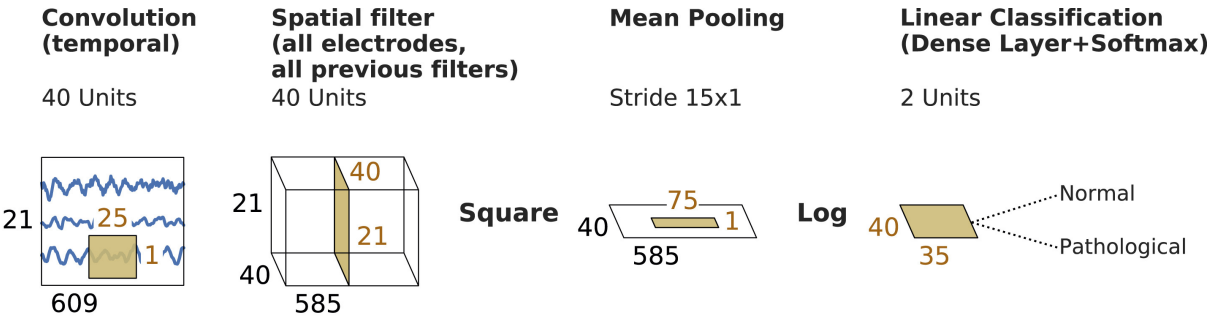

Figure 3.1: Architecture of the Shallow ConvNet, showing the temporal and spatial convolution layers, followed by mean pooling and a simple dense layer [SSF+17]. From the paper of Gemein et al. [GSC+20].

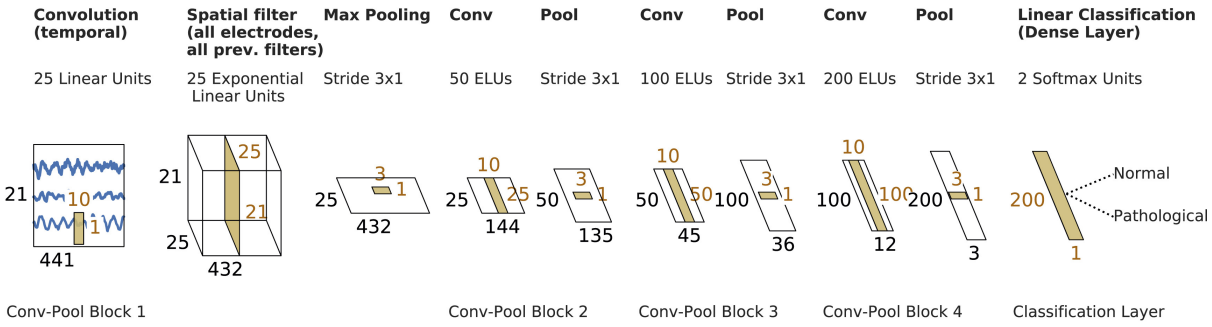

Figure 3.2: Architecture of the Deep ConvNet, showing the temporal and spatial convolution layers, followed by three convolution-max-pooling blocks [SSF+17]. From the paper of Gemein et al. [GSC+20].

3.1.2 EEGNet

EEGNet, as introduced by Lawhern et al. [LSW+18], is a compact convolutional neural network designed to generalize to different BCI paradigms. The first layer consists, as for the ConvNets, of a temporal convolution, but EEGNet then achieves the small parameter count by using a depth-wise convolution [Cho17], followed by a separable convolution, as seen in Figure 3.3.

In their paper, EEGNet was validated on two ERP-based and two oscillatory-based BCI datasets. In their results, they show that EEGNet performed similarly to Deep ConvNet, but outperformed it in the within-subject case, which they attribute to the smaller parameter count leading to EEGNet being less data intensive [LSW+18].

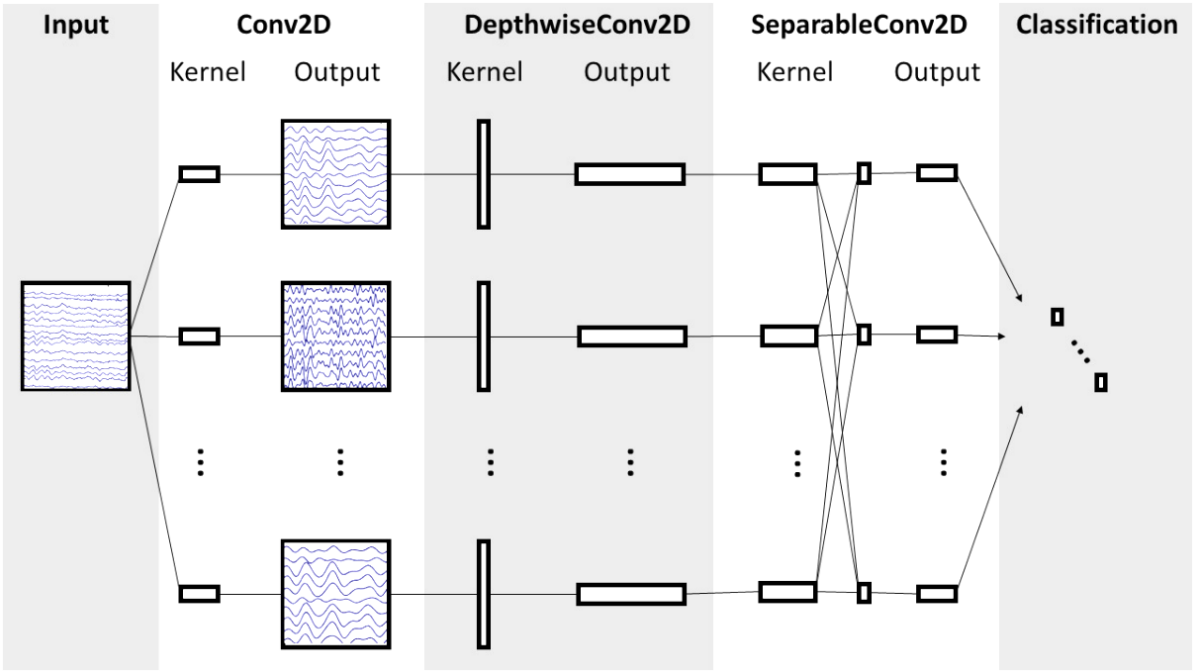

Figure 3.3: Architecture of the EEGNet model. [SSF+17]

3.1.3 Training Process

As one promise of deep learning methods is their ability to be used without domain knowledge, and in order to avoid overfitting to the dataset or having to split away a test set for nested cross-validation, we decided to use the models as implemented in the Braindecode toolbox [SSF+17] with minimal hyperparameter tuning.

The models were optimized using Adam [KB14] with a cross entropy loss function. Cosine annealing [LH16] without restarts was used to schedule the learning rate. As several of our components are unbalanced datasets, class weights and stratified cross-validation were used. All scores reported are the balanced accuracy, which is the arithmetic mean of sensitivity and specificity, as we care equally about false positives and false negatives. For reproducibility, the code and all models will be uploaded to GitHub.

3.2 Model Comparison

With the purpose of assessing the predictive power of our three models, we used ten stratified random splits with 20% of the data as validation, ignoring subject structure, following the guidelines by Varoquaux et al. [VRE+17]. For this first benchmark the medium preprocessing was used. As a sanity check and comparison, we also included a linear support vector machine (SVM) [CV95] in this benchmark.

The results, in Figure 3.4, show the validation balanced accuracy of the models cross-validated on the entire training data. As can be seen, EEGNet and Deep ConvNet are relatively close for most tasks, with an average of 75.1% and 75.8% respectively. Shallow ConvNet falls behind in some tasks, leading to an average of 74.1%. The linear SVM performs worse on all tasks, especially on P3 and ERN, with an average of 64.2%.

When comparing the performance of the individual tasks, we have to keep in mind that decoding accuracy can not be equated with the effect size of the signal [HB18].

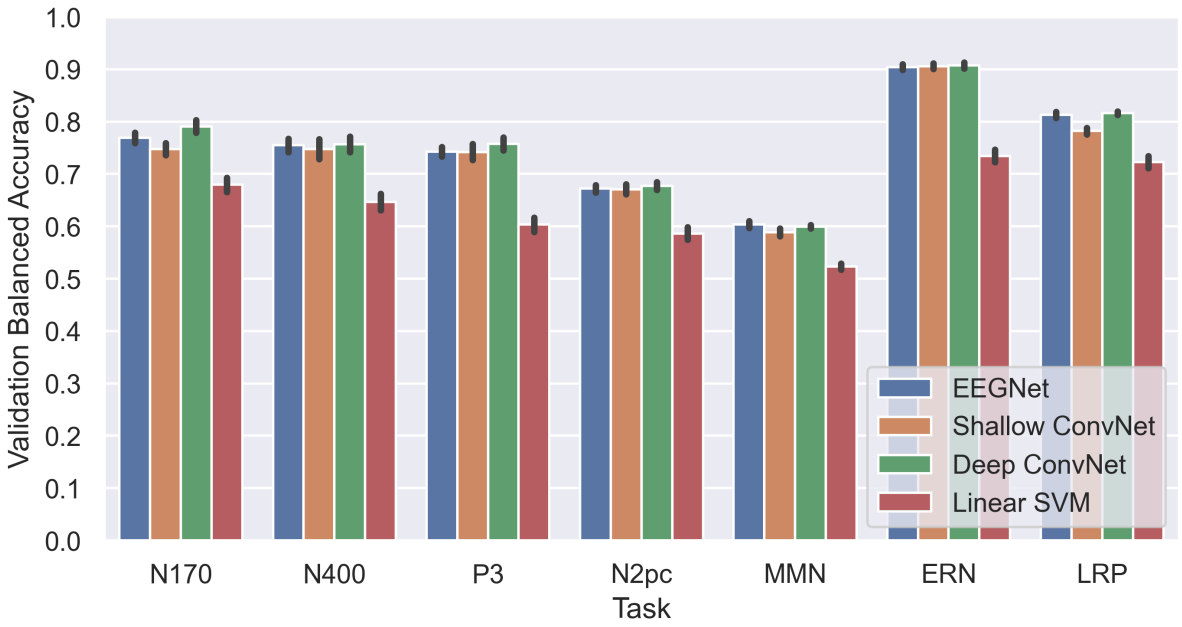

Figure 3.4: Bar plot showing the average balanced accuracy and standard deviation of ten cross-validation splits of all subjects combined for each task.

3.3 Sample Count Comparison

The difficulty of producing large EEG datasets can be an obstacle for the use of deep learning models. To get a sense of how much the training data size impacts the accuracies and variance, we trained the four models on a subset of the data. To better compare the difficulty in decoding the different tasks, we chose to use equal and balanced sample counts for all tasks.

As we can see from the average balanced validation accuracy and standard deviation of all tasks in Table 3.1, the Deep ConvNet achieves the highest accuracies across the board. For the lower sample counts this was surprising, going against the expectation that Deep ConvNet is more data intensive [LSW+18] with, in our case, over 25 times more parameters than EEGNet.

Table 3.1: Average balanced validation accuracy and average standard deviation of all tasks per model and sample counts. The sample counts for the complete datasets can be seen in Table 3.1.

	Sample Count			
Model	200	600	2000	All
EEGNet	62.4 ± 5.8%	69.0 ± 4.0%	72.1 ± 2.2%	75.1 ± 0.9%
Shallow ConvNet	59.1 ± 6.7%	63.2 ± 4.4%	67.8 ± 2.3%	74.1 ± 1.1%
Deep ConvNet	**62.9** ± 6.9%	**69.2** ± 4.2%	**72.6** ± 2.1%	**75.8** ± 0.9%
Linear SVM	60.6 ± 6.5%	63.7 ± 3.5%	65.6 ± 2.1%	64.2 ± 1.3%

Looking at the individual task accuracies in Figure 3.5, we can clearly see that having lower sample counts affects the tasks differently. Going from 200 samples to the full dataset barely changes the decoding accuracy of MMN, while showing a very strong improvement for other tasks, like N170, N2pc and LRP.

Interestingly, the change in sample count also affects the four models in different ways, with Shallow ConvNet benefitting a lot from the full dataset. At 200 samples, EEGNet is barely decoding N170 above chance, whereas Deep ConvNet is almost 10% more accurate, and for LRP the opposite is true.

We are not able to tell from this data how much of this effect is from the models being fundamentally different at decoding the tasks, and how much is from the hyperparameters fitting the individual tasks better or worse. It would be interesting to repeat this experiment with the hyperparameters optimized for each individual task.

For linear SVM, the boost in sample count does not always lead to an increased accuracy, here we clearly hit diminishing returns very quickly. On the heavily imbalanced tasks, like P3 and ERN, the linear SVM performs worse on the full dataset than the balanced subset of 2000 samples, even though class weights were used.

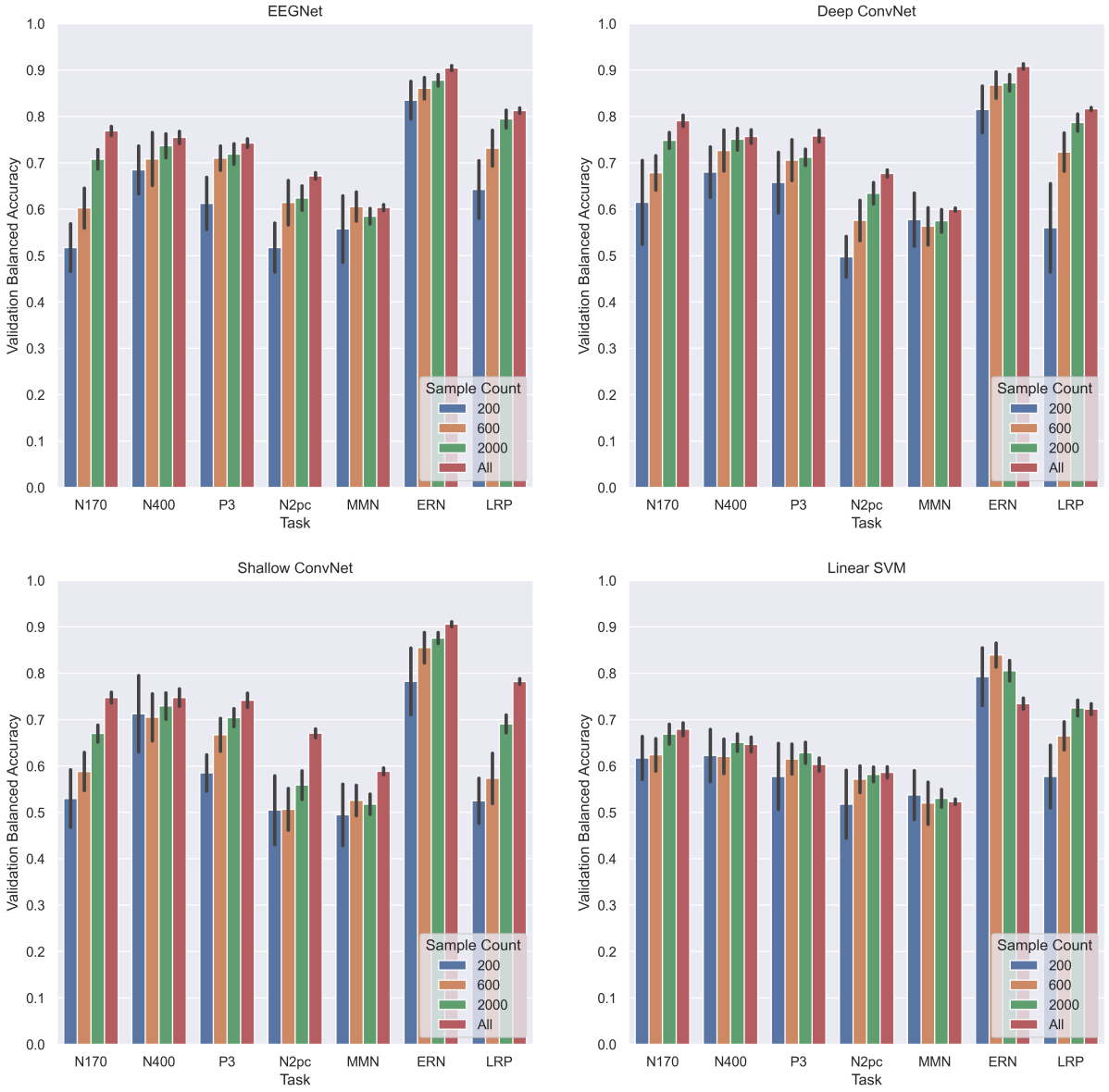

Figure 3.5: Bar plots for each model showing the average balanced accuracy and standard deviation over ten cross-validation splits for a balanced subset of 200, 600, and 2000 samples, as well as the entire dataset. The sample count for the full dataset can be seen in Table 3.1.

3.4 Preprocessing Comparison

In order to see if preprocessing the data can influence the decoding accuracies, we repeated the tenfold cross validation of all models on the light, medium, and heavy preprocessing levels. As can be seen in Figure 3.6, there was no significant difference between the preprocessing levels for any of the models. While we anticipated the deep learning models to be able to handle the more noisy data, we at least expected a positive impact of the heavy preprocessing, due to rejecting 3% to 16% of the noisy data, depending on the task. However, as can be seen in Chapter 5.3, preprocessing did have an impact on how the models decided on their predictions.

3.5 Conclusion

By looking at the decoding accuracies of the three deep learning models, we showed that EEGNet and Deep ConvNet have very similar performance across tasks, that large sample sizes are required for good decoding accuracies, and that our preprocessing had little impact on the validation scores. These findings were mostly in line with our expectations, nevertheless, they can serve as a baseline for what to expect from deep learning decoding methods on ERP data, and for our further analysis.

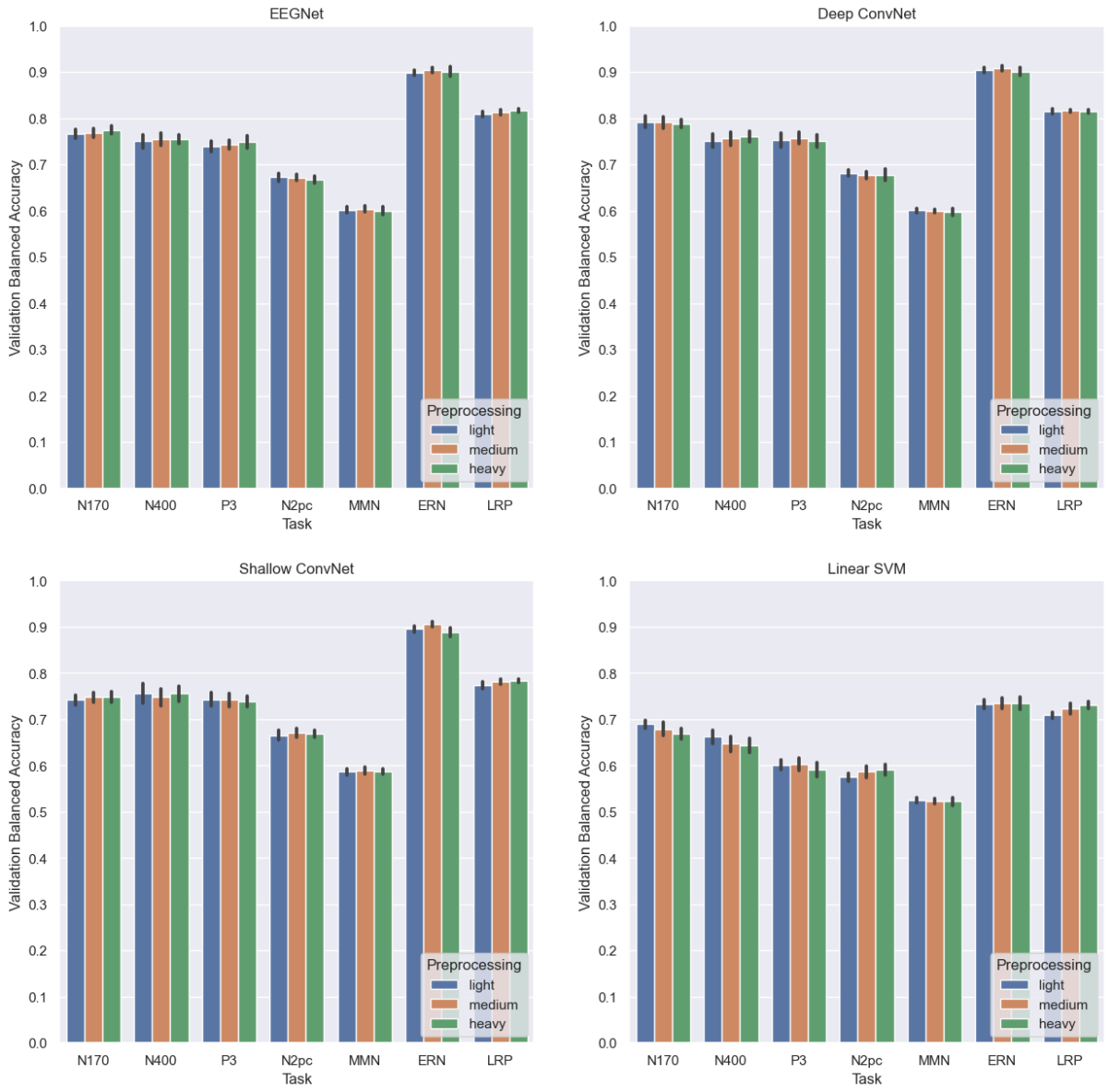

Figure 3.6: Bar plots for each model showing the average balanced accuracy and standard deviation over ten cross-validation splits for light, medium and heavy preprocessing of the entire dataset.

4 Subject Analysis

4.1 Overview

As all the recorded paradigms of the ERP CORE dataset are from the same forty subjects, we have the opportunity to look into how the different subjects perform across all seven tasks. For this analysis, we decided to restrict to the Deep ConvNet model, as it achieved the best overall scores in the previous chapter.

4.1.1 BCI Illiteracy

In BCI research, there exists a phenomenon often called “BCI Illiteracy”, where around 20% of subjects are not able to use a particular BCI system [AN10]. There can be many causes for a subject to not reach decoding accuracies that are sufficient for BCI systems, ranging from subjects not producing clear EEG signals through excessive muscle artifacts or variations in brain structure, subjects misunderstanding instructions, to errors caused by the researchers [AN10].

As this phenomenon has shown up in many studies [AN10], there is also a lot of research into the causes [BSH+09] and possible solutions [VB10].

However, there is also critique of the entire concept of “BCI Illiteracy”. For example, the performance thresholds for what counts as effective use of BCI are often ill-defined [Tho19]. In addition, labeling the subjects as “illiterate” can put the fault of the poor performance of a BCI system entirely onto the user [Tho19].

The paradigms used in ERP CORE were not made with the same intention of creating a BCI system, and have varying decoding accuracies, so establishing a threshold for what counts as “illiterate” in our case is difficult. They also include button presses in some tasks, which could influence decoding. Nevertheless, the phenomenon of significantly lower decoding accuracies for some subjects most likely also extends to ERP decoding for cognitive research.

4.2 Cross Subject

Firstly, we looked at the cross subject decoding accuracies, using leave-one-subject-out training, where the model is validated on the data of each subject, while trained on the data from all the remaining subjects. This maximizes the amount of training and validation data, while effectively simulating using a pre-trained model on an unseen subject.

Figure 4.1 shows the balanced accuracy scores of all subjects on all tasks. The color map shows the standardized accuracies for each of the tasks to get a better visual indication of how different the subjects are from the average task accuracies. The subjects were also ordered by their averaged task accuracy scores.

Notably, no single subject performs below average in all of their tasks. Even subject 29 with the lowest average balanced accuracy of 68% still achieves slightly above average results in three of the seven tasks. On the other hand, there are multiple subjects performing above average in all the tasks, with subject 20 having the highest average score across all tasks of 80%.

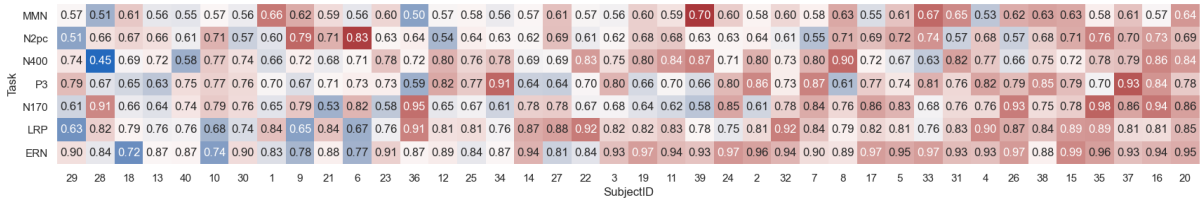

Figure 4.1: Balanced accuracy scores for each subject and task of the cross-subject case. The color map was standardized for each task, so that red scores indicate above average decoding accuracy for that particular task.

With these unexpectedly noisy results, we wanted to check for correlations between individual tasks. Especially with some ERPs like the N400 and P3 sharing similar temporal and spatial activity [ASD11], we expected there to be a relatively strong correlation between the subject accuracies of these ERPs.

In Figure 4.2, we plotted the individual tasks against each other. Visually, no strong correlation was apparent. The highest Spearman correlation coefficient is 0.33, which would indicate a moderate monotonic correlation between LRP and ERN, as well as N170 and N2pc, with a p-value of 0.04 at our sample size. However, as we did multiple comparisons, this p-value is not significant when applying Bonferroni correction.

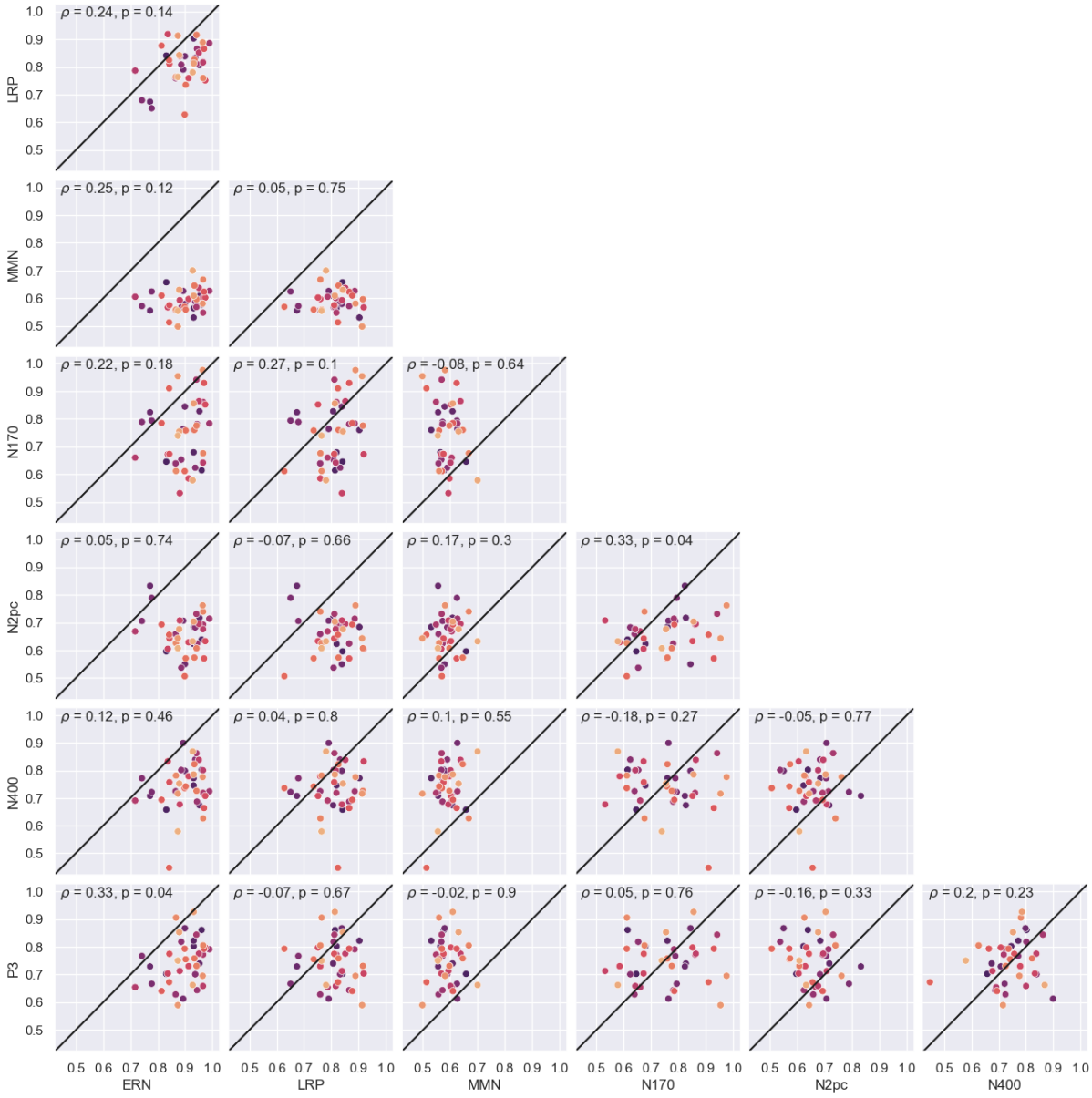

Figure 4.2: Pair plot of the balanced accuracy of all task combinations in the cross-subject case. Spearman correlation ρ and p-value p for each pair is included in the top left.

4.3 Within Subject

In order to double-check our findings and to eliminate the influence of other subjects, we also looked at the within-subject accuracies. Here, we used the average balanced accuracy of tenfold cross-validation to mitigate the high variance at lower sample counts.

As can be seen in Figure 4.3, there are now some subjects that have lower than average decoding scores in all tasks, with the worst subject having an average balanced accuracy of 62%. Interestingly, the subject with the highest average decoding accuracy is now at 83%, higher than the best score in the cross-subject case.

Looking at the correlations between the within-subject scores in Figure 4.4, the highest correlation is now between LRP and N170, with a Spearman correlation coefficient of 0.39 and p-value of 0.01. The previously highest correlation in the cross-subject analysis between LRP and ERN, as well as N170 and N2pc now only has a correlation coefficient of 0.14 and 0.04 respectively.

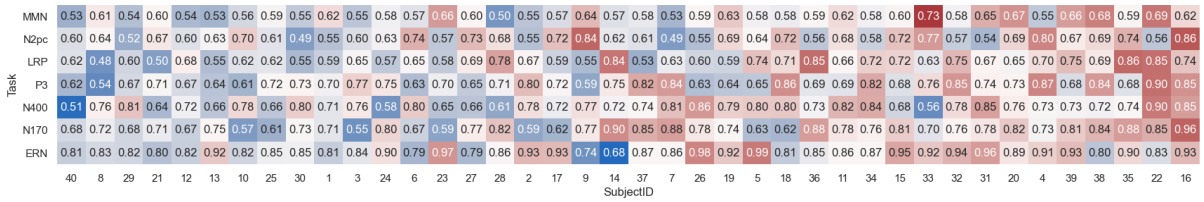

Figure 4.3: Averaged balanced accuracy scores of ten cross-validation folds for each subject and task in the within-subject case. The color map was standardized for each task, so that red scores indicate above average decoding accuracy for that particular task.

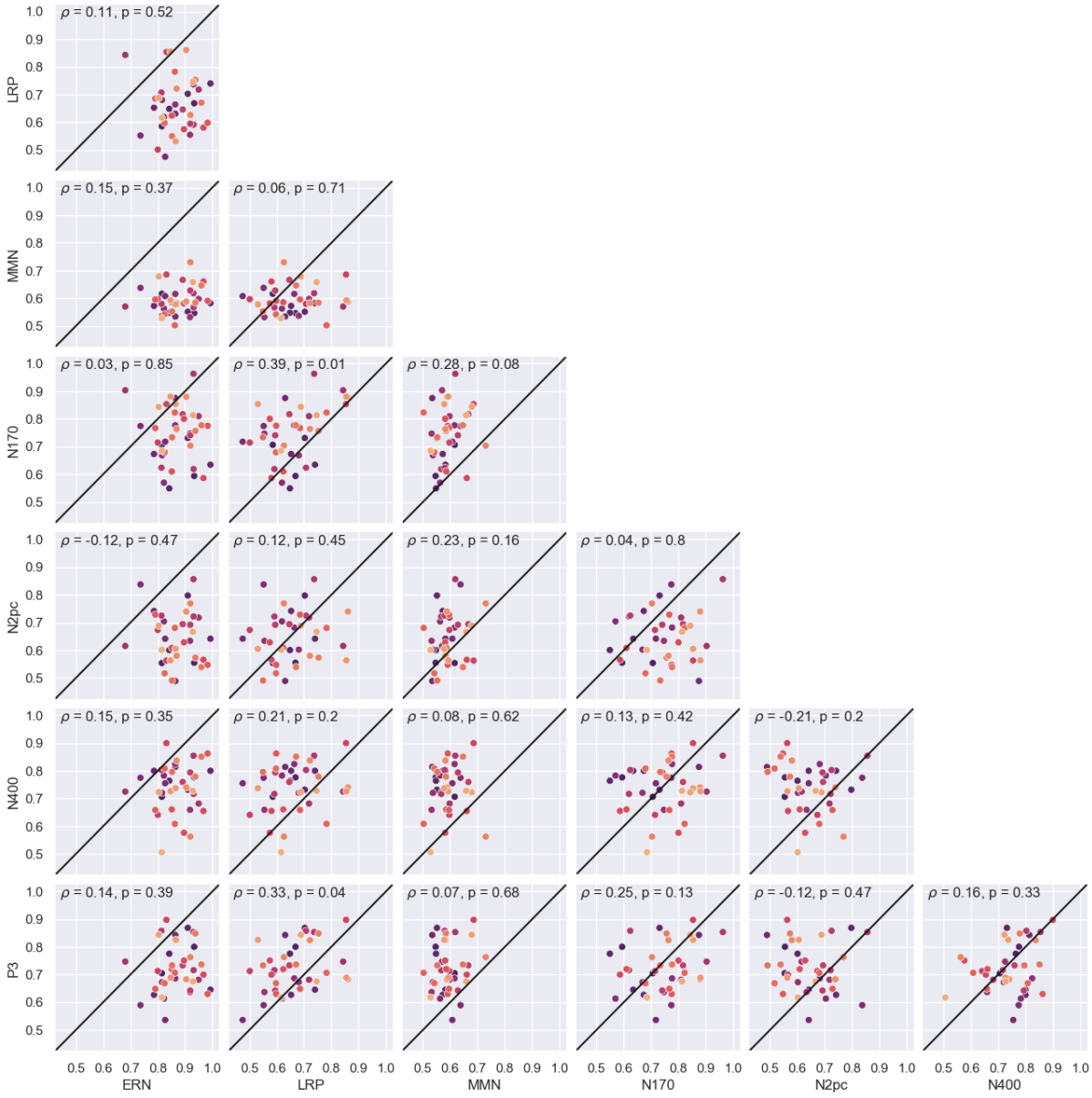

Figure 4.4: Pair plot of the averaged balanced accuracy over ten cross-validation splits of all task combinations in the within-subject case. Spearman correlation ρ and p-value p for each pair is included in the top left.

4.4 Cross vs. Within Subject

Plotting the cross-subject against the within-subject scores, we can see that there is a statistically relevant monotonic correlation for all tasks, even when Bonferroni corrected. From the scatter plots, we can also see that LRP benefitted especially from cross-subject training. The fact that N170 has the least correlation between the cross-subject and within-subject training might suggest that the N170 component has the most variation between subjects.

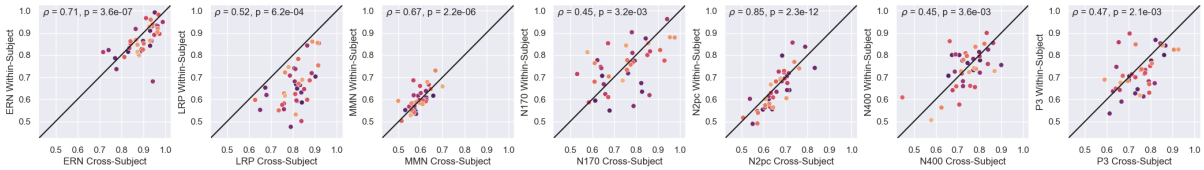

Figure 4.5: Scatter plot showing the within-subject balanced accuracies plotted against the within-subject averaged balanced accuracies. Spearman correlation ρ and p-value p for each pair is included in the top left.

4.5 Conclusion

For the cross-subject training, no single subject had below average decoding accuracies on all tasks, which together with the fact that there is no statistically significant correlation between any of the tasks, could indicate, that “BCI Illiteracy” is not necessarily a property of the subject, but more of a mismatch between subject and BCI system. This would be in line with the study by Lee et al. [LKK+19], which used three different BCI paradigms, and showed that no subject was “universally illiterate”, meaning that each subject was at least sufficiently accurate in one of the paradigms. They also showed no statistically significant correlation between the subject accuracies of their ERP, steady-state visually evoked potential (SSVEP) and motor imagery (MI) paradigms.

5 Feature Attribution

The field of explainable artificial intelligence (XAI) tries to open up the black box nature of deep learning models. Understanding what factors play into the decision-making of deep learning models can not only help to find errors and help improve the models, but is also important in establishing trust in them [AB18]. Especially interesting for cognitive science is that these explanations could also help discover novel and unforeseen insights.

There are two main approaches to making deep learning models more explainable, one is to make less complex or inherently interpretable models, the other is to try to explain how a model made its decision [AB18].

We will be focusing on feature attribution methods, where the aim is to explain a model by providing a relation between the input and output. Feature attribution methods that give explanations for a single input are called local methods, compared to global methods, which provide, for example, a constructed sample with the highest probability for that class [TRP21].

As there is still a lot of research into feature attribution methods and their use for EEG data is still relatively novel, there is no established method to use. Researchers use a variety of different local feature attribution methods, from Saliency [VMS+20], input perturbation [SSF+17], SHAP Values [NFM+19], LRP [SLSM16], to DeepLift [LSW+18].

The decision on which feature attribution method to use is also complicated by the fact that evaluating what counts as a good explanation is not always easy to quantify [DK17]. Removing features that are deemed as important and then assessing the change in accuracy is one way to gauge the performance of different attribution methods. For computer vision and natural language processing tasks, such evaluations have shown that backward decomposition based methods, such as layer-wise relevance propagation [BBM+15] and DeepLift [SGK17], are generally recommended [TRP21]. For time series classification, layer-wise relevance propagation and DeepLift were also recommended over Saliency, LIME, and SHAP [SAE+19]. With layer-wise relevance propagation and PatternAttribution having issues with deeper layers [SGL20], and assigning attribution to random noise, that were not exhibited by DeepLift [KSJ+19], we decided to use DeepLift for our analysis.

As with other backward decomposition based methods, DeepLift starts at the output layer and moves backwards to the input while distributing an attribution score to each neuron [ACÖG17]. One difference with DeepLift compared to other methods is that it explains the difference in output from a reference output with the difference in input from a reference input, which allows the attribution to propagate even if the gradient is zero [SGK17]. An input of zeros was chosen for reference, as in the EEGNet paper [LSW+18], however further investigations on what is a good reference value for ERP data might be needed. For this thesis, the DeepLift implementation with the rescale rule from the Captum toolbox [KMM+20] was used.

5.1 Interpretation of Feature Attribution

Before looking at the attribution weights, we need to establish how they can be interpreted in the context of decoding EEG signals. As thoroughly explained by Hebart and Baker [HB18], there can be several confusions that emerge due to the mix of the activation-based philosophy of the firm statistical background of EEG analysis and the information-based philosophy of machine learning.

In univariate analysis, the difference between the mean parameters of the conditions, called the signal strength, is used to determine if there is a statistically significant effect. In activation-based philosophy, noise is seen as an error of the underlying signal.

In multivariate decoding, suppressor variables that are not directly correlated with the output, like a signal only containing noise, can still provide information that helps the prediction, if for example it can be used to subtract noise from other signals. Many feature attribution methods would assign importance to such a suppressor variable that contains information [WBMH21]. From an information-based standpoint, that suppressor variable clearly helped the model make its prediction, and removing it might lower the decoding accuracy.

This means that when interpreting feature attribution, we have to keep in mind that importance is determined by the information and does not necessarily reflect an underlying brain signal.

There are promising solutions for multivariate decoding methods that allow for better neurophysiological interpretability, such as converting backward models to forward models [HMG+14], or feature attribution methods that assign less importance to suppressor variables like PatternAttribution [WBMH21].

For the following visualizations, blue will indicate a positive attribution. A positive attribution as assigned by DeepLift can be interpreted as the variable containing information

that is used by the decoder to decide for condition one. In the case of N2pc and LRP this means the right side condition, and for the other tasks it means that the condition that contains the component, for example the face stimulus for N170.

The color scale of the visualization was chosen for each plot individually by the absolute maximum attribution in order to avoid clipping and to keep zero attribution centered. It is also unclear how well the units can be compared between different DeepLift attributions.

5.2 Averaging Trials

Local feature attribution methods give us the ability to look at the attribution of a single trial. While it can be informative to look at single trials to see exactly how the decoder decided on a specific sample, in order to get more generalizable insights on the whole task or the deep learning model, we need to look at many single trials.

In the Figure 5.1, we see signals of the PO8 electrode, where we can expect the N170 component to occur [KFZ+21], together with the attribution to the prediction of the Deep ConvNet model as derived from DeepLift.

In the top of Figure 5.1, we can see a single correctly classified positive trial, where the subject was looking at a face, and a single true negative trial, where the subject was looking at a car. The true positive trial of the top figure is a good example of how a single trial can look vastly different from the average. It only has a very minor peak at around 90 milliseconds that is attributed positively, indicating that it contains information leading to the classification as a face or positive condition.

The bottom figure shows the average of all true positives and true negatives of the validation set. Here we can clearly see the difference between the two conditions, and how the average peak in attribution of true positive trials is now at 150 milliseconds, right before the peak in voltage.

To avoid further complexity, we mostly looked at the average attribution of multiple trials in this thesis. More advanced statistical methods to analyze the attribution of multiple single trials could provide more insights.

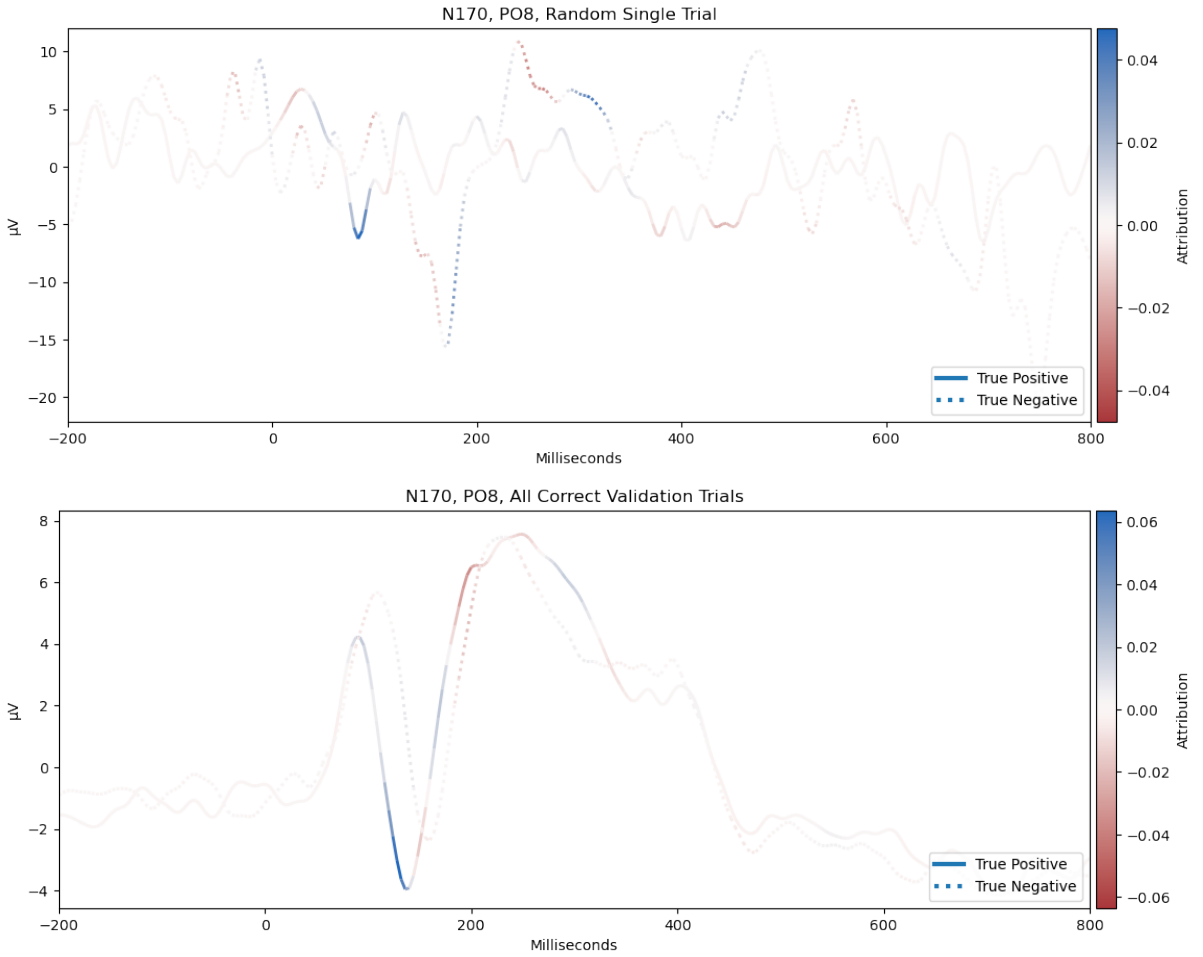

Figure 5.1: Line plots showing the voltage and attribution of the PO8 electrode in the N170 task. Top figure shows a single random correctly classified positive and negative trial. Bottom figure shows the average of all true positive and true negative trials of the validation set.

5.3 Preprocessing

While we showed no changes in decoding accuracies between the different preprocessing levels in Chapter 3.4, there was a clear difference in the attribution assigned.

Figure 5.2 shows the difference between medium and heavy preprocessing on the P3 task. Light preprocessing was not included, as it looked very similar to medium preprocessing. As can be seen, on the medium preprocessing the FP1 and FP2 electrodes, that sit very close to the eyes, have by far the highest attribution. After removing artifacts with the heavy preprocessing, most of the attribution of these two electrodes is removed. While some other tasks also showed reduced FP1 and FP2 attribution, P3 showed by far the strongest reduction.

One possible interpretation of this is that on light and medium preprocessing the model mostly decided which condition the trials belong to based on eye artifacts, perhaps by the subjects looking for the buttons to press when the rare event of the P3 paradigm occurs. However, the fact that the decoding accuracies did not change even though the heavy preprocessing removed these seemingly important eye artifacts, is unexpected.

We decided to use heavy preprocessing for all further analysis, as it seemed to reveal attribution distributions that are more in line with plausible brain activity, while removing attribution to possible artifacts.

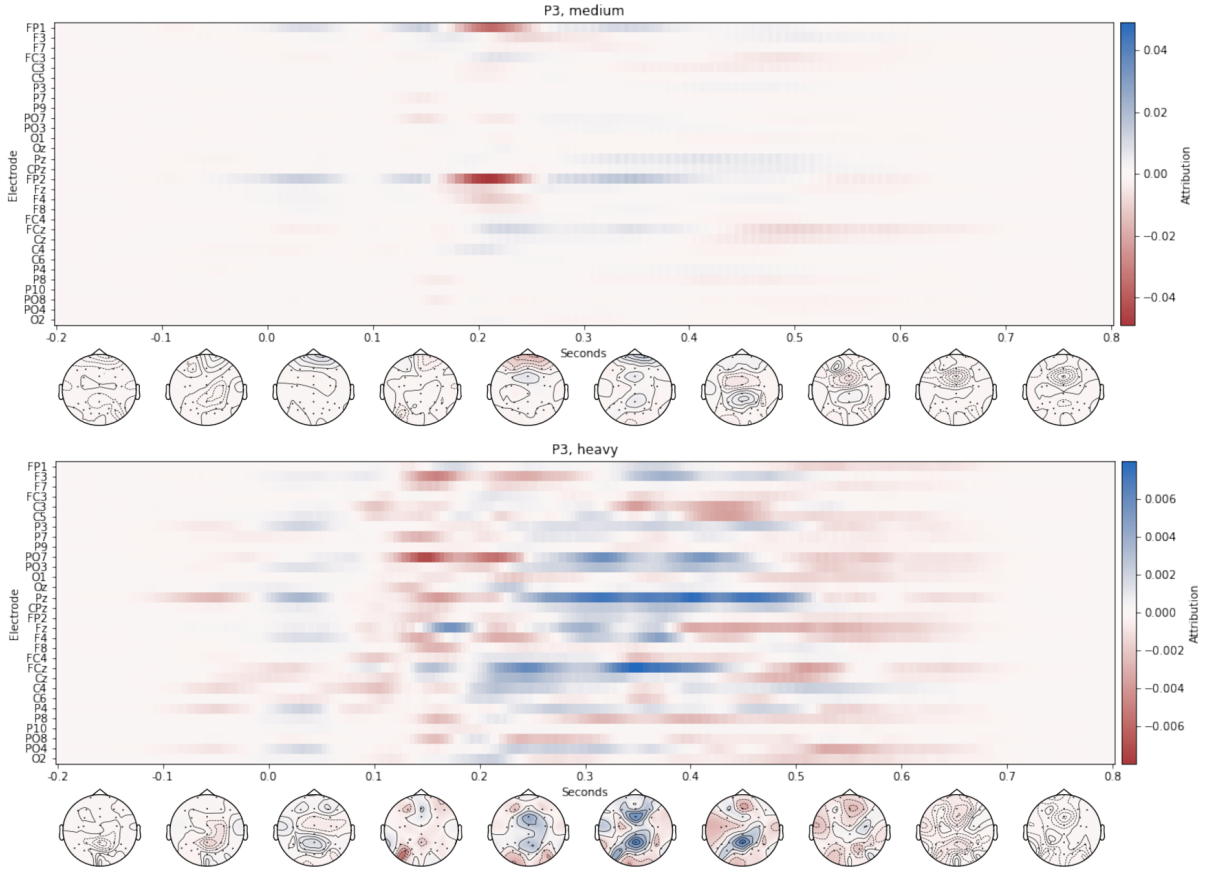

Figure 5.2: Attribution map of the P3 task for medium and heavy preprocessing. Shows average DeepLift attribution of 1000 random trials at all time points and electrodes, together with a topographic map showing the average attribution distribution on the scalp for each 100ms window. The color scale between the plots is very different.

5.4 Models

As the attribution of what is important information for a network is of course highly dependent on the network architecture as well, we looked at how the different deep learning models would handle the same tasks.

Looking at the average attribution of the N400 task of the different deep learning models in Figure 5.3, we can see a very clear difference between the Shallow ConvNet and the two other networks, whereas Deep ConvNet and EEGNet have fewer differences. While the N400 task is a more obvious example of this, it seems that the Deep ConvNet and

EEGNet are able to focus on more specific time points and electrodes compared to the Shallow ConvNet.

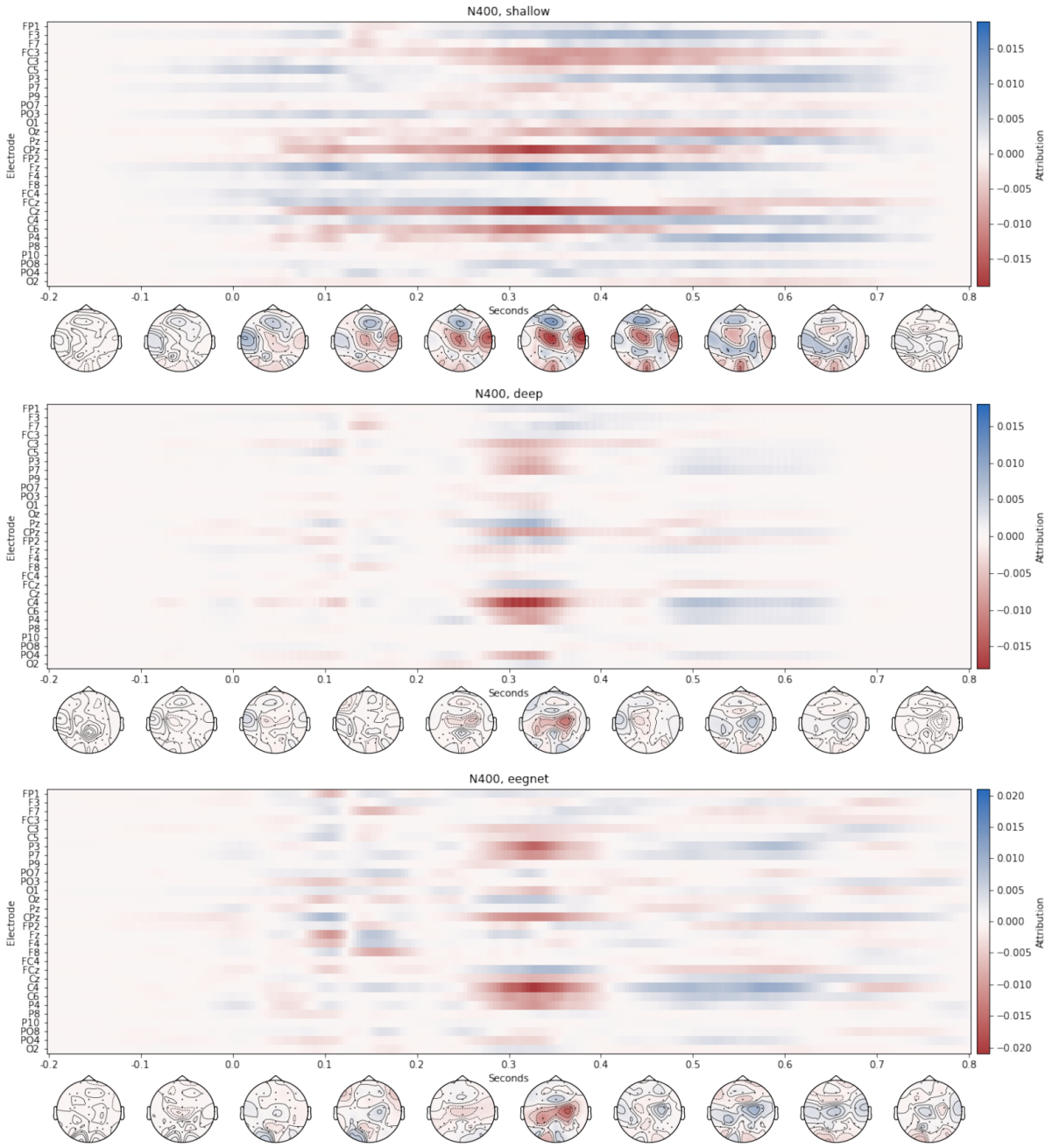

Figure 5.3: Attribution map of the N400 task for the Shallow ConvNet, Deep ConvNet and EEGNet. Shows average DeepLift attribution of 1000 random trials at all time points and electrodes, together with a topographic map showing the average attribution distribution on the scalp for each 100ms window.

5.5 Tasks

In this section, we will look at the average attribution of all trials of the validation set for all tasks in Figure 5.4 and 5.5.

For N170, the peak attribution might seem surprisingly early at around 130ms, that does however line up with the peak of the difference wave, as seen later in Chapter 5.6. Looking at the topography, the electrodes with the highest attribution are P07, PO8, P9 and P10, which lines up with previous research into the N170 component [ZB07].

For the N400 component, there is a high negative attribution around the C4 electrode starting at 250ms and lasting until 350ms. Later at 500ms to 650ms there is a positive attribution for the same electrodes. The location and timeframe of the attribution does roughly fit with previous research [KF11]. While impossible to tell from the average of both conditions, when we look at the average of each condition individually in Figure 5.6, we can see that the mostly positive early and late attributions belong to the true positives, whereas the negative attributions of the true negatives are mostly in the 250ms to 350ms window.

As seen in Chapter 5.3, after the heavy preprocessing, the topography of the P3 task looks plausible compared to previous research [LKF+09]. However, the attribution also seems very noisy, assigning attribution to areas that are implausible, such as before the stimulus gets shown to the subject. This might be the suppressor variable effect explained in Chapter 5.1, or artifacts caused by the temporal convolutions or max-pooling of the Deep ConvNet architecture.

The timeframe and topography of the N2pc attribution again lines up very well with the timing and location of the N2pc component established by previous research [ASW+12].

For the MMN task, the attribution does not seem to align very well with 100-250ms timeframe of previous research [GKSF09], and the electrodes with the highest attribution could also suggest that eye and muscle artifacts played a significant role in the decision-making of the Deep ConvNet. MMN was also the task with the lowest balanced accuracies.

The location of the ERN attribution perfectly matches the expected electrodes for this task [OH08], however there is also significant attribution outside the expected 0ms to 100ms window. The average of both condition hides the very clear positive attribution of the true positives trials in the expected timeframe, as can be seen in Figure 5.6.

For the LRP at -50ms to 0ms, the positive attribution of the C3 electrodes on the left side of the brain corresponds to the right-hand movement, and as expected the same is mirrored for the C4 electrode [LKF+09]. The opposing attribution at the FC3 and FC4

electrodes is harder to interpret, and might be again related to the suppressor variables mentioned in Chapter 5.1.

Overall, the attribution maps fit the expected brain activity by the components very well, and while interpretation of the attribution values is made difficult by the problems outlined in Chapter 5.1, the results do look promising.

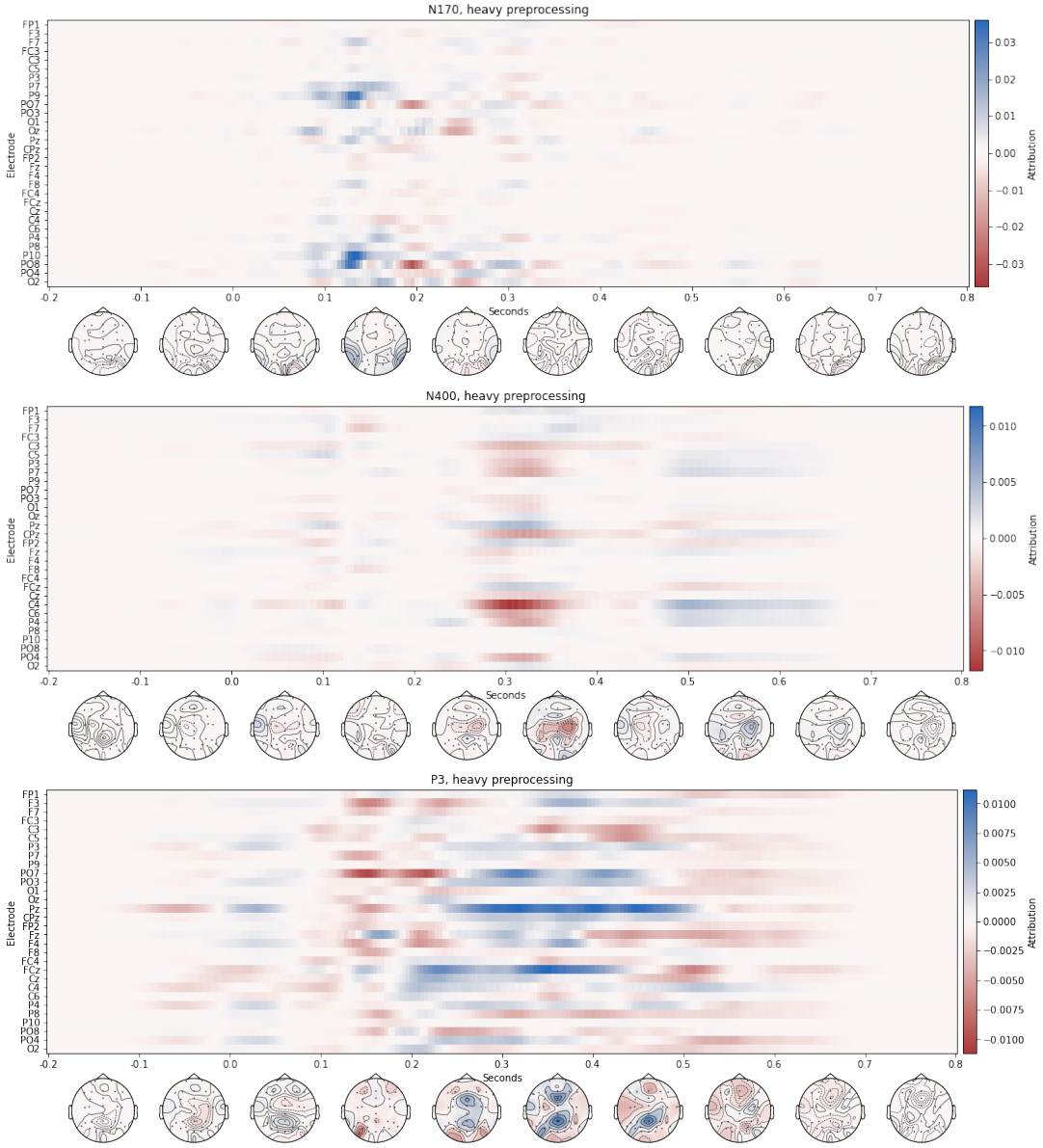

Figure 5.4: Attribution map of the N170, N400 and P3 task for the Deep ConvNet. Shows average DeepLift attribution of all validation set trials at all time points and electrodes, together with a topographic map showing the average attribution distribution on the scalp for each 100ms window.

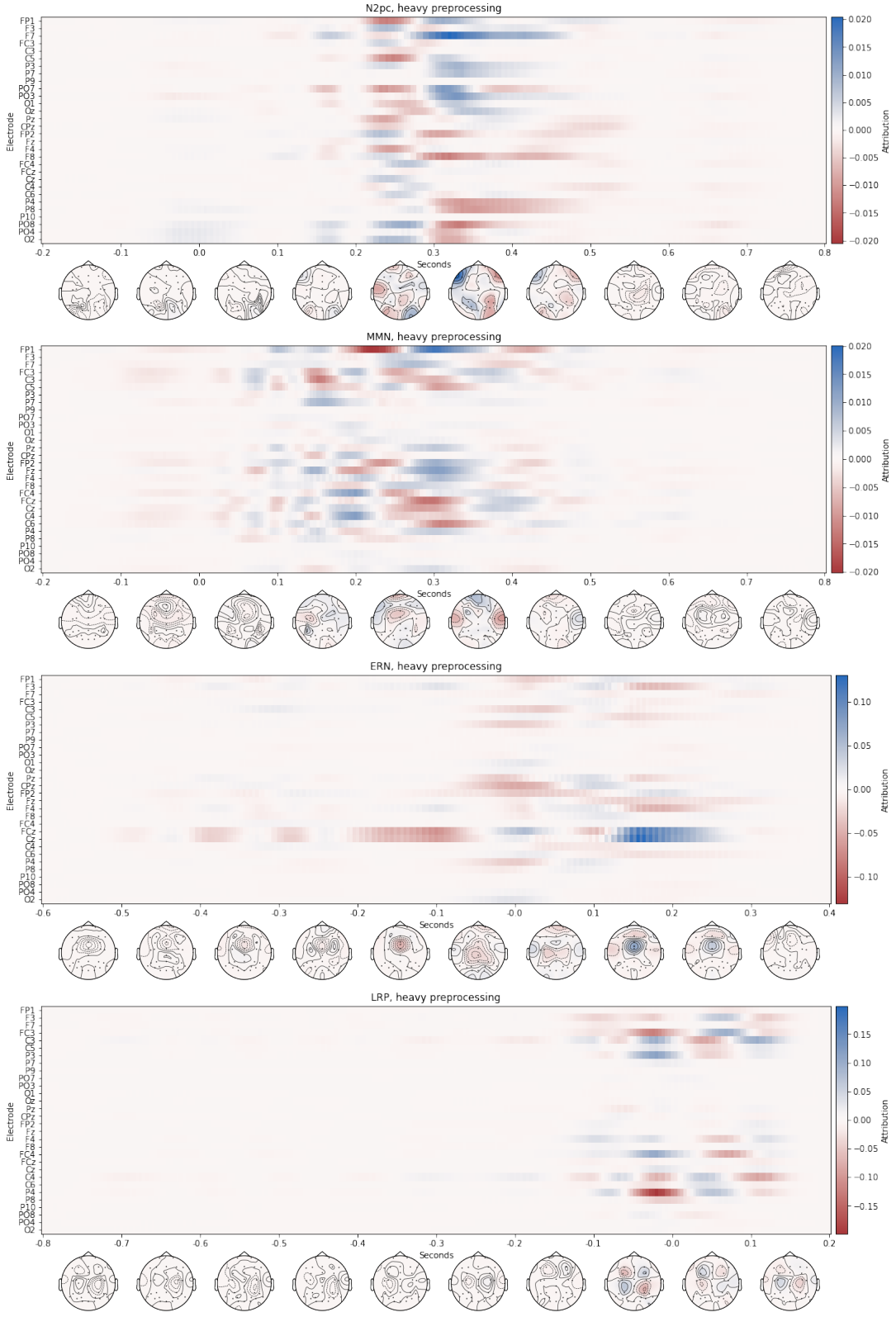

Figure 5.5: Attribution map of the N2pc, MMN, ERN and LRP task. See figure 5.4

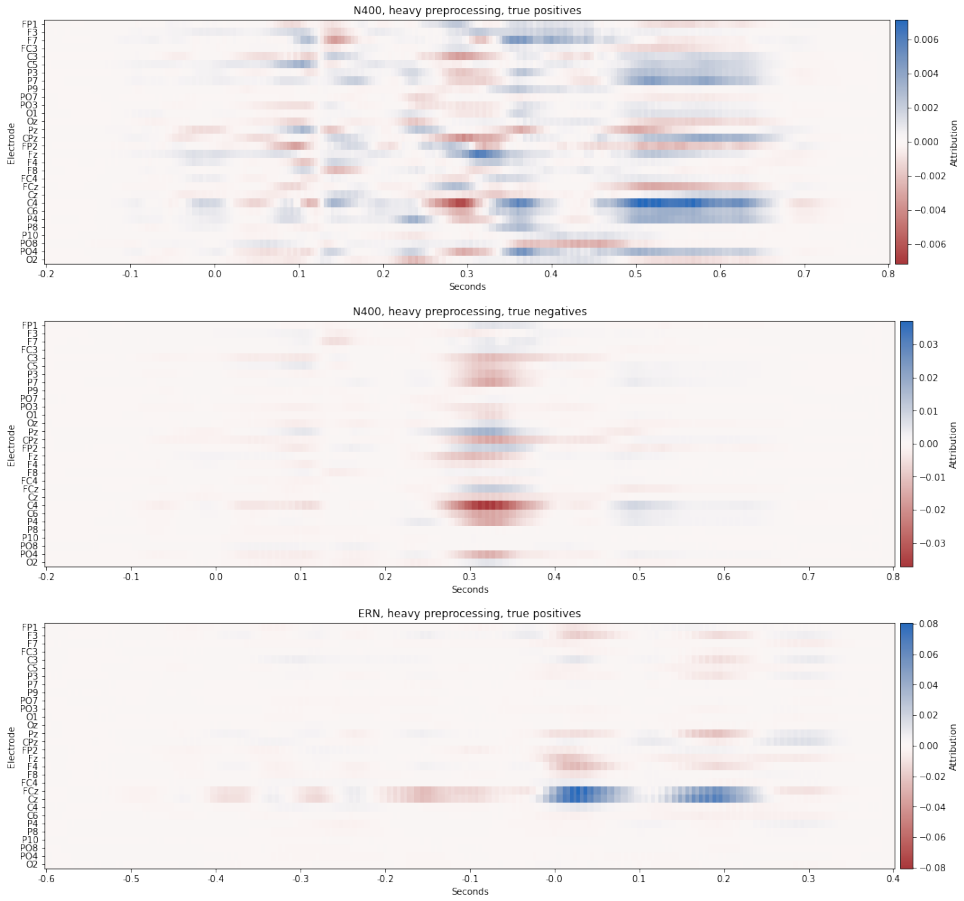

Figure 5.6: Additional attribution maps for only true positives and negative validation trials.

5.5.1 Positive and Negative Predictions

To see the variability of single trials and the difference between positive and negative predictions, we can look at the average attribution over time in Figure 5.7 and 5.8. Each quadrant shows 80 random trials of the two conditions, either correctly or incorrectly classified. For the P3 task, we can see that while there is a lot of variability, the point of maximum positive attribution is between 200ms and 400ms for all trials. Similarly, for LRP two small bands at -100ms and 0ms are clearly visible, even with the variation between trials. Interestingly, the true negative trials for the P3 task do not have flipped signs compared to the true positives, as is the case in the LRP task. This could be because while the LRP task has two distinct responses for the left and right condition, the P3 only has a decreased activity for the negative trials. Another explanation would be that this is related to the DeepLift reference of a zero input, as the negative condition of

P3 is close to the reference. As stated previously, further research into what is a good reference input for EEG data is needed.

This difference between how the conditions are assigned attribution can also be seen in Figure 5.9. Here we can see the average attribution over all true positive and true negative validation trials of the relevant electrodes for P3 and LRP. For the Pz electrode of the P3 task, the peak has a very positive attribution, whereas the lack of a peak in the true negative peak does not get assigned a negative attribution. For the C3 electrode on the left side of the brain, and the C4 electrode on the right side, the attributions between true positives and true negatives are sign flipped. This indicates that, a negative voltage on the C3 electrode at around 100ms is information for the trial being positive, whereas negativity of the C4 electrode at the same time would be information against the trial being positive.

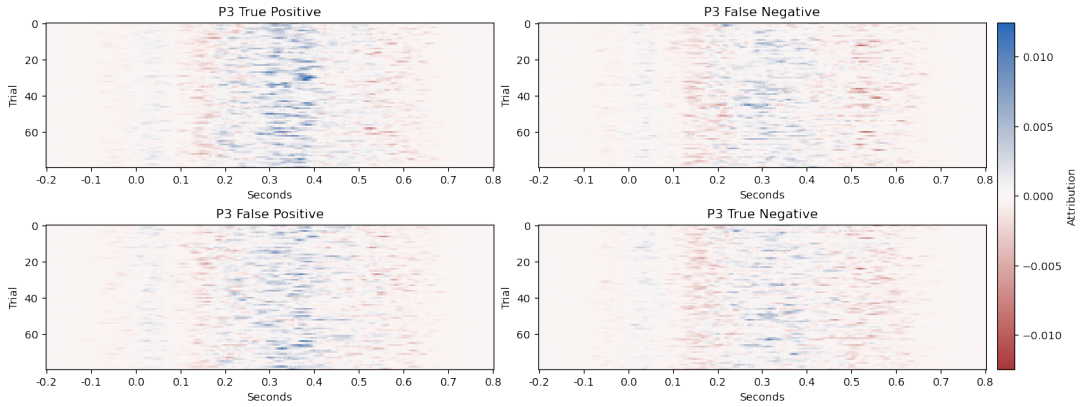

Figure 5.7: Average attribution over time for 80 random true/false positive/negative trials of the P3 validation data.

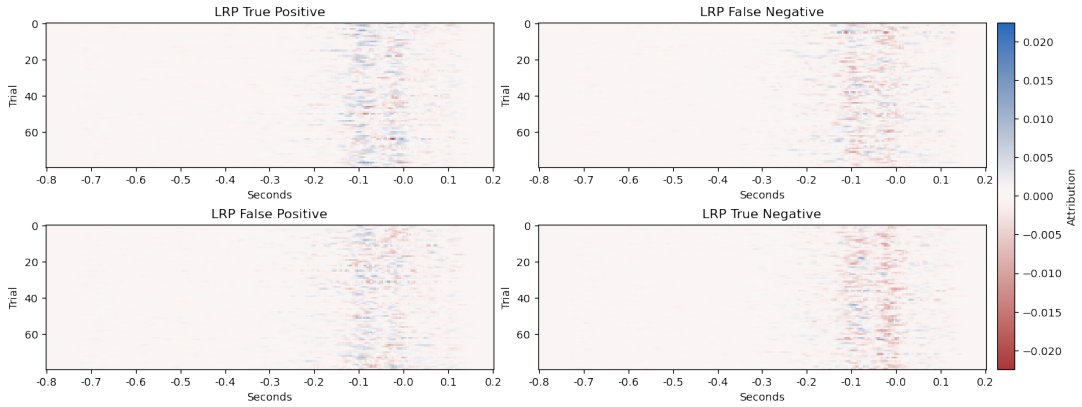

Figure 5.8: Average attribution over time for 80 random true/false positive/negative trials of the LRP validation data.

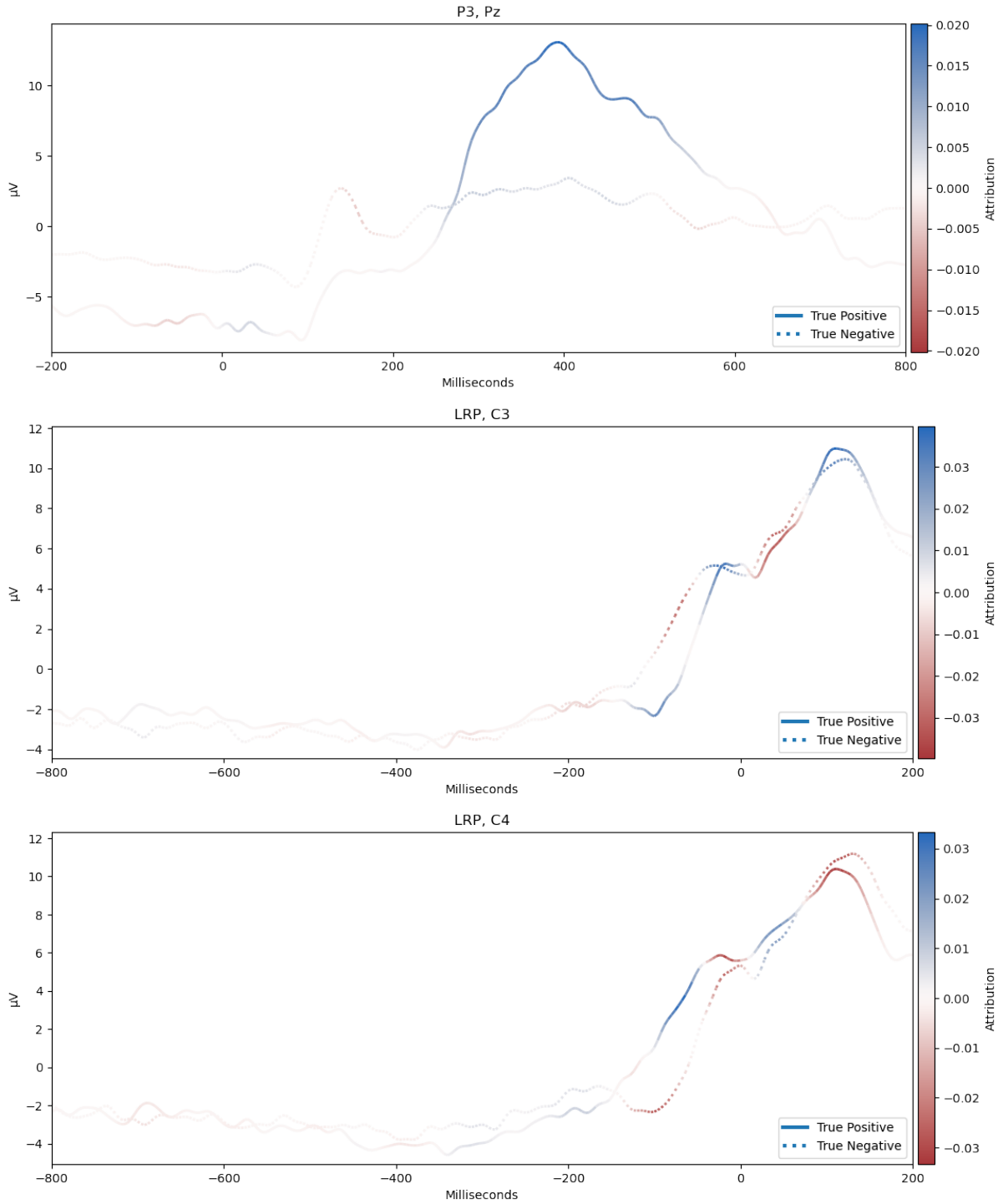

Figure 5.9: Line plots showing the average voltage and attribution of all true positive and true negative trials of the validation data for a single electrode. Top plot shows the Pz electrode of the P3 task. Bottom two plots show the C3 and C4 electrode of the LRP task.

5.6 Difference Waves

Figures 5.10 and 5.11 show the difference wave as well as the attribution of all validation trials for one electrode of each task. The electrodes were chosen to be the same as in the ERP CORE paper [KFZ+21]. While for some tasks, like the N400 and N2pc, the attribution follows the difference wave closely, for other tasks, like ERN and LRP, the attribution diverges visibly from the difference wave. This divergence could be explained by the model finding non-linear or multivariate effects. For other tasks, like N170, MMN and ERN, the difference wave and attribution are sign flipped because the positive condition shows up as a negative voltage on the electrode. Interestingly, there is also a significant positive attribution at 0ms to 100ms of the P3 task, where the difference wave does not show any effect.

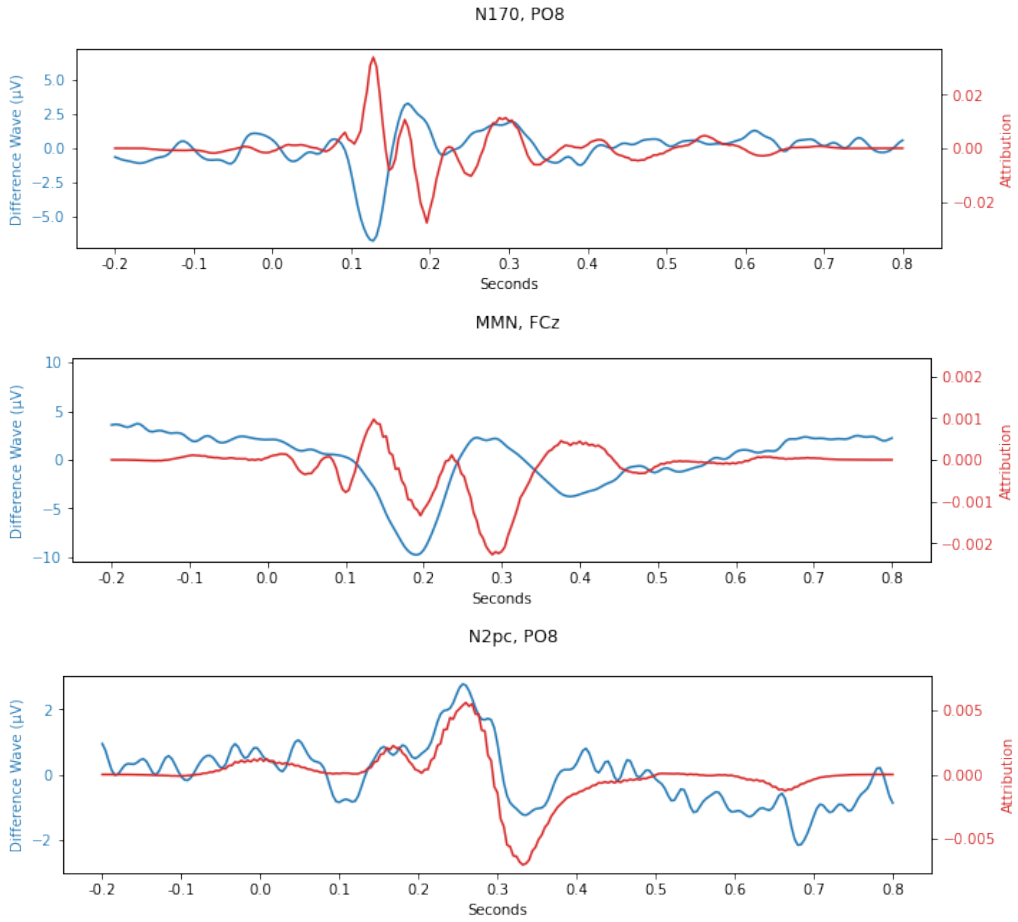

Figure 5.10: Line plots showing the non-baseline corrected difference wave between the true positives and true negatives of the validation data for each task, as well as their attribution on a single relevant electrode.

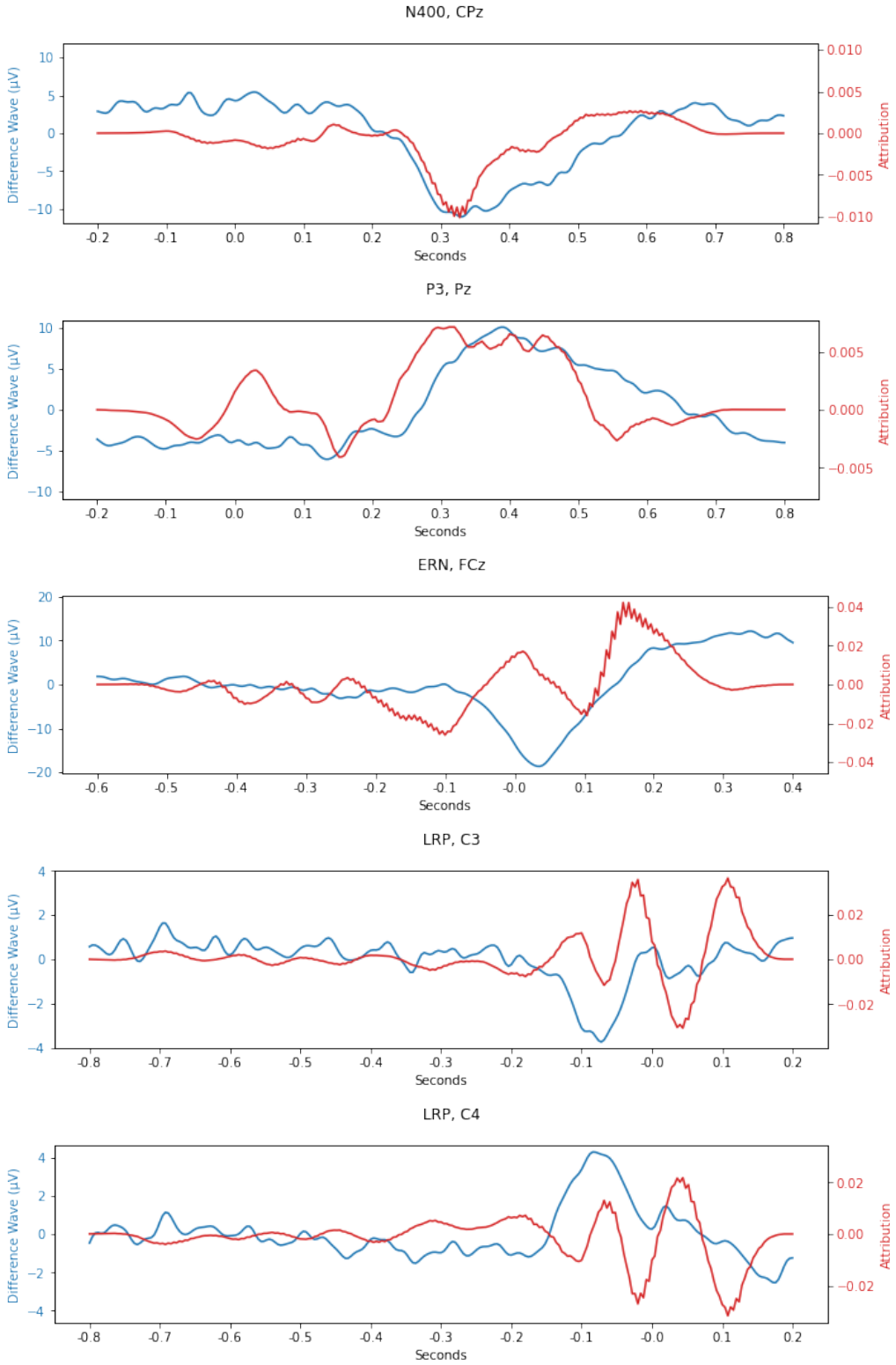

Figure 5.11: Line plots showing the non-baseline corrected difference wave between the true positives and true negatives of the validation data for each task, as well as their attribution on a single relevant electrode.

5.7 Subject Comparison

In order to see how the attribution changes between subjects, we looked at the average attribution over time and all electrodes in Figure 5.12. For the cross-subject data in the top two plots, we can see that the temporal and spatial distribution of the average attribution is roughly the same for all subjects, with only a few outliers. This is not surprising, as for the cross-subject case, the models train on mostly the same data.

For the within-subject case in the bottom two plots, we are looking at the average attribution of all validation trials of a single model for each subject that was trained on 80% of its trials. As the within-subject case had a high variance in balanced accuracy between splits, this data might be too noisy to draw conclusions from. In addition, due to the absolute differences in maximum attribution the color maps of the plots had to be adjusted, making visual comparisons harder. Still, when looking at the bottom right plot, we can still see similarities to the attribution distribution between the electrodes of the cross-subject case. This is at least an indication that even in the within-subject case the model places importance on the same electrodes.

5.8 Conclusion

The attributions assigned by DeepLift do mostly highlight regions and timeframes that are compatible with the expected brain activity. This does help to establish some trust into how the deep learning models come to their decisions. It can also help reveal problems, like the possible reliance on eye artifacts in the P3 tasks for medium preprocessing in Chapter 5.3.

However, for cognitive science, feature attribution methods are held back by the fact that attribution values can not be interpreted as brain activations, as described in Chapter 5.1.

To find out if there are more suitable explainable artificial intelligence methods, a rigorous evaluation on a synthetic EEG dataset, which would provide ground truth data, might be required.

Nevertheless, the attributions assigned by DeepLift provided a different view onto the ERP components and revealed some interesting aspects about how they are decoded by deep learning models.

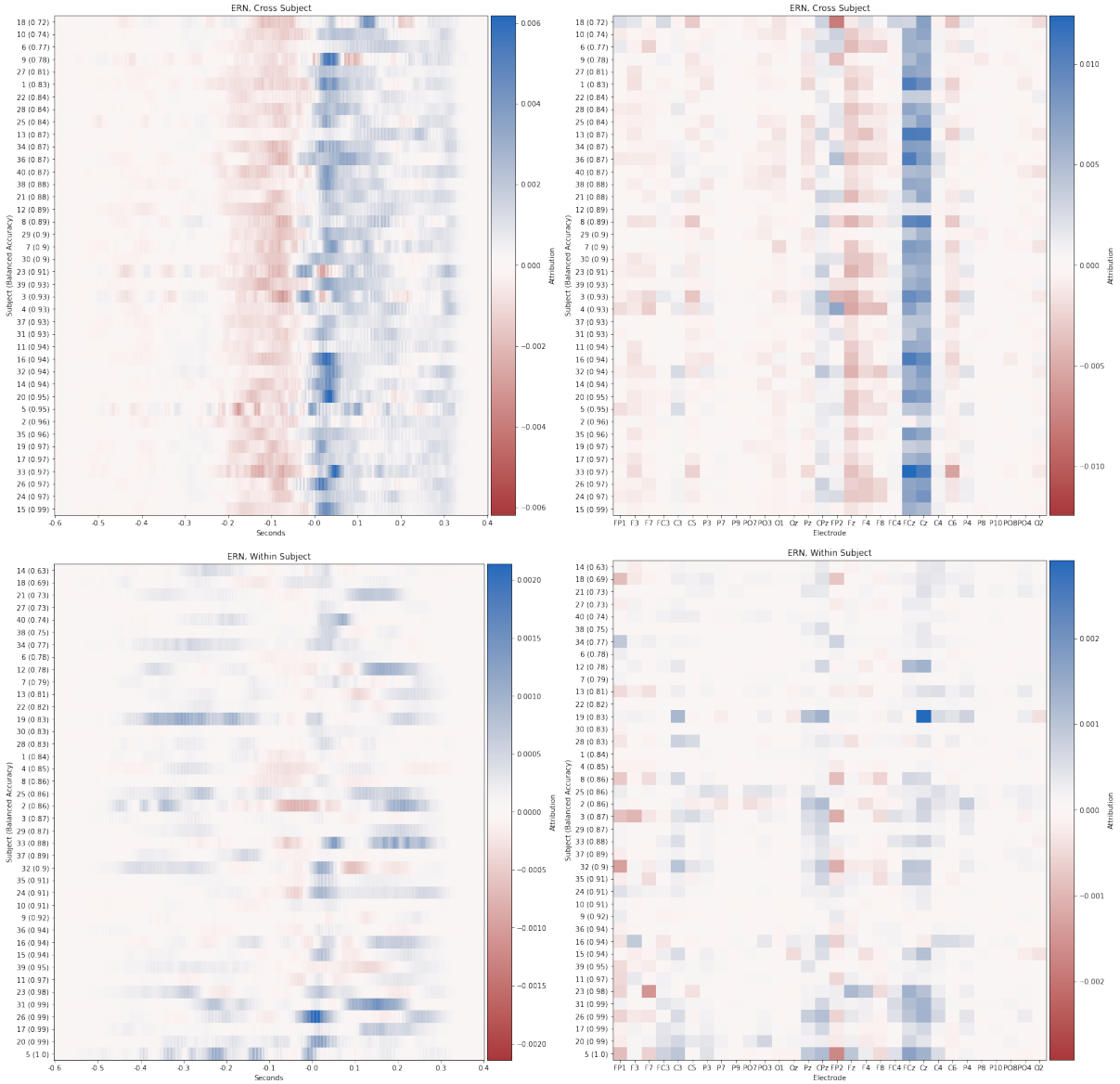

Figure 5.12: Attribution distribution of all subjects in the ERN task, sorted by balanced accuracy. Top plots show cross-subject case, the bottom plots show the within-subject case. Left plots show the average attribution over time, right plots show the average attribution over the electrodes.

6 Discussion

The goal when starting this thesis was to see how deep learning decoding methods and explainable artificial intelligence could be used for cognitive science. By looking at the seven different ERP components of the ERP CORE dataset [KFZ+21], we got insights into the decoding performance of deep learning methods. Starting with the model comparison, we showed that EEGNet and Deep ConvNet are relatively similar in performance, and that the different tasks have widely varying decoding accuracies. By repeating the tests on smaller subsets of the data, we showed that large datasets are required to use these methods. Next, we looked at cross-subject and within-subject performance, where a correlation between of the task performance between the subjects could not be found. This was unexpected and goes against the concept of "BCI Illiteracy" of individual subjects, and implies more of a mismatch between decoder, task, and subject. Finally, we looked at explainable artificial intelligence, specifically feature attribution methods, and how they could help by making deep learning models interpretable.

There were many possible additions to the benchmark part of the thesis that we had to forego due to time constraints, such as analyzing more preprocessing methods, other datasets, and additional decoding models. While we tested relatively straightforward convolutional neural networks, there are more advanced deep learning architectures being used for EEG decoding [IHW+20][SJYX21], which could improve the overall decoding accuracy.

The within-subject analysis was limited by the low sample count per subject, which caused high variability in the decoding accuracy. In order to reduce the need for large datasets, data augmentation methods like cropped decoding [SSF+17], or pre-trained models could be used. With the seven different tasks of the ERP CORE dataset, cross-classification [KMG15], might also reveal relationships between the different components.

The biggest limitation of this thesis, was due to the issues outlined in Chapter 5.1. The attribution assigned by many feature attribution methods is hard to interpret, as they can assign attribution to variables that are uncorrelated with the output. As we do not have ground truth data for the underlying brain activity that is recorded by the ERP CORE data, and could only compare to the univariate analysis of the ERP CORE study [KFZ+21]

and other established research into the ERP components, it was difficult to judge the validity of the DeepLift attributions. Potentially the best way to validate XAI methods for cognitive research, would be to establish a synthetic EEG decoding benchmark, where attributions can be compared to ground truth data. Due to time-constraints, we could not use other validation schemes, like analyzing the change in decoding accuracy when randomizing the input or model parameters, or removing features with the highest attribution. Taking into account these limitations, the attributions assigned by DeepLift were still promising, and showed explanations of the decisions by the decoding models, that seem plausible and in line with the expected brain activity.

Visualizing the attribution of many single ERP trials also proved to be time-consuming, due to the amount of possible parameters and the high dimensionality of the data. An interactive visualization tool specifically for visualizing attribution scores of EEG data might be helpful. While we mostly analyzed the average of the attribution from many trials visually, analyzing them with statistical methods might reveal more insights by the distribution of the single trials.

Removing artifacts from the signal seemed to improve explainability of the model, even if it did not impact decoding accuracy, as seen in Chapter 5.3. How exactly noise and artifacts influence deep learning decoding methods, and what it means for cognitive science if a decoder uses information from sources that are not brain signals, like blinks or muscle artifacts, is a topic that requires further research. As stated in Hebart and Baker [HB18], these differences in noise between conditions could also reflect processing strategies by the brain. However, until these factors are better understood, using preprocessing to remove artifacts is most likely preferable.

We looked at well known ERPs that are evoked in established paradigms, in order to have a comparison for our results. However, once the interpretability issues have been sufficiently addressed and deep learning decoders for cognitive science have been better understood, their ability to find complex features could provide insights that traditional univariate methods are unable to detect. Single trial decoding also allows for novel paradigm designs, as it allows for real-time analysis and does not require time-locking the responses. Overall, the potential benefits of deep learning models for cognitive science makes further research into combating their current drawbacks worthwhile.

Bibliography

[AB18] A. Adadi, M. Berrada. “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI).” In: *IEEE access* 6 (2018), pp. 52138–52160 (cit. on p. 27).

[ACÖG17] M. Ancona, E. Ceolini, C. Öztireli, M. Gross. “Towards better understanding of gradient-based attribution methods for deep neural networks.” In: *arXiv preprint arXiv:1711.06104* (2017) (cit. on p. 28).

[ACW+12] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, H. Zhang. “Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b.” In: *Frontiers in neuroscience* 6 (2012), p. 39 (cit. on p. 13).

[AN10] B. Z. Allison, C. Neuper. “Could anyone use a BCI?” In: *Brain-computer interfaces*. Springer, 2010, pp. 35–54 (cit. on pp. 8, 21).

[ASB+19] S. Appelhoff, M. Sanderson, T. L. Brooks, M. van Vliet, R. Quentin, C. Holdgraf, M. Chaumon, E. Mikulan, K. Tavabi, R. Höchenberger, et al. “MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis.” In: *The Journal of Open Source Software* 4.44 (2019) (cit. on p. 10).

[ASD11] Y. Arbel, K. M. Spencer, E. Donchin. “The N400 and the P300 are not all that independent.” In: *Psychophysiology* 48.6 (2011), pp. 861–875 (cit. on p. 22).

[ASW+12] A. An, M. Sun, Y. Wang, F. Wang, Y. Ding, Y. Song. “The N2pc is increased by perceptual learning but is unnecessary for the transfer of learning.” In: *PLoS One* 7.4 (2012), e34826 (cit. on p. 34).

[BBM+15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.” In: *PloS one* 10.7 (2015), e0130140 (cit. on p. 27).

[BSH+09] B. Blankertz, C. Sanelli, S. Halder, E. Hammer, A. Kübler, K.-R. Müller, G. Curio, T. Dickhaus. “Predicting BCI performance to study BCI illiteracy.” In: *BMC Neurosci* 10.Suppl 1 (2009), P84 (cit. on p. 21).

[Cho17] F. Chollet. “Xception: Deep learning with depthwise separable convolutions.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258 (cit. on p. 14).

[CV95] C. Cortes, V. Vapnik. “Support-vector networks.” In: *Machine learning* 20.3 (1995), pp. 273–297 (cit. on p. 16).

[DK17] F. Doshi-Velez, B. Kim. “Towards a rigorous science of interpretable machine learning.” In: *arXiv preprint arXiv:1702.08608* (2017) (cit. on p. 27).

[GKSF09] M. I. Garrido, J. M. Kilner, K. E. Stephan, K. J. Friston. “The mismatch negativity: a review of underlying mechanisms.” In: *Clinical neurophysiology* 120.3 (2009), pp. 453–463 (cit. on p. 34).

[GLL+13] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al. “MEG and EEG data analysis with MNE-Python.” In: *Frontiers in neuroscience* 7 (2013), p. 267 (cit. on p. 10).

[GSC+20] L. A. Gemein, R. T. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, T. Ball. “Machine-learning-based diagnostics of EEG pathology.” In: *NeuroImage* 220 (2020), p. 117021 (cit. on pp. 8, 13, 14).

[HB18] M. N. Hebart, C. I. Baker. “Deconstructing multivariate decoding for the study of brain function.” In: *Neuroimage* 180 (2018), pp. 4–18 (cit. on pp. 7, 16, 28, 46).

[HMG+14] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, F. Bießmann. “On the interpretation of weight vectors of linear models in multivariate neuroimaging.” In: *Neuroimage* 87 (2014), pp. 96–110 (cit. on p. 28).

[IHW+20] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, L. Benini. “EEG-TCNet: An Accurate Temporal Convolutional Network for Embedded Motor-Imagery Brain–Machine Interfaces.” In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2020, pp. 2958–2965 (cit. on p. 45).

[JEB+17] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, A. Gramfort. “Autoreject: Automated artifact rejection for MEG and EEG data.” In: *NeuroImage* 159 (2017), pp. 417–429 (cit. on p. 10).

[KB14] D. P. Kingma, J. Ba. “Adam: A method for stochastic optimization.” In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 15).

[KF11] M. Kutas, K. D. Federmeier. “Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP).” In: *Annual review of psychology* 62 (2011), pp. 621–647 (cit. on p. 34).

[KFZ+21] E. S. Kappenman, J. L. Farrens, W. Zhang, A. X. Stewart, S. J. Luck. “ERP CORE: An open resource for human event-related potential research.” In: *NeuroImage* 225 (2021), p. 117465 (cit. on pp. 8–10, 29, 40, 45).

[KMG15] J. T. Kaplan, K. Man, S. G. Greening. “Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations.” In: *Frontiers in human neuroscience* 9 (2015), p. 151 (cit. on p. 45).

[KMM+20] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. “Captum: A unified and generic model interpretability library for pytorch.” In: *arXiv preprint arXiv:2009.07896* (2020) (cit. on p. 28).

[KSJ+19] B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, T. Jeon. “Why are saliency maps noisy? cause of and solution to noisy saliency maps.” In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 4149–4157 (cit. on p. 27).

[LH16] I. Loshchilov, F. Hutter. “Sgdr: Stochastic gradient descent with warm restarts.” In: *arXiv preprint arXiv:1608.03983* (2016) (cit. on p. 15).

[LKF+09] S. J. Luck, E. S. Kappenman, R. L. Fuller, B. Robinson, A. Summerfelt, J. M. Gold. “Impaired response selection in schizophrenia: Evidence from the P3 wave and the lateralized readiness potential.” In: *Psychophysiology* 46.4 (2009), pp. 776–786 (cit. on p. 34).

[LKK+19] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, S.-W. Lee. “EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy.” In: *GigaScience* 8.5 (2019), giz002 (cit. on p. 26).

[LSW+18] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance. “EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces.” In: *Journal of neural engineering* 15.5 (2018), p. 056013 (cit. on pp. 8, 13, 14, 17, 27, 28).

[Luc12] S. J. Luck. “Event-related potentials.” In: (2012) (cit. on p. 7).

[McK+10] W. McKinney et al. “Data structures for statistical computing in python.” In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56 (cit. on p. 11).

[NFM+19] N. Nagabushan, T. Fisher, G. Malaty, M. Witcher, S. Vijayan. “A comparative study of motor imagery based BCI classifiers on EEG and iEEG data.” In: *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2019, pp. 1–5 (cit. on pp. 8, 13, 27).

[OH08] D. M. Olvet, G. Hajcak. “The error-related negativity (ERN) and psychopathology: Toward an endophenotype.” In: *Clinical psychology review* 28.8 (2008), pp. 1343–1354 (cit. on p. 34).

[PBL19] A. Pedroni, A. Bahreini, N. Langer. “Automagic: Standardized preprocessing of big EEG data.” In: *NeuroImage* 200 (2019), pp. 460–473 (cit. on p. 10).

[PGM+19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035 (cit. on p. 11).

[PKM12] J. A. Palmer, K. Kreutz-Delgado, S. Makeig. “AMICA: An adaptive mixture of independent component analyzers with shared components.” In: *Swartz Center for Computatonal Neursoscience, University of California San Diego, Tech. Rep* (2012) (cit. on p. 10).

[PKM19] L. Pion-Tonachini, K. Kreutz-Delgado, S. Makeig. “ICLabel: An automated electroencephalographic independent component classifier, dataset, and website.” In: *NeuroImage* 198 (2019), pp. 181–197 (cit. on p. 10).

[SAE+19] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim. “Towards a rigorous evaluation of xai methods on time series.” In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 4197–4201 (cit. on p. 27).

[SGK17] A. Shrikumar, P. Greenside, A. Kundaje. “Learning important features through propagating activation differences.” In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3145–3153 (cit. on pp. 8, 27, 28).

[SGL20] L. Sixt, M. Granz, T. Landgraf. “When explanations lie: Why many modified bp attributions fail.” In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9046–9057 (cit. on p. 27).

[SJYX21] Y. Song, X. Jia, L. Yang, L. Xie. “Transformer-based Spatial-Temporal Feature Learning for EEG Decoding.” In: *arXiv preprint arXiv:2106.11170* (2021) (cit. on pp. 8, 13, 45).

[SLSM16] I. Sturm, S. Lapuschkin, W. Samek, K.-R. Müller. “Interpretable deep neural networks for single-trial EEG classification.” In: *Journal of neuroscience methods* 274 (2016), pp. 141–145 (cit. on p. 27).

[SSF+17] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball. “Deep learning with convolutional neural networks for EEG decoding and visualization.” In: *Human brain mapping* 38.11 (2017), pp. 5391–5420 (cit. on pp. 8, 13–15, 27, 45).

[Tho19] M. C. Thompson. “Critiquing the concept of BCI illiteracy.” In: *Science and engineering ethics* 25.4 (2019), pp. 1217–1233 (cit. on p. 21).

[TRP21] A. W. Thomas, C. Ré, R. A. Poldrack. “Challenges for cognitive decoding using deep learning methods.” In: *arXiv preprint arXiv:2108.06896* (2021) (cit. on pp. 8, 27).

[VB10] C. Vidaurre, B. Blankertz. “Towards a cure for BCI illiteracy.” In: *Brain topography* 23.2 (2010), pp. 194–198 (cit. on p. 21).

[VMS+20] A. Vahid, M. Mückschel, S. Stober, A.-K. Stock, C. Beste. “Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control.” In: *Communications biology* 3.1 (2020), pp. 1–11 (cit. on p. 27).

[VRE+17] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, B. Thirion. “Assessing and tuning brain decoders: cross-validation, caveats, and guidelines.” In: *NeuroImage* 145 (2017), pp. 166–179 (cit. on p. 16).

[WBMH21] R. Wilming, C. Budding, K.-R. Müller, S. Haufe. “Scrutinizing XAI using linear ground-truth data with suppressor variables.” In: *arXiv preprint arXiv:2111.07473* (2021) (cit. on p. 28).

[ZB07] E. Zion-Golumbic, S. Bentin. “Dissociated neural mechanisms for face detection and configural encoding: evidence from N170 and induced gamma-band oscillation effects.” In: *Cerebral Cortex* 17.8 (2007), pp. 1741–1749 (cit. on p. 34).

[ZLLG21] B. Zang, Y. Lin, Z. Liu, X. Gao. “A deep learning method for single-trial EEG classification in RSVP task based on spatiotemporal features of ERPs.” In: *Journal of Neural Engineering* 18.4 (2021), p. 0460c8 (cit. on pp. 8, 13).

All links were last followed on January 23, 2021.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature