# Investigating EEG Biomarkers for Attention with Deep Learning: Final Report
## Project Category: Healthcare

**Ziad Ali, Tolga Celik, and Kevin Chen**
Department of Electrical Engineering
Stanford University
ziadaali@stanford.edu, otcelik@stanford.edu, and kevinch@stanford.edu

## Abstract

Electroencephalography (EEG) enables researchers to non-invasively measure semi-localized brain activity and correlate that activity with specific cognitive functions. We investigate how deep learning models can be applied to the study of attention and the extent to which electrophysiological biomarkers obtained from EEG signals can be measured and used to classify a person's state of attentiveness. We determined that artificial neural networks are able to learn an individual's EEG characteristics well enough to perform binary classification tasks but struggle with learning population-wide datasets. Conversely, more complex CNN and RNN architectures are able to better process multi-subject data, but struggle with single-subject data. We also determined that the frontal cortex, parietal lobe, and the occipital lobe regions of the brain are most highly correlated with attention.

## 1 Introduction

Electroencephalography (EEG) is a non-invasive measurement technology to electrically record brain activity. EEG measurements have high temporal resolution and variable (medium to low) spatial resolution, depending on the density of the electrode net used. EEGs are well-suited to enable researchers to study cognitive activity, including attention, given the ease with which they can obtain large amounts of data from subjects. Our interests lie in applying different deep learning models to EEG data in order to classify and predict the state of a subject's attentiveness. Explicitly, the inputs to our system are the recorded EEG time-series voltage data (which consists of one channel for each electrode used) one second prior to some stimulus onset. The outputs of our neural networks predict information about subjects' attention levels, measured based on reaction time to stimuli, and this output can be a classification of the relative speed of the reaction time or an analog prediction of the subject's exact reaction time. Our aims then extend to investigating which deep learning networks are best suited to handle EEG data as well as analyzing the effects of restricting the number of EEG recording channels to demonstrate the merit of extrapolating attention information from a smaller form-factor, (import for reference-less EEG and function localization). An application of these concepts of focus would be a miniature EEG patch placed on the head that would be able to record and report subject attention levels.

## 2 Related work

Attention is a common research area linked to EEG, and studies have been conducted to analyze several event-related potentials (ERPs, e.g. P1, N1, P3, etc.) to extrapolate subject attentiveness. It is believed that the alpha (8-12 Hz) and theta (4-8 Hz) EEG frequency bands are most linked to the allocation of attentional resources [1]. There also exist hypotheses about the specific phase of

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

EEG components affecting the development of an ERP component, thus affecting the corresponding reaction time [2]. As such, both time- and frequency-domain information could be considered as inputs to our model.

Deep learning (DL) for EEG is most commonly associated with classification tasks, such as sleep scoring or seizure/anomaly detection. Popular choices for architecture include CNNs and RNNs, to leverage the inherent structure of EEG data, both spatially and temporally. Because of the peculiarities of EEG processing, including low SNR and high inter-subject or inter-trial variability, pre-processing is fairly common, and can include artifact removal (e.g. eye blinks), filtering, interpolation, and feature extraction. However, raw EEG data is still commonly used for the input to the DL model.

There are few architectures typically chosen for DL using EEG data. Among previous DL EEG studies, 53% employ CNNs or hybrid CNNs, and either directly input raw signals or spectral information [3]. Similarly, 18% of studies employ Deep Belief Networks and 10% use RNNs [3]. It is also typical to choose between 3-5 convolutional layers or 1-2 LSTM layers [3]. In particular, Alhagry et al. report 85% accuracy using two LSTM layers with dropout for an emotion recognition dataset [4]. It is much less common to use GRUs or a combination of GRU and CNNs.

## 3 Dataset and Features

The dataset we used comes from an experiment in which EEG (electroencephalogram) signals were recorded from different subjects during a simulated driving setup [5]. Subjects spent an average of 90 minutes in a virtual reality environment and were instructed to keep a car centered in a specific lane of the road. At random intervals, the car would begin to drift left or right, and the subject would need to correct the car's position to re-center it in its lane. The data was collected using a 32-channel EEG sampled at 500 Hz, 16-bit resolution, with Ag/AgCl electrodes placed in a modified 10-20 system (2 reference channels - Figure A.1) [6]. In addition to the raw EEG signals, the dataset also includes labelled time indices for the drift events and corresponding correction responses. The experiment had 27 unique subjects and 62 total sessions. Fig. A.2 depicts a visualization of the data.

The dataset is labelled with 4 different types of events: "Deviation onset left" (251), "Deviation onset right" (252), "Response onset" (253), "Response offset" (254). These events respectively occur 13522, 13670, 27192 and 27192 times over a total of 93 hours of recordings. The reaction time of the driver is also included in the dataset, and is defined as the time difference between deviation onset and response onset. The large number of events in the dataset suggest an accurate general deep model should be attainable (if biomarkers are common across subjects), but the limited number of events per subject (on the order of several thousand each) indicate training individual models may be difficult.

The average median response time was 1.037 s across all drivers with a standard deviation of 0.478 s. Histograms showing the reactions times for subjects 44 and 41 are depicted in Fig. A.3.

We extract signal sequences of uniform time duration from immediately prior to the deviation onset to feed into our models that dealt with the time-domain data. For some models, rather than utilizing all data points from our 1-second EEG signal preceding stimulus onset, we preprocessed all such signals to extract the relative band-powers within 5 frequency bands as compared to the whole signal, with the intention of lowering the dimensionality of our input data for easier model learning. See Appendix B for details, and A.4 for a plotted example.

## 4 Methods

### 4.1 Time Domain

In our raw signal analysis, we use a combination of CNNs (convolutional neural networks), RNNs (with GRU layers), and combined CNN-RNNs to process data. We combined CNNs and RNNs because CNNs tend to uncover (in our case) temporal relationships (given the nature of our signals) that RNNs can make use of. We also incorporated a spectrogram layer as part of our CNN architecture. Spectrograms apply Fourier transforms using a rolling time window to data to generate a 2D plot of frequency vs. time, and are used to analyze the frequency content of data in event detection problems [7]. An example network using LSTMs for trigger word detection was given in the course lectures. The 2D image generated by the spectrogram is well-suited for processing by a CNN.

We used these networks to perform both binary classification and continuous-value estimation. For binary classification we used a binary cross-entropy loss function (Equation 1) whereas for continuous-value estimation we used a smooth L1 loss function (Equation 2 and 3). Binary cross-entropy loss is designed such that only one term of the equation is ever evaluated (given that $y_i$ is always either 0 or 1). If the evaluated term has diametrically opposed values for $y_i$ and $\hat{y}_i$ the loss is high; otherwise, the loss is almost negligible. Smooth L1 loss is a loss function that uses mean-squared error for a narrow region of interest of the data and L1 loss (average of magnitude of errors) for outliers. This was used because the nature of the response times is such that most are concentrated around 1s but others extend in time up to 10+ seconds, and we are not concerned with accurately predicting these long reaction times precisely.

$$\text{BCE}_{Loss}(y, \hat{y}) = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y_i}) + (1 - y_i) \cdot \log(1 - \hat{y_i}) \tag{1}$$

$$\text{SL1}_{Loss}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N} \text{smooth}_{L_1}(y_i - \hat{y_i}) \tag{2}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3}$$

Our best performing model scheme is shown in Fig. 1. A GRU was used instead of an LSTM because GRUs learn faster and we did not have enough data to train a complex LSTM model. Before choosing this setup, we tried using LSTMs with various numbers of layers with and without convolutional layers. This model was trained for regression on the response times, and was evaluated as a binary classifier with respect to the median response time at test stage as shown in Fig. A.5.

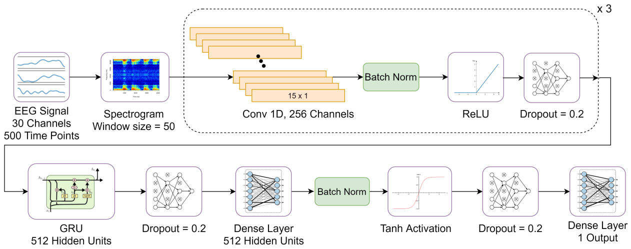

Figure 1: Diagram of the Neural Network Model combining GRU and CNNs.

### 4.2 Frequency Domain

As previously described, periodograms were used to compute band powers for 5 bands from 0-30 Hz. This serves as a reduction in dimensionality and a general frequency-domain representation of our time-domain signal. These 5 data points were extracted for all 30 EEG channels and flattened into a single 150-element (for 30 channels) input vector for a standard artificial neural network (ANN).

The ANN's architecture consisted of an input layer, 2 hidden layers, and an output layer, with batch normalization and ReLU outputs applied at both hidden layers and dropout applied at the second hidden layer. A sigmoid function was used to calculate the output, and binary cross entropy was used for the loss function.

The ANN was trained on an individual subject's data (subject 44) as well as the entire dataset. In addition the ANN was trained separately for 6 clusters (A-F) of 4 electrodes each as seen in Figure A.1. This was performed to determine whether the bulk of "attention information" could be localized to a specific region of the brain. Lastly, the ANN was also trained on differential cluster data in which one electrode of the four became a "reference" and the signals from the other three were modified to be the difference between those electrodes and the reference.

We evaluated all binary classification models using accuracy, precision, recall, and F1 score. We evaluated the regression models using mean-squared error.

# 5 Experiments/Results/Discussion

## 5.1 Time Domain

The time domain system was particularly designed for added complexity as our initial models failed to reach high training accuracy even with extended periods of training. The spectrogram, repeated convolutional layers, GRU and the repeated dense layer are incorporated for this purpose. The tanh activation was preferred before the final dense layer (instead of a ReLU or other activation) because the regression results tended towards higher values, and we chose to provide the network with an activation that can yield negative values. The dropout layers were added to prevent overfitting to the training set, due to the added complexity, and 0.2 dropout probability yielded the best regularization. Batch normalization layers were used, and a GRU was preferred over an LSTM to improve the network's performance on limited data. Learning rate decay was used in training, starting from $\alpha = 10^{-3}$ and reducing the learning rate by a factor of 0.9 once the validation cost reached a plateau for two consecutive epochs. The batch size was chosen as 64 samples to speed up training.

The hyperparameters are given in Table 1; results for classification are provided in Table 2, and the confusion matrix is given in Table 3. Class 0 corresponds to response times shorter than the median time, and Class 1 corresponds to longer responses times. We also report MSE for regression although we use regression only to classify the level of attention. The results show that the model can predict the attention level of an individual from data before the event onset 35% better than the baseline of 0.5 (evenly split data). There is negligible bias towards either class.

| | |
|---|---|
| Learning Rate | $10^{-3}$ |
| Conv 1D Layer 1 Channel # | 256 |
| Conv 1D Layer 2 Channel # | 256 |
| Conv 1D Layer 3 Channel # | 256 |
| GRU Hidden Unit # | 512 |
| Optimizer | Adam |
| Dropout Probability (Global) | 0.2 |
| Batch Size | 64 |
| Number of Epochs | 200 |
| Pre-onset time analyzed | 1 second |

Table 1: Time domain CNN+RNN hyperparameters

| | |
|---|---|
| Test Accuracy | 0.674 |
| Precision | 0.678 |
| Recall | 0.677 |
| F1 Score | 0.678 |
| Regression MSE | 0.444 |
| Training Accuracy | 0.809 |

Table 2: Time domain CNN+RNN results

| True/Predicted | Class 0 | Class 1 |
|---|---|---|
| Class 0 | 1684 | 823 |
| Class 1 | 828 | 1737 |

Table 3: Confusion Matrix

## 5.2 Frequency Domain

The hyperparameters for the ANN were the learning rate, hidden units per layer, dropout probability, number of training epochs, and optimizer function (stochastic gradient descent vs. Adam). A random search for all hyperparameters was used to find the optimal ANN settings (that maximize classification accuracy). Hidden units were maximized at 2000, the learning rate was between 0.1 and 1e-6, dropout was between 0 and 0.9, and epochs were between 5 and 30. The number of layers was also varied, but it was determined early on that 2 hidden layers was optimal, so it was not included in the random search.

The results for the 4 different examined model use-cases are shown in Tables 4-7. The highest performing model for subject 44 had 1657 units in layer 1, 1574 in layer 2, a 0.67 dropout rate, a learning rate of 2.2e-5, and 21 epochs. The results for this model are shown in the row labeled 'All' in Table 4. The accuracy baseline is 50% (classification categories are split by median), so this model (which attained an accuracy of 77.7%) was relatively successful at learning a single subject's attention biomarkers (55% improvement). However, the highest performing cross-subject model attained an accuracy of only 63.6%. This demonstrates that while frequency domain features may be unique to individuals, they do not generalize well across subjects - the RNN and CNN models perform better for multi-subject test sets.

With regards to clusters, the best performing cluster was A, corresponding to the frontal cortex. This makes sense as the frontal cortex is hypothesized to be the site of attention and decision making processes in the brain. Cluster D also performed well for cross-subject models; cluster D corresponds to the occipital lobe, which is the site of vision processing in the body. Lastly, cluster C also performed well in the cross-subject models; cluster C corresponds to the parietal lobe, another region of the brain thought to be responsible for attention. While certain clusters performed better than others, the differences between clusters are relatively small - this suggests attention may be more of a 'whole-brain' process than we initially assumed.

The differential cluster models performed nearly as well for subject 44 as the common reference models, and just as well (and better in some cases) for the cross-subject models. This suggests that a patch electrode system localized to only a single area of the brain could obtain useful measurements even without a distant reference electrode. However, like the larger 30-channel model, both the common reference and differential reference cluster models had better performance in the single-subject case, and did not seem to generalize well across subjects. Again, this is likely due to the nonstationary nature of EEG, and it is possible that our best models may not generalize well across cognitive conditions, even among the same subject. There was insufficient data to draw any conclusions in this regard, but the overall performance of the frequency-domain models were satisfactory.

| Cluster | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| A | 0.7511 | 0.7794 | 0.6943 | 0.7344 |
| B | 0.7121 | 0.7927 | 0.5677 | 0.6616 |
| C | 0.7078 | 0.7611 | 0.5983 | 0.6699 |
| D | 0.7316 | 0.7838 | 0.6332 | 0.7005 |
| E | 0.7251 | 0.7656 | 0.6419 | 0.6983 |
| F | 0.6991 | 0.7557 | 0.5808 | 0.6568 |
| **All** | **0.7771** | **0.7944** | **0.7424** | **0.7675** |

Table 4: Frequency domain ANN electrode cluster results (subject 44, common reference)

| Cluster | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| A | 0.7381 | 0.7903 | 0.6419 | 0.7084 |
| B | 0.7229 | 0.7440 | 0.6725 | 0.7064 |
| C | 0.7381 | 0.8333 | 0.5895 | 0.6905 |
| D | 0.7208 | 0.8086 | 0.5721 | 0.6701 |
| E | 0.7165 | 0.7634 | 0.6201 | 0.6843 |
| F | 0.7164 | 0.7356 | 0.5581 | 0.7002 |

Table 5: Frequency domain ANN electrode cluster results (subject 44, differential reference)

| Cluster | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| A | 0.5880 | 0.5991 | 0.5311 | F1: 0.5631 |
| B | 0.5899 | 0.6153 | 0.4793 | 0.5388 |
| C | 0.5916 | 0.6156 | 0.4873 | 0.5440 |
| D | 0.6045 | 0.6211 | 0.5353 | 0.5750 |
| E | 0.5845 | 0.5955 | 0.5265 | 0.5589 |
| F | 0.5824 | 0.5986 | 0.4996 | 0.5447 |
| **All** | **0.6355** | **0.6567** | **0.5675** | **0.6089** |

Table 6: Frequency domain ANN electrode cluster results (all subjects, common reference)

| Cluster | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| A | 0.5814 | 0.5950 | 0.5096 | 0.5490 |
| B | 0.5690 | 0.5758 | 0.5230 | 0.5482 |
| C | 0.5963 | 0.6087 | 0.5430 | 0.5736 |
| D | 0.6046 | 0.6239 | 0.5265 | 0.5711 |
| E | 0.5895 | 0.6058 | 0.5119 | 0.5549 |
| F | 0.5707 | 0.5866 | 0.4781 | 0.5268 |

Table 7: Frequency domain ANN electrode cluster results (all subjects, differential reference)

When used for the attention classification task, the time domain model gives more accurate results than the frequency domain model; however, the time domain model is complex and takes longer to train repeatedly with different clusters, so the frequency domain model gives more insight on the relevance of a subset of the EEG channels for the attention task. Both models suffered from overfitting, likely due to the sparsity of data and the need to use complex models which created a higher order function space.

## 6 Conclusion/Future Work

After experimentation with different model architectures, we achieved the best results using two general models, taking in time-domain-based and frequency-domain-based inputs respectively. Our models were able to predict reactions times, as a measure of attention, based on one second of data prior to the stimulus onset with good performance. Our frequency-domain model worked especially well for single-subject attention prediction, while our time-domain model was better at generalizing attention across different subject data. In the future, we believe that improvements may be made upon the input data time window, extending beyond one second before the deviation. Furthermore, our ability to experiment with some model types were limited by the amount of data that was available, One direction of interest is to collect higher-order data to classify attention, as opposed to simply predicting a reaction time, in order to test the viability of models to analyze deeper facets of attention.

# 7 Contributions

- Tolga Celik: Implementation, tuning and testing of the time domain model
- Ziad Ali: EEG signal conditioning and processing, ANN implementation, frequency domain processing, cluster models (equally split with Kevin)
- Kevin Chen: EEG signal conditioning and processing, ANN implementation, frequency domain processing, cluster models (equally split with Ziad)

# References

[1] Tim Lomas, Itai Ivtzan, and Cynthia Fu. A systematic review of the neurophysiology of mindfulness on EEG oscillations. *Neuroscience and biobehavioral reviews*, 57, 10 2015.

[2] Wolfgang Klimesch. $\alpha$-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, 16, 11 2012.

[3] Alexander Craik, Yongtian He, and José Contreras-Vidal. Deep learning for Electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16, 02 2019.

[4] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *Emotion*, 8(10):355–358, 2017.

[5] Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific Data*, 6(1):19, Apr 2019.

[6] St Louis Erik K., Lauren C. Frey, and Jeffrey W. Britton. *Electroencephalography (EEG): an introductory text and atlas of normal and abnormal findings in adults, children, and infants*. American Epilepsy Society, 2016.

[7] H. Zhang, I. McLoughlin, and Y. Song. Robust sound event recognition using convolutional neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 559–563, 2015.

[8] Saeid Sanei and J.A. Chambers. *Fundamentals of EEG Signal Processing*, chapter 2, pages 35–125. John Wiley  Sons, Ltd, 2013.

[9] P. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

[10] Ernst Niedermeyer and Lopes da Silva Fernando Henrique. *Electroencephalography basic principles, clinical applications, and related fields*. Lippincott Williams  Wilkins, 2005.

# A Appendix

# B Appendix: PSD Calculation

The method for computing this referenced power spectral density (PSD) was by computing an averaged periodogram over overlapping windows of the input time-domain data. [8] Welch's method allows for an estimation of the spectral content of an analyzed signal, and has the advantage of reducing the effects of noise on the spectrum of the signal. The analysis is performed by applying a window function (e.g. Hanning) to segments of the time-domain signal, computing the periodogram of these windowed signals (i.e. taking a Fourier Transform of the signal's autocorrelation), and then averaging the resulting periodograms into a generalization of the signal's power spectral density [9]. The reason this methodology is more robust to noise is due to the overlapping nature of the windowed segments taken, which after averaging compounds the effect of the window function's spectral emphasis on lower frequency content with its emphasis on the middle of the windowed signal.

As an alternative to Welch's method, we also computed PSDs using the multitaper method, which employs orthogonal filters upon the signal of interest to cancel the effects of noise. After deriving

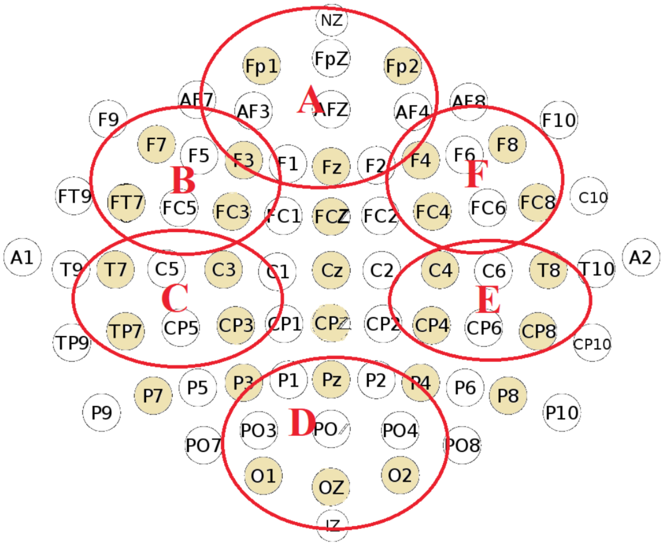

Figure A.1: Electrode map for 32-channel EEG. Electrodes used are colored yellow. Red circles labeled A-F correspond to clusters A-F used for localization training.

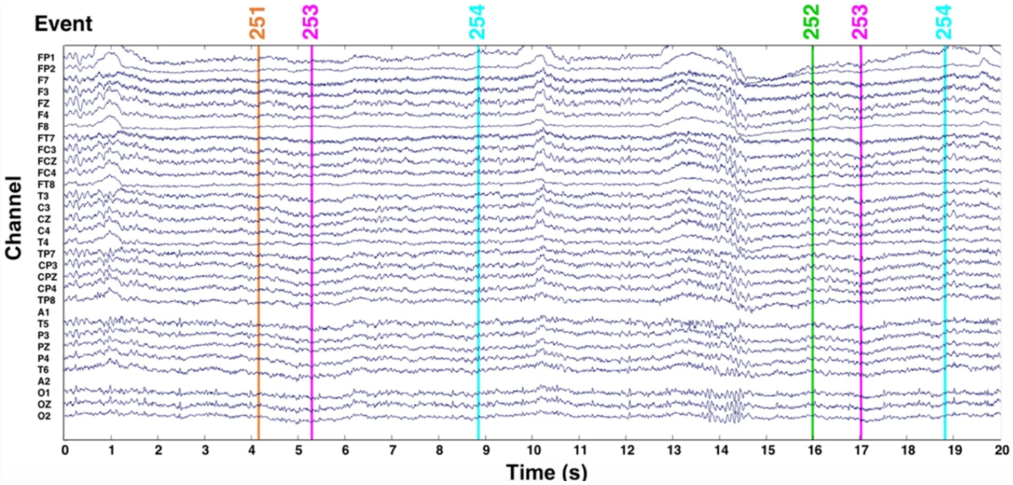

Figure A.2: 32 EEG signal waveforms representing 32 channels collected during a single experiment from a participant. Colored lines represent onsets of events (stimulus, reaction onset, reaction offset).

power spectral densities with either method, these are then collapsed into the cognition bands of interest. The 5 bands were: delta (0.5 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 12 Hz), beta (12 - 30 Hz), and gamma (30 - 100 Hz) [10].

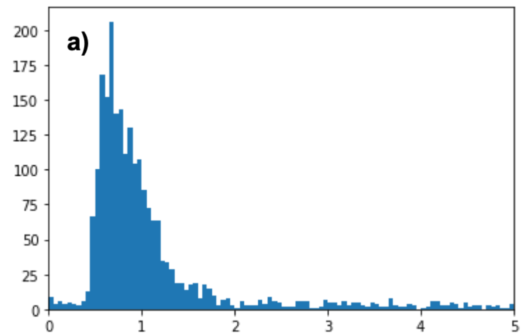

Figure A.3: Histogram of reaction times (time between deviation onset and response onset) for a) subject 44 and b) subject 41 (frequency vs. time in seconds).

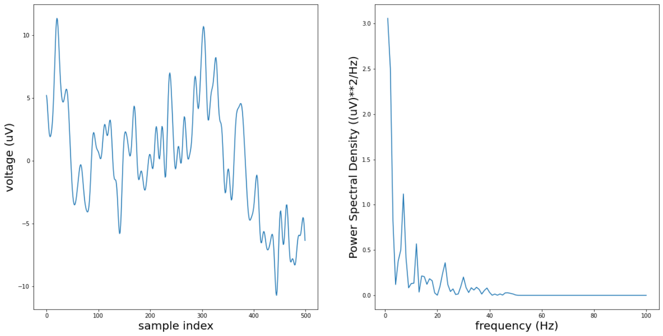

Figure A.4: Side by side comparison of the EEG signals in different domains. The left figure denotes the recorded time-domain data, while the right figure showcases the spectral content of this underlying one second of data. As expected, most of the signal power lies from 0 to 30 Hz.

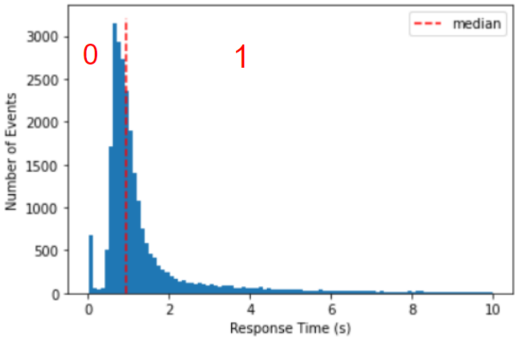

Figure A.5: Binary classification of estimated response time data with respect to the median response time.