MASTER THESIS

Utrecht University

Predictive Markers in EEG data for Susceptibility: A Data Driven Approach

Jesse Wouter Grootjen

August 2019

MASTER THESIS

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in the subject of Artificial Intelligence

Predictive Markers in EEG data for Susceptibility: A Data Driven Approach

Jesse Wouter Grootjen

August 2019

First examiner: dr. Christian P. Janssen
Department of Experimental Psychology
Utrecht University

Second examiner: dr. Ben M. Harvey
Department of Experimental Psychology
Utrecht University

Utrecht University

Jesse Wouter Grootjen
Predictive Markers in EEG data for Susceptibility: A Data Driven Approach
August 2019
Reviewers: dr. Christian P. Janssen and dr. Ben M. Harvey

Utrecht University
Department of Experimental Psychology
Faculty of Social and Behaviour Science
Heidelberglaan 1
3584 CS Utrecht
The Netherlands

Abstract

In this thesis, I test whether data-driven approaches can predict cognitive load based on EEG data. Previous research has tested in an oddball paradigm how the magnitude of an Event-Related Potential relates to cognitive load. However, an intervention is not always desirable in everyday scenarios. I therefore test, whether it is possible to predict the different cognitive load conditions before the intervention (i.e., oddball stimulus) is presented in three data-driven experiments. In experiment 1, I used machine learning to train a model that classifies the data for different conditions. In experiment 2, I test which characteristic/features of the data have the highest predictive power for each condition. In experiment 3, I tested how a Fourier transformation and data-driven approach can be used to complement each other. The combined results show that using a data-driven approach; I can predict cognitive load for the experiment at hand. However, the machine learning approach (experiment 1 and 2) require a long processing time. The combined approach (experiment 3) provides a consistent pattern that can differentiate between the cognitive load conditions. Further research is needed to test the generality of these findings for different datasets.

Acknowledgements

Working on this project was inspiring, interesting, exciting and fun, but a real challenge at times. As much as I enjoyed working on the topic of predicting Cognitive Load, it was not always easy to see the next step. Several people helped me to kind the best thing to do next while keeping my head up and enabled me to finish this thesis.

I would like to first and foremost thank my supervisor *Christian P. Janssen*. Thank you for the inspiring discussions, pushing me just a bit more, the critial and extensive feedback on my writing towards the end and all meetings from start till the end.

I would also like to thank my dad and his girlfriend for keeping up with me for such a long time, knowing when to bother me and when to leave me working.

To conclude, I would like to thank my friends and family for always helping me relax and have fun outside of work.

“*Man will only become better when you make him see what he is like.*”

- Anton Chekhov

Contents

Introduction I

Every day, we use our cognitive capacity to acquire knowledge and understand through thought, experience and the senses. An example of this is making a decision about what to wear. We make this decision based on information we receive (e.g. weather and occasion) that our brain processes. It allows us to understand and to relate to the world more effectively. Even though our cognitive capability exceeds that of every other animal, there are limits to our cognition.

For example, imagine you are reading a study book. If someone calls your name while you are reading, you might notice it. However, if you are very deeply involved in the material and thinking about it hard, you might not notice it, as you might experience a so-called "high cognitive load".

A high cognitive load suggests that the task at hand is difficult and, therefore, little cognitive capacity remains for processing other signals. For example, in the experiment of Van der Heiden et al. (2018), they show that when participants are tasked with driving in a simulation, that these participants are less susceptible to auditory signals than a participant who is stationary in that simulation. Thus, measuring cognitive load provides insight into the difficulty of the task at hand, enabling us to identify tasks where we have less cognitive capacity remaining and th erefore being less susceptible to other signals even though being susceptible to other signals is in some cases preferred (e.g. safety-critical situations).

Traditional research conducted into cognitive load concludes results through interventions and/or manipulations in an experiment (e.g. Klimesch et al. (1993) and Scheer et al. (2016)). For example, in a study by Van der Heiden et al. (2018), frequent sounds are played to test how susceptible people are to sounds under different task load conditions. While using interventions is a proper way to test cognitive load, it is less suitable for everyday scenarios like driving a car in real traffic. In such cases, providing additional stimuli (as intervention) can distract from the task that one needs to focus on. Instead, it would be beneficial if the cognitive load could be assessed passively *without an intervention*. I will, therefore, explore whether data-driven techniques can be used to detect patterns in data that predict cognitive load. An additional advantage of such an approach is that it can also find new characteristics that attribute to different conditions (Shi et al., 2018), and when caution is taken to prevent overfitting (Pitt and Myung, 2002), provide novel insights.

1.1 Problem Definition

In this thesis, I test whether a data-driven (bottom-up) approach can predict cognitive load. Specifically, I will use data from an experiment by Van der Heiden et al. (2018), in which participants were exposed to three conditions that offer a variety of task load and associated cognitive load. Their study showed that when an intervention was provided (i.e., when cognitive load was probed through the presentation of an oddball stimulus), that the different conditions could be detected in an event-related potential signal. My question is new whether these conditions can be predicted based on the EEG signal *before* the intervention (oddball stimulus) is presented. Specifically, my research question is:

Can a data-driven approach be used to, without interventions and manipulations, predict human cognitive load?

Although there is no guarantee that patterns will be found, the approach is strengthened by successes in other labs where machine learning was able to detect specific cognitive processes within EEG data (e.g. Obermaier et al. (2001) and Doroshenkov et al. (2007))

The remaining part of the thesis proceeds as follows. Chapter one provides further background on cognitive load, how cognitive load can be measured, the application of machine learning techniques to measured data from experiments, and a description of the experiment of which I used the data in this thesis. Chapter 3 discusses the General Method, where I provide insight into the generic processing of the data for the different experiments. This is followed by three experimental chapters. Chapter 4 looks at my first experiment, where I use machine learning to classify different amounts of cognitive load. Chapter 5 refines this work by testing what characteristics are most predictive for the categorization into different cognitive load conditions. Chapter 6 tests whether a top-down approach using proven preprocessing methods can provide insight into what differentiate the cognitive load conditions. Lastly, Chapter 7 discusses the combined results with their implications and limitations.

Literature Review II

This section provides a literature review of the fields in which this thesis is placed. At first, I start with a description of cognition and cognitive load. Next, I discuss the experiment from which the data is used in this thesis and provide arguments as to why this data can be used. Finally, we look at the application of machine learning techniques to EEG data.

2.1 Cognition and Cognitive Load

Most of the time, when we talk about cognition functions, we are referring to the cognitive skills in order to receive, select, transform, develop, and recovery information that we've received from external stimuli. Although we can study these as separate ideas, we need to keep in mind that these cognitive skills are always interrelated and that these sometimes overlap. The main cognitive functions are defined as memory, executive functions, and lastly, attention (Marvel and Paradiso, 2004).

Cognitive load refers, in general, to the load a specific task places on one or more of the cognitive functions named above. Thus, as there are different cognitive functions, we can say that there are different *domains* of cognitive load in accordance with the functions. Having a high cognitive load in the attention domain placed upon a person with a specific task might have that person miss other stimuli. A prime example of this is the video where one is tasked with counting the number of passes the team in white cloths makes, while this is happening most people miss the gorilla walking through the background (Simons and Chabris, 1999). Because we are tasked with counting the passes, our attention is focussed on this. We place a cognitive load on ourselves; by doing this, we are less susceptible to other stimuli that require attention. To some extent, this is because of the highly efficient filtering our brain does, filtering through all the stimuli leaves only the relevant/'important' stimuli to be perceived. For the remainder of this thesis, I define cognitive load as the load placed upon the cognitive function of attention.

In an attempt to measure the cognitive load of a human, various methods have been employed. Among the *non-invasive* methods are, for example, functional Magnetic Resonance Imaging *fMRI* (e.g. Herzmann et al. 2017; Young et al. 2018), Magnetoencephalography *MEG* (e.g. Colclough et al. 2015; Basti et al. 2018) and Electroencephalography *EEG* (e.g. Artoni et al. 2018). While all of the previously named methods are candidates for measuring workload, fMRI and MEG require expensive setups that prohibit them from being implemented in consumer-grade vehicles.

EEG, on the other hand, is a cheaper option and provides insight into cognitive load (e.g. Antonenko et al. (2010), Anderson et al. (2011), and Leppink et al. (2013)) and susceptibility to auditory signals which can be used as metric for cognitive load (Wester et al., 2008; Van der Heiden et al., 2018). These last two studies show that susceptibility to auditory signals varies over different tasks and that humans are less susceptible to auditory signals while having a high task load than when having a low task load (i.e. when there is a high cognitive load human are less susceptible to auditory stimuli than when humans have a low cognitive load).

In more detail, in Van der Heiden et al. (2018) they showed that while in a driving simulation; when a participant was tasked with manual driving (i.e. a high task load) they where less susceptible to auditory stimuli, when a participant was tasked with driving in an autonomous car (i.e. a medium task load) they where more susceptible to auditory signals, and when a participant was placed in a stationary car (i.e. a low task load) they where even more susceptible to auditory signals. They also showed that participants who were tasked to react (i.e. press a button) to an auditory stimulus were more susceptible to this stimuli than participants who were asked to ignore this auditory stimulus. Because the data of Van der Heiden et al. (2018) is available through the university, the data has different levels of cognitive load, and because the researcher who did this study is easily approachable, I chose this dataset to use for my thesis.

However, research into cognitive load is almost exclusively done in an experimental setup where they use interventions and/or manipulations to measure cognitive load. While interventions and manipulations are fine for experimental setups, this is not acceptable in everyday scenarios. I would like to test whether this is possible to determine the cognitive load of a person without using interventions and/or manipulations by using a bottom-up (machine learning) approach.

2.2 Background on Machine Learning Technique(s)

Machine learning has recently gained popularity, spurred by well-publicized advances like deep learning (Lecun et al., 2015) and the advancement in hardware and acquired data that allow for big data analytics (Chen et al., 2014). According to Lee et al. (2018), there are already a number of significant commercial applications that have appeared, including but not limited to recommendation engines, speech and handwriting recognition systems, content identification, image classification/retrieval, automatic captioning, spam filters, and demand forecasting. One prime example is IBM Watson (Ferrucci, 2012); this question-answering computer system, communicating in natural language, entered the public stage through its winning performance in the quiz show Jeopardy, and is now used in commercial applications like lung cancer treatment (AOCNP, 2015) and heart failure identification (Guidi et al., 2016).

Machine learning is the field of science that 'gives computers the ability to learn without being explicitly programmed' (Samuel, 1959). It, therefore, differs from human learning (true learning), where variables of learning have meaning and are not just features. Following Shalev-Shwartz and Ben-David (2014) and Trevor et al. (2017), almost all of machine learning can be split into three different types of learning; namely supervised learning, unsupervised learning and reinforcement learning. In supervised learning, learning occurs from a labelled dataset. This creates the opportunity to train a classifier to label new data in the same manner as in the learning dataset (Trevor et al., 2017, Chapter 2). In unsupervised learning, no labelled data is needed, in this case, learning occurs by means of finding a group that has similar features (clustering) (Trevor et al., 2017, Chapter 14). For reinforcement learning a way of feedback is introduced to give positive and negative feedback according to how good something is, this results in an online learning tool (Sutton and Barto, 1998). Because the data from the experiment that I use for my thesis is labelled, we will in the remainder of this thesis only look at supervised learning.

2.2.1 Random Forest

Random Forests is one of these supervised learning methods and was introduced by Breiman (2001). A Random Forest (see algorithm 1) is a classifier consists of a collection of decision trees, each tree is built from a sample ($\mathbf{Z}^*$) drawn with replacements (i.e., a bootstrap sample) from the training set S. A tree T_b is build following the following three steps recursively until the minimum node size n_{min} is reached: 1) Select m variables at random from the p variables. 2) Pick the best variable/split-point among the m. 3) Split the node into two daughter nodes. After doing this for all trees, we have a forest $\{T_b\}_1^B$. In contrast to its original publication Breiman (2001), I use the implementation of Pedregosa et al. (2011), where their implementation combines classifiers by averaging the probabilistic prediction of each tree, instead of letting each classifier vote for a single class. Thus, to now predict a new point x; let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = confidence \;\; vote\{\hat{C}_b(x)\}_1^B$. As the number of trees in a Random Forest increases, the test set error rates converges to a limit, meaning that there is no overfitting in large Random Forests (Breiman 2001 and Shalev-Shwartz and Ben-David 2014, Chapter 18).

Algorithm 1: Creating a Random Forest

Input: training set S

while $b < B$ **do**
    Draw a bootstrapped sample $\mathbf{Z}^*$ of size N from the training data ;
    while *nodesize* $> n_{min}$ **do**
        Select m variables at random from the p variables;
        Pick the best variable/split-point among the m;
        Split the node into two daughter nodes;
    end
end
return *Forest* $\{T_b\}_1^B$

A reason as to why Random Forests are popular is because of its interpretability. Random Forests can be used to determine feature importance (e.g. Menze et al. (2009)), and its decision making can be viewed. This might not be feasible to do by hand when working with a large forest, but every decision tree can be viewed, its root, node and leaves. Because of its popularity successes with random forest are in abundance, in Yoshida et al. (2014) they successfully classify a driver's cognitive state in a real driving situation with eye-movements. Bashivan et al. (2015) showed that using random forests it is also possible to rank the 192 features and select the 64 features that contribute to cognitive load in a memory task.

Despite its success, there are two major contributors to wether it succeeds or whether it fails. Correlation between the different trees is key. Uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this effect is that the trees protect each other from their individual errors. While some trees may be wrong, many other trees will be right, resulting in that as a group the trees are able to classify correctly. So the prerequisites for a random forest to perform well are: 1) There needs to be some actual signal in the features so that models built using those features do better than random guessing. 2) The predictions (and therefore, the errors) made by the individual trees need to have low correlation with each other (Trevor et al., 2017, Chapter 15). To make sure that there is enough diversity between the trees we need to take the correlation between features into consideration, we can select features of importance which do not correlate using feature selection.

2.2.2 Recursive Feature Selection

Recursive feature selection is such a feature selection method, it trains a classifier for all features - 1 and repeats this process such that each feature is left out exactly once.

This provides with an insight into the feature that is approximated to be least significant. This feature is then removed from the features pool, and the process repeats until only a set number of features remains. This algorithm can be adapted to use different models to evaluate the features; in algorithm 2, I show an implementation from Guyon et al. (2002) adapted with a Random Forest.

Algorithm 2: Recursive Feature Selection

Input: an instance a vector of feature values (p^*) and class labels

Output: ranking of features

Find the optimal values for the tuning parameters of the RF model;
Train the RF model with full feature set;
$p \leftarrow p^*$;
while $p \geq 2$ **do**
    $RF_p \leftarrow RF$ with the optimal tuning parameter for the p variables and overservations in **Input**;
    $w_p \leftarrow$ calculate weight vector of the $RF_p(w_{p1}, \ldots, w_{pp})$;
    $rank.criteria \leftarrow (w_{p1}^2, \ldots, w_{pp}^2)$;
    Remove variable with lowest value in *rank.criteria* vector from **Input**;
    $Rank_p \leftarrow$ variable with lowest value in *rank.criteria* vector;
    $p \leftarrow p - 1$;
end
$Rank_1 \leftarrow$ variable in **Input** $\notin (Rank_2, \ldots, Rank_{p^*})$;
return $(Rank_1, \ldots, Rank_{p^*})$

This procedure is an instance of backward feature elimination. For computational efficiency, more than one feature can be removed at a time. This, however, is at the expense of possible classification performance degradation. If features are removed one at a time, features can be ranked correspondingly. Features that remain among the top-ranked are not necessarily ones that are individually most relevant, only together these features are optimal. Although this algorithm is relatively old, it still gets used in new research (e.g. Wang et al., 2018; Liu et al., 2018).

General Method III

3.1 Overview

For the experiments reported here, Van der Heiden et al. (2018) provided the data from their experiment. This dataset was used because it was easily available, labelled and contained conditions that vary in cognitive load. The data contains EEG recordings of 18 participants (11 F) with an age range of 20 to 25 years old (M = 22.06 years, SD = 1.39 years). In their experiment, they tested how susceptible human drivers are to auditory signals in three situations: when stationary, when driving, or when being driven by an autonomous vehicle. They measured susceptibility using a three-stimulus auditory oddball paradigm (*intervention*) while participants experienced these different situations and studied this specifically through the frontal P3 (fP3) Electroencephalography Event-Related Potential response (EEG ERP response). In their results, they showed that the fP3 ERP response is reduced in autonomous conditions compared to stationary conditions, but not as strong as when participants drove themselves. In addition, the fP3 component is further reduced when the oddball task does not require a response (i.e., in a passive category, versus active). I will test if it is possible to differentiate between these different conditions and categories without intervention and using a bottom-up approach.

3.2 Preprocessing

For each of the experiments, these data were processed offline using the MNE v0.17.0 package (Gramfort et al., 2014) in Python (Rossum, 1995). To compensate for noise from the mains, a 50 Hz notch filter was applied and the data was referenced to the mean of left and right mastoid signal. The data was also corrected for a 50ms delay when logging the stimulus onset, see Van der Heiden et al. (2018) for details. Trials containing false alarms (e.g. button-presses to the non-target auditory signal), misses (e.g. failed responses to the deviant target tone), and invalid responses (e.g. responses outside the 100-950 ms interval, after correction for the delay) following Van der Heiden et al. (2018).

After following the preprocessing steps from Van der Heiden et al. (2018), I also rejected trials that are contaminated with blinks using the EOG electrodes. Lastly, because I want to look at the interval of 500 ms preceding the stimulus, a high-pass filter is applied to filter out frequencies < 2 Hz because for lower frequencies I cannot capture a complete sine wave in this time interval. The interval of 500 ms preceding the stimulus was chosen because it limits the amount of data that is used/needed to check for predictive markers in the EEG data while still allowing for a large spectrum of wavelengths.

Experiment 1 IV

In this experiment, I wanted to evaluate whether it is possible to train a machine learning model that predicts the driving condition (stationary, autonomous, or manual driving) based on EEG. To do this, I compared the validation score of this model to a random model and a most common model (this model always chooses the label that is most common in the data). If the machine learning model is better than the other two models, it might be usable in online use-cases where the driving condition is associated with a difference in susceptibility. If a model can be used to predict this driving condition, it could be used to determine when a person is most susceptible to auditory stimuli.

4.1 Method

Using different preprocessing techniques, I wanted to create more features than the 64 features from the electrodes. Because I was using a bottom-up approach, I did not know what features mattered, and by using Random Forests, I could be able to determine what feature(s) matter to classify the different driving conditions. After preprocessing, according to the general method (Chapter 3), I copied the data into a total of four datasets and processed them further. For dataset 1, a baseline correction was applied for an interval of 100 ms preceding the stimulus onset, and the data was selected to be in the interval of 500 - 0 ms preceding the stimulus. For dataset 2, the same baseline correction was applied, but instead of the interval, an average was placed upon each data point containing the average of 500 ms preceding that datapoint. Dataset 3 had no baseline correction applied and contained the data in the interval of 500 - 0 ms preceding the stimulus. The last dataset, dataset 4, contained no baseline correction, and each data point contained the average of 500 ms preceding that point. Finally, I downsampled all datasets from 2048 Hz to 200 Hz.

The data was separated to account for features, features that the dataset consist of, represent the *EEG bands* which have been estimated from each electrode. Frequencies were extracted for the following bands: δ (<4 Hz), θ (4-7 Hz), α (8-12 Hz), β (12-30 Hz), and γ (31-50 Hz). Additionally, the following four features were also extracted: $\delta + \theta$, $\theta + \alpha$, $\alpha + \beta$ and $\beta + \gamma$ (such that $\delta + \theta$ contained <7 Hz, etc.), these four features counter overlap in the different bands. Nine features in total were extracted from each of the 64 EEG electrodes, resulting in a total of 576 features for each dataset. The features were denoted as "proprocessing_electrode_feature". For example, feature "nByA_FCz_al", represented the alpha band (8 to 12 Hz) for electrodes "FCz" without being baselined and but with an average.

4.1.1 Random Forest

As a machine learning technique I chose to use a Random Forest model from Pedregosa et al. (2011) because previous research (Chapter 2) shows promising results by using Random Forests on EEG data, they can hardly be overfitted and are fast to train. A total of 2304 features were then used to train this model, setting the random state (seed) to 1636 to make sure the experiment could be repeated, maximum depth of 25% of all features to make sure that a tree wouldn't turn out to have a single node at each depth. I put the total amount of trees to 64 trees to limit calculation time. All the remaining settings remained on default because a preliminary test showed that these settings had minimal impact on the resulting score.

Stratified 10-fold cross-validation was used to compute the final result to make sure that each fold contains approximately the same percentage of samples of each target condition as the complete set and as validation for the stability of the machine learning model trained. Using Stratified 10-fold cross-validation the model was trained with 8 out of the total of 10 folds, where the remaining 2 folds were used to test the model. This was repeated such, that every fold was used exactly 8 times to train and 2 times to test to reduce the chance that the model was biased.

The labels used for the training and testing of the model were the different driving conditions. I decided to leave the category (active vs passive) out because the participants were asked to react to auditory stimuli after the stimuli were presented. However, the data used in this experiment were selected to exclude data after the auditory signal was given.

To evaluate the score of Random Forest model to something, I also created two additional models. The most common model picks the label that is most common in the dataset. In the case of this experiment, the most common label is *Driving*. Lastly, I created a random model, which was also set to a random state (seed) of 1636 to make sure the result could be repeated.

4.2 Interim Results

Fig. 4.1 shows the accuracy of the proposed method compared to a model that is completely random and a model that picks the most common label. In this figure, I show that I was able to get an accuracy of 41,1% using a Random Forest model, while the model that always picked the label that is most common in the data was able to score 35,4%. Using a completely random model decreases the accuracy by ∼2% and results in an accuracy of 33,4%, which is what one would expect from a completely random model.

FIGURE 4.1: The Accuracy of the model compared to random and most common

4.3 Discussion of Results

In this experiment, I used the preprocessed EEG data to assess the ability of a Random Forest model to classify the three driving conditions. Results showed that this classification was done to a better extent than randomly applying a label to data or classifying everything as the label that is most common. However, because of its overall accuracy, this model is not a solution for classifying cognitive load in safety-critical situations.

Random Forest models are vulnerable to features that correlate and feature that do not contribute to the labelling of the data. This vulnerability is because in each node in a tree from a random forest, a random pick selects a set of features which could be used in that node. The tree then selects the feature that leads to the most substantial Gain (see Algorithm 1, step 5). Reducing the Gini Impurity is one of the ways you can calculate the highest Gain, it means that the decision tree tries to form nodes containing a high proportion of samples (data points) from a single label by finding the features that cleanly divide the data into classes. When the random pick of features in the initial stage doesn't contain features that can be used to label the data clearly, the decision tree loses its ability to perform well.

Even though I was able to create a Random Forest model to classify the driving conditions, this model is not suitable for safety-critical situations. Neither was I able to determine what features were ranked high among the trees. A next experiment should be to reduce the number of features such that only features that contribute to labelling the data are considered in a next Random Forest model, neither should redundant features be considered to cut down on time used to create and evaluate the Random Forest model.

Experiment 2 V

In this experiment, I followed up on one of the discussion points from experiment 1. Namely, that a reduction in features that correlate with each other and the elimination of features that do not contribute to the label should improve the score for a Random Forest model. Using recursive feature selection, I want to reduce the features that are used in the Random Forest model. Using recursive feature selection provides insight into the correlation between features and the contribution of a feature to the label. It will allow for selecting the features that contribute most to the label and eliminating features where the correlation with the chosen features is high. This should result in giving the Random Forest model a better chance of picking features that can be used to distinguish between the labels.

5.1 Method

For this experiment, I used the same preprocessing method as in experiment 1 (Chapter 4). However, instead of using stratified 10-fold cross-validation, I used a 5-fold to reduce calculation time. I also decided oly to use the manual driving condition and stationary label instead of all 3 labels to reduce training and testing time, by only using 2 of the 3 labels the calculation time could improve with 33%. I chose these two labels because these were the two most extreme opposites in previous experiments (Van der Heiden et al., 2018). Saving time is important for this experiment because to eliminate one feature at a time for a complete ranking, I needed to do $4.22 * 10^{6748}$ iterations of the algorithm.

Next, I searched for settings for the Random Forest algorithm to minimalize the time it takes to do the 5-fold cross-validation while making sure that the test scores are still better than random and the most common label. This was done by changing the parameters is the Random Forest model one at a time, while leaving the other parameters on the default setting. The number of trees was tested for 1, 2, 4, 8, 16, 32 and 64 trees (default = 8), the maximum depth of the trees in the random forest was tested for $0.1, 0.2, \ldots, 1$ of all the features (no maximum depth as default). I also tested with the minimal sample split which is required to split a node and changed it to $0.1, 0.2, \ldots, 1$ of all the datapoint (default = 2). Lastly, I changed the number of features to consider when looking for the best split (default

= $\sqrt{total\ number\ of\ features}$) and changed it to $0.1, 0.2, \ldots, 1*$ total number of features. Afterwards, I want to use recursive feature selection to eliminate features one at a time.

5.2 Interim Results

Preliminary research showed (figure 5.1 and 5.2) that there are a vast amount of features that correlate heavily with each other. Thus, feature selection should pick the feature that contributes the most to the classification of the label and leaves the remaining features out. In figure 5.1 on the left, I show the correlation between the features that haven been preprocessed such that they have not been baselined and have been averaged with the delta-frequency band. The brighter the colour, the more correlated the individual features are. Features on the diagonal line are correlated with themselves; thus, they are the brightest. In figure 5.1 on the right, I show the correlation between the features that haven been preprocessed such that they have not been baselined and have been averaged. In this figure, we can clearly see that there are some bright spots among the features; most of them are around the diagonal line.

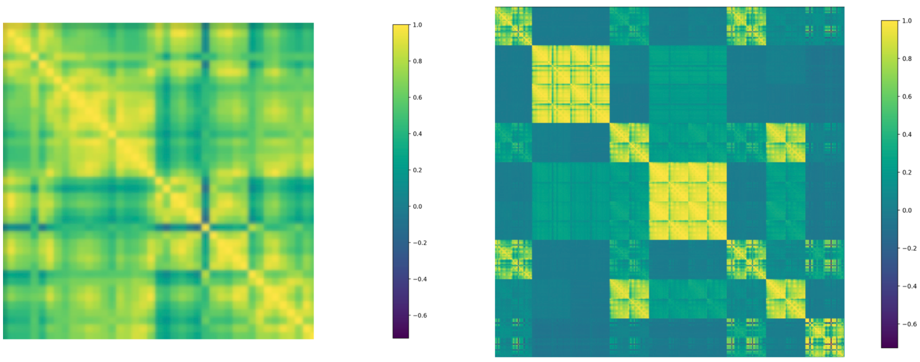

FIGURE 5.1: **Left:** Correlation matrix of the delta frequncy band for the no baseline but with average features. **Right:** Correlation matrix of the no baseline but with average features

In figure 5.2, I show a complete overview of the correlations between features. Again, there are some large bright spots in the figure. All these bright spots show features that are highly correlated with each other.

The EEG signal extracted from the recording was processed, filtered and segmented into 500 ms epochs as described in the General Method (Chapter 3) and the specific method for this experiment. The complete data set was divided into 5 sets in accordance with the stratified cross-validation to avoid bias and to make sure that each fold contains approximately the same percentage of samples of each target condition as the complete set. Next, the model was trained with 4 out of the total of 5 sets,

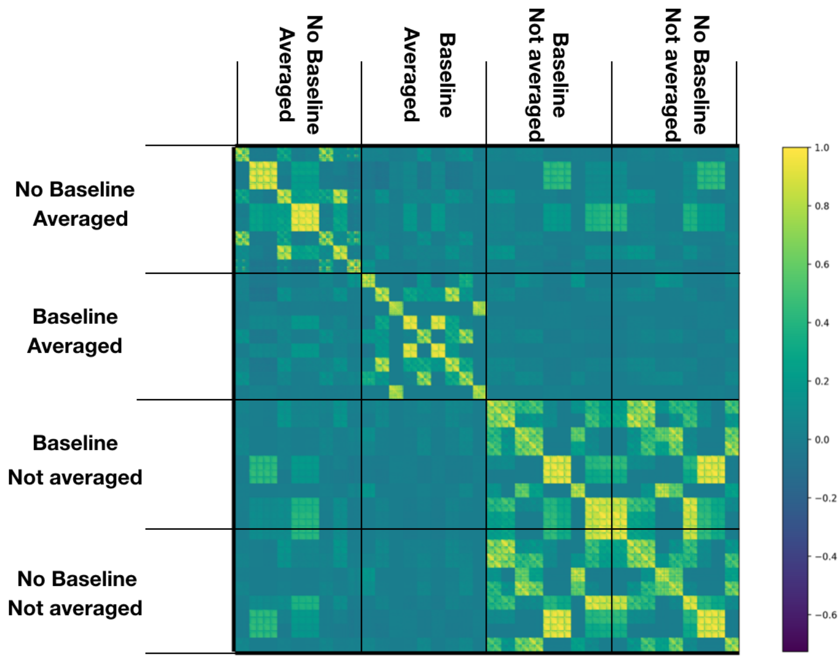

FIGURE 5.2: Correlation matrix that shows the correlation between features. (best shown in color)

where the remaining 1 set was used to test the model. This was repeated such, that every set was used exactly 4 times to train and 1 time to test. Figure 5.3 shows the train en test scores of Random Forest models with a different number of trees and time it takes to train and test these models.

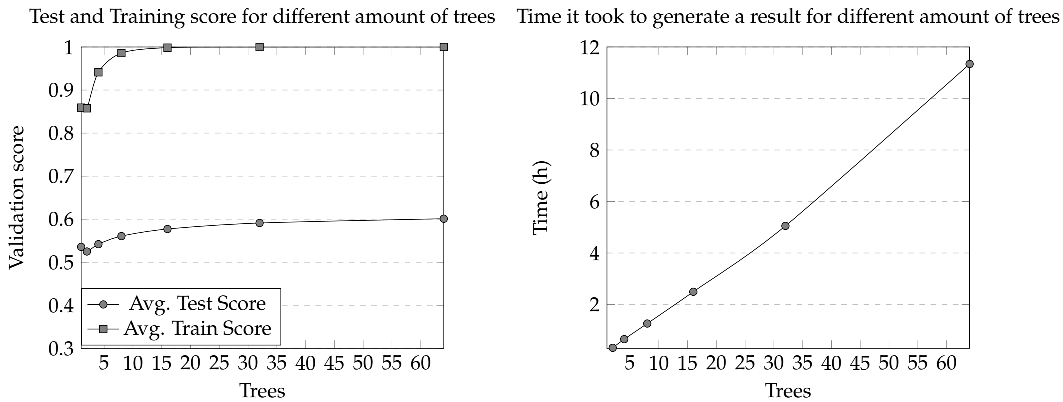

FIGURE 5.3: **Left:** this graph shows the train and test score for different amount of trees in the Random Forest with all other settings on default from a 5-fold cross validation. **Right:** this graph shows the time it took for different amount of trees to generate a result from a 5-fold cross validation.

This result highlights a problem for a bottom-up (data-driven) approach. The score required to be an improvement on the most common label is 0.54 (from all the labels, 54% of the labels is "stationary"). Thus only a model with 4 or more trees in the forest satisfied the requirement of having a better classification than most common. To train and test this model it takes 2359 seconds (roughly 39 minutes). Using recursive feature selection would mean that this model should be trained and tested for a total of 2304 times (each time the model would leave one feature out to train). The time it would take to calculate the worst feature would take roughy $5.4 * 10^6$ seconds (63 days). As shown in Algorithm 2 in Chapter 2, the process would repeat until there were only two features left (or another chosen number of features).

While the remaining parameters of the random forest algorithm also contributed to the score and time it takes to train and test the model, the performance and calculation time changes I found were marginal.

5.3 Discussion of Results

In this experiment, I tried to use the preprocessed EEG data to assess which features contributed most to the labelling of the data. Results show, however, that using the recursive feature elimination with a Random Forest is not a viable option because of time constraints. The complete algorithm would take approximately $1.89 * 10^{6746}$ years when there is no improvement in calculation speed when there are fewer features to consider. The algorithm can be adapted to drop more than one feature each time (e.g. drop the two worst features) but this can also result in inaccuracy.

A next experiment should find another way of preprocessing the initial data to produce fewer features and look for manners of preprocessing EEG data that has already been used to produce significant results.

Experiment 3 VI

In this experiment, I wanted to evaluated the power of the different EEG frequency bands and calculated if there is a significant difference between the different categories (active and passive), the different conditions (stationary, autonomous, and manual driving), and if there is an interaction effect between the category and conditions before the auditory stimulus. By calculating the power of a frequency band, I hoped to acquire fewer features than in my previous experiments, but still, be able to distinguish between conditions before the auditory stimulus.

6.1 Method

After preprocessing, according to the general method (Chapter 3), I applied a 50 Hz low-pass filter to the data to remove all the frequencies that were out of scope for the EEG frequency bands. Next, I selected the interval of 500 - 0 ms preceding the stimulus. The dataset was then downsampled from 2048 Hz to 500 Hz to reduce the density of the data (data reduction) but have enough data remaining to processes the data further. Afterwards, I reduced the electrodes such that they still cover the whole scalp and conform to the 10/20 standardized layout. I also decided to add the FCz as an additional electrode because this electrode yielded a significant result in the experiment of Van der Heiden et al. (2018). The remaining electrodes were as follows: Fp1, Fp2, F3, F4, Fz, Cz, C3, C4, T7, T8, Pz, P3, P4, O1, O2, Oz.

6.1.1 Fourier transform

For each trial and electrode, I applied a fast Fourier transform to calculate the power for the frequency spectrum. Next, I took the calculated averages from the output from the fast Fourier transform for each participant in the three different conditions. Thus, resulting in three averaged fast Fourier transforms for each participant, in total 54 for each electrode (18 participants $\times$ 3 conditions). Next, I grouped and averaged the frequency ranges to account for the frequency bands found in EEG data, namely; δ (<4 Hz), θ (4-7 Hz), α (8-12 Hz), β (12-30 Hz), and γ (31-50 Hz).

6.1.2 ANOVA

Statistical analyses were conducted in R. I used a 2 (categories: active and passive) x 3 (conditions: stationary, autonomous, and manual driving) ANOVA, with the active and passive category as a between-subjects factor and driving condition as a within-subjects factor. Through-out all analyses, a significance level of $\alpha = .05$ for the conditions was used. In cases where a main effect of driving was found, I used Holm-corrected pairwise t-tests to determine the differences between the three driving conditions.

6.2 Interim Results

Table 6.1 reports the F-values of the main effect of driving condition. As can be seen, the theta frequency band produced significant differences. More specifically, for 11 out of the 17 electrodes, there was a significant effect. For electrodes that are located on the somatosensory cortex, parietal and occipital lobe, there was consistently a significant difference in theta power between the three driving conditions. The effect was the strongest pronounced on the electrode P4, where a Holm-corrected pairwise comparisons yielded significant differences between stationary and driving ($p = .002, d = 0.72$), between autonomous and driving ($p = .006, d = 0.54$) and between stationary and autonomous ($p = .016, d = 0.30$). More generally, on all of the electrodes, a similar pattern was found; namely, the stationary condition consistently had the highest theta power and the manual driving conditions the lowest.

Electrode	δ-power	θ-power	α-power	β-power	γ-power
Fp1	0,121	0,918	0,241	0,715	0,849
Fp2	1,563	1,223	0,219	0,28	0,157
F3	0,825	1,879	0,250	0,601	0,887
F4	1,983	2,433	0,622	0,325	0,158
Fz	1,381	3,690*	0,003	0,192	0,408
FCz	1,642	5,149*	0,076	0,161	0,435
Cz	1,959	7,848**	0,327	0,13	0,492
C3	1,602	7,785**	1,208	0,069	0,604
C4	2,225	9,286***	0,209	0,385	0,374
T7	1,463	0,006	2,924(.)	1,503	0,041
T8	2,415	2,969(.)	0,210	0,300	0,591
Pz	2,104	4,533***	1,734	0,203	0,673
P3	2,108	10,370***	1,802	0,176	0,505
P4	2,259	10,816***	0,946	0,144	0,642
O1	2,255	5,494*	5,489**	1,109	0,307
O2	2,358	6,126*	0,245	0,588	2,072
Oz	2,490(.)	6,924**	2,506(.)	0,530	0,459

TABLE 6.1: This table shows the F-values from the main effect of driving condition of each electrode for the different frequency bands when F(2, 16). Signif. codes: '***'$<$ .001, '**'$<$ .01, '*'$<$ .05, '(.)'$<$ .1.

For the main effect of response category (active versus passive), there were only consistent significant effects on the gamma frequency band. Specifically, for 11 out of the total of 17 electrodes, there was a significant effect. Electrodes that yielded a significant effect tended to be located around the midline region of the brain. Again, the effect was strongest pronounced on electrode P4, with an F-value of $F(1,16) = 7.958, p < 0.05, d = 1.03$. On all electrodes with a significant effect, a similar pattern was found, where the passive category consistently yielded a higher gamma power than the active category. Table 6.2 reports the F-values of the main effect of category (active versus passive).

Electrode	δ-power	θ-power	α-power	β-power	γ-power
Fp1	2,178	3,724(.)	3,040	2,994	5,184*
Fp2	0,210	2,138	2,694	2,976	5,158*
F3	0,885	1,100	1,349	1,325	2,447
F4	0,616	2,169	1,670	2,121	1,533
Fz	0,674	0,816	1,775	2,285	5,404*
FCz	0,630	0,586	1,793	2,410	6,136*
Cz	0,602	0,422	1,608	2,633	6,459*
C3	0,612	0,201	1,587	2,447	3,467(.)
C4	0,526	0,887	1,533	1,016	1,393
T7	0,654	0,599	3,928(.)	5,271*	7,479*
T8	0,462	1,225	1,951	2,019	3,256(.)
Pz	0,517	0,000	1,580	3,547(.)	6,624*
P3	0,533	0,000	2,418	4,358(.)	7,325*
P4	0,537	0,212	1,975	3,36(.)	7,958*
O1	0,460	0,004	0,692	2,564	6,523*
O2	0,488	0,086	0,183	0,002	1,283
Oz	0,472	0,001	0,859	2,946	7,055*

TABLE 6.2: This table shows the F-values from the main effect of active / passive category of each electrode for the different frequency bands when F(1, 16). Signif. codes: '***'< .001, '**'< .01, '*'< .05, '(.)'< .1.

There was no significant interaction between the driving mode and response category for any electrode or frequency band, see table 6.3.

Electrode	δ-power	θ-power	α-power	β-power	γ-power
Fp1	0,680	0,136	0,532	0,807	0,819
Fp2	0,479	0,140	0,411	0,224	0,138
F3	0,164	0,173	0,547	0,841	0,803
F4	0,377	0,731	0,695	0,429	0,027
Fz	0,290	0,162	0,232	0,356	0,430
FCz	0,351	0,215	0,176	0,325	0,463
Cz	0,354	0,223	0,111	0,245	0,486
C3	0,292	0,031	0,106	0,189	0,640
C4	0,321	0,844	0,526	0,625	0,467
T7	0,319	0,233	0,008	0,074	0,708
T8	0,336	1,368	0,737	0,766	0,800
Pz	0,256	0,303	0,288	0,115	0,364
P3	0,289	0,292	0,283	0,103	0,314
P4	0,277	0,004	0,095	0,175	0,516
O1	0,284	0,641	0,929	0,535	0,174
O2	0,377	0,354	0,706	0,205	0,067
Oz	0,278	0,427	0,261	0,260	0,269

TABLE 6.3: This table shows the F-values from the interaction effect of the active / passive category with the driving condition (stationary, autonomous, and driving) of each electrode for the different frequency bands when F(2, 16).

6.3 Discussion of Results

In this experiment, I investigated if the output of a Fourier transform can be used as a means to distinguish between the different conditions and categories. I show in table 6.1 that a reduction in θ-power is significant between the driving conditions for 11 of the 17 electrodes and in table 6.2 that a reduction is γ-power for 11 of the 17 electrodes is significant between the categories. I also showed in table 6.3 that I did not find an interaction effect between the conditions and categories. After finding a pattern using a bottom-up (data-driven) approach, I need to validate my findings in theory.

Previous EEG studies have identified the frontal midline region as an optimal location for detecting load-sensitive EEG signals in the theta band (Ishii et al., 1999;

Gevins et al., 1997; Inouye et al., 1994). They have also shown that alpha-band signals over the frontal and parietal regions tend to be relatively more sensitive to the attention demands of tasks than to the alpha band signals recorded over other regions (Gevins et al., 1997; Klimesch et al., 1993). In Klimesch (1997) and Klimesch et al. (1997), they emphasize that alpha activity, especially in the higher frequency range (10–13 Hz, referred to as upper alpha), is associated with semantic information processing, in particular with searching, accessing, and retrieving information from long-term memory. Since most cognitive tasks draw on these processes, alpha Event-Related Desynchronization can be observed in a wide range of task demands (Klimesch et al., 2006). Theta activity, in contrast, has been frequently related to episodic and working memory as theta Event-Related Synchronization increases parametrically with working memory load and is sustained during the retention period (e.g. Jensen and Tesche (2002) and Kahana et al. (2001)).

The results in this experiment also reflect that the theta band is ideal for detection differences in cognitive load. However, the results reflect that the parental frontal region is more suitable for the detection of differences in cognitive load, then the aforementioned frontal midline region. The results also don't reflect any relation to the alpha band for detecting differences in cognitive load. Which I did not consider is the fact that some electrodes might be placed over the primary visual, auditory, or somatomotor cortices, this might influence the resulting indices to be disproportionately affected by task-related activation of these regions. While this is not a problem when detecting cognitive load in a task-related environment, the same model cannot be used to generalize because the task-related activation of regions might differ.

Muller et al. (2000) showed that gamma-band signals differ when subjects attend to a particular stimulus, compared to when the same stimulus was ignored. Participants in the study of Van der Heiden et al. (2018) were asked to either press a button when an auditory signal was presented (attend to a stimulus) as a participant in the active category or to ignore the auditory signal (ignore the stimulus) as participants in the passive category. The results from table 6.2 of this thesis imply that attending to or ignoring a stimulus can be detected beforehand.

General Discussion VII

7.1 Summary of Results

This thesis started with the research question "Can a data-driven approach be used to, without interventions and manipulations, predict human cognitive load?" and I can confirm that this is possible. Using a Random Forest machine learning model, I was able to outperform a random model (which randomly selected a label) and the most common model (which always selected the most common label). However, the model created during this thesis is not accurate enough to be used in safety-critical situations. Afterwards, I tried to improve my Random Forest model by using Recursive Feature Selection to select features that are important to label the data and remove redundant features to speed up the training and testing of the model. This was, however, not feasible because of time constraints, calculating the optimal features would take over a thousand years.

In the last experiment, I used a Fourier transform to calculate the power for the different bands for each condition. This yielded consistently significant results that I was able to distinguish between the driving conditions without an intervention. It also proved that it was possible to consistently distinguish between the active and passive category without an intervention.

7.2 Implications

In the existing theory, I did not find research that tried to predict cognitive load in a driving simulator. However, research into cognitive load, in general, is not novel. Klimesch et al. (1993) showed that alpha-band signals over the frontal and parietal regions tend to be relatively more sensitive to the attention demands over other regions. Again in Klimesch et al. (1997) and Klimesch (1997), they emphasized that alpha activity, especially in the higher frequency range, is associated with semantic information processing. In contrast with this, I did not find any significant distinguishing between the different conditions or categories, which might be explained by the fact that this is associated with semantic information processing.

I did find a consistent significance in the theta frequency band, which frequently has been related to episodic and working memory as theta Event-Related Synchronization increases parametrically with working memory load (Jensen and Tesche, 2002; Kahana et al., 2001). However, I found the opposite to their findings, where they report on a decrease in theta power when cognitive load increases, I found the opposite. Where I found a decrease in theta power when cognitive load decreased, they reported on an increase in theta power. However, again, they did not apply cognitive load in a driving simulator but with memory tasks.

7.3 Future Work

A follow-up could investigate if there are other machine learning algorithms that are less vurnerable to these vast amounts of irrelevant features. But, what might be more interesting is to continue with my third experiment. A follow-up could be to test whether the phenomenon that I found is not only consistent over the dataset used for this thesis but if it is consistent over multiple data sets.

Another way to continue is to experiment if it is possible to detect this phenomenon in real-time uses in a simulator, and, if so, is it possible to manipulate the participant into having more or less cognitive load. It could be even possible to test if when a participant has a low cognitive load, is that participant quicker to press a, e.g., button in case of a warning signal.

7.4 Conclusion

This study aimed to use a data-driven approach to see whether it was possible to predict different cognitive load conditions. As the results and discussion have shown, it is possible to distinguish between different cognitive load conditions. Even though the results from this study are not completely accounted for by the existing literature, I believe that further improvements in this field could bring an explanation as to why the results and existing literature do not match. Even though it was not the major goal of this study, it also showed that it is possible to distinguish between active and passive categories. It shows the importance of trying a non-traditional approach, and how it can provide new insights.

Bibliography

Anderson, E. W., K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva (2011). "A user study of visualization effectiveness using EEG and cognitive load". In: *Computer Graphics Forum* 30.3, pp. 791–800. ISSN: 14678659. DOI: 10.1111/j.1467-8659.2011.01928.x.

Antonenko, Pavlo, Fred Paas, Roland Grabner, and Tamara van Gog (2010). "Using Electroencephalography to Measure Cognitive Load". In: *Educational Psychology Review* 22.4, pp. 425–438. ISSN: 1573-336X. DOI: 10.1007/s10648-010-9130-y. URL: https://doi.org/10.1007/s10648-010-9130-y.

AOCNP, DCC (2015). "Watson will see you now: a supercomputer to help clinicians make informed treatment decisions". In: *Clinical journal of oncology nursing* 19.1, p. 31.

Artoni, Fiorenzo, Arnaud Delorme, and Scott Makeig (2018). "Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition". In: *NeuroImage* 175.February, pp. 176–187. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2018.03.016. URL: https://doi.org/10.1016/j.neuroimage.2018.03.016.

Bashivan, Pouya, M Yeasin, and Gavin Bidelman (Dec. 2015). "Single trial prediction of normal and excessive cognitive load through EEG feature fusion". In: *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. DOI: 10.1109/SPMB.2015.7405422.

Basti, Alessio, Vittorio Pizzella, Federico Chella, Gian Luca Romani, Guido Nolte, and Laura Marzetti (2018). "Disclosing large-scale directed functional connections in MEG with the multivariate phase slope index". In: *NeuroImage* 175.February, pp. 161–175. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2018.03.004.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Chen, Min, Shiwen Mao, and Yunhao Liu (2014). "Big Data : A Survey". In: *Mobile Netw Appl* January, pp. 171–209. DOI: 10.1007/s11036-013-0489-0.

Colclough, G. L., M. J. Brookes, S. M. Smith, and M. W. Woolrich (2015). "A symmetric multivariate leakage correction for MEG connectomes". In: *NeuroImage* 117, pp. 439–448. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2015.03.071.

Doroshenkov, LG, VA Konyshev, and SV Selishchev (2007). "Classification of human sleep stages based on EEG processing using hidden Markov models". In: *Biomedical Engineering* 41.1, pp. 25–28.

Ferrucci, D. A. (2012). "Introduction to "This is Watson"". In: *IBM Journal of Research and Development* 56.3.4, 1:1–1:15. ISSN: 0018-8646. DOI: 10.1147/JRD.2012.2184356.

Gevins, Alan, Michael E Smith, Linda McEvoy, and Daphne Yu (1997). "High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice." In: *Cerebral cortex (New York, NY: 1991)* 7.4, pp. 374–385.

Gramfort, Alexandre et al. (2014). "MNE software for processing MEG and EEG data". In: *NeuroImage* 86, pp. 446–460. ISSN: 10959572. DOI: 10.1016/j.neuroimage.

2013.10.027. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1016/j.neuroimage.2013.10.027.
Guidi, Gabriele, Roberto Miniati, Matteo Mazzola, and Ernesto Iadanza (2016). "Case study: IBM Watson analytics cloud platform as analytics-as-a-service system for heart failure early detection". In: *Future Internet* 8.3, p. 32.
Guyon, Isabella, Jason Weston, Stephen Barnhill, and Vladimir Vapnik (2002). "Gene Selection for Cancer Classification using Support Vector Machines". In: *Machine Learning* 46, pp. 389–422. ISSN: 10970282. DOI: 10.1002/bip.360320308.
Herzmann, Charlotte et al. (2017). "Using Functional Magnetic Resonance Imaging to Detect Preserved Function in a Preterm Infant with Brain Injury". In: *Journal of Pediatrics* 189, 213–217.e1. ISSN: 10976833. DOI: 10.1016/j.jpeds.2017.06.063. URL: https://doi.org/10.1016/j.jpeds.2017.06.063.
Inouye, Tsuyoshi, Kazuhiro Shinosaki, Akinori Iyama, Yuko Matsumoto, Seigo Toi, and Tsutomu Ishihara (1994). "Potential flow of frontal midline theta activity during a mental task in the human electroencephalogram". In: *Neuroscience Letters* 169, pp. 145–148.
Ishii, Ryouhei et al. (1999). "Medial prefrontal cortex generates frontal midline theta rhythm". In: *NeuroReport* 10.4, pp. 675–679.
Jensen, Ole and Claudia D Tesche (2002). "Frontal theta activity in humans increases with memory load in a working memory task". In: *European Journal of Neuroscience* 15, pp. 1395–1399.
Kahana, Michael J, David Seelig, and Joseph R Madsen (2001). "Theta returns". In: *Curr. Opin. Neurobiol* 11, pp. 739–744.
Klimesch, W (1997). "EEG-alpha rhythms and memory processes". In: *International Journal of Psychophysiology* 26, pp. 319–340.
Klimesch, W, H Schimke, and G Pfurtscheller (1993). "Alpha Frequency , Cognitive Load and Memory Performance". In: *Brain Topography* 5.3, pp. 241–251.
Klimesch, W, M Doppelmayr, Th Pachinger, and B Ripper (1997). "Brain oscillations and human memory : EEG correlates in the upper alpha and theta band". In: *Neuroscience Letters* 238, pp. 9–12.
Klimesch, Wolfgang, Michael Doppelmayr, and Simon Hanslmayr (2006). "Upper alpha ERD and absolute power: their meaning for memory performance". In: *Progress in brain research* 159, pp. 151–165.
Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521. DOI: 10.1038/nature14539.
Lee, Jay H., Joohyun Shin, and Matthew J. Realff (2018). "Machine learning: Overview of the recent progresses and implications for the process systems engineering field". In: *Computers and Chemical Engineering* 114, pp. 111–121. ISSN: 00981354. DOI: 10.1016/j.compchemeng.2017.10.008.
Leppink, Jimmie, Fred Paas, Cees PM Van der Vleuten, Tamara Van Gog, and Jeroen JG Van Merriënboer (2013). "Development of an instrument for measuring different types of cognitive load". In: *Behavior research methods* 45.4, pp. 1058–1072. DOI: 10.3758/s13428-013-0334-1.
Liu, Shuang et al. (2018). "Study on an effective cross-stimulus emotion recognition model using EEGs based on feature selection and support vector machine". In: *International Journal of Machine Learning and Cybernetics* 9.5, pp. 721–726. ISSN: 1868808X. DOI: 10.1007/s13042-016-0601-4.
Marvel, Cherie L. and Sergio Paradiso (2004). "Cognitive and neurological impairment in mood disorders". In: *Psychiatr Clin North* 27.1, pp. 1–17. DOI: 10.1016/S0193-953X(03)00106-0.Cognitive.

Menze, Bjoern H et al. (2009). “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data”. In: *BMC bioinformatics* 10.1, p. 213.
Muller, Matthias M, Thomas Gruber, and Andreas Keil (2000). “Modulation of induced gamma band activity in the human EEG by attention and visual information processing”. In: *International Journal of Psychophysiology* 38, pp. 283–299.
Obermaier, Bernhard, Christoph Guger, Christa Neuper, and Gert Pfurtscheller (2001). “Hidden Markov models for online classification of single trial EEG data”. In: *Pattern recognition letters* 22.12, pp. 1299–1309.
Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
Pitt, Mark A and In Jae Myung (2002). “When a good fit can be bad”. In: *Trends in cognitive sciences* 6.10, pp. 421–425.
Rossum, Guido (1995). *Python Reference Manual*. Tech. rep. Amsterdam, The Netherlands, The Netherlands.
Samuel, AL (1959). *Some studies in machine learning using the game of checkers. Reprinted in EA Feigenbaum & J. Feldman (Eds.)(1963). Computers and thought*.
Scheer, Menja, Heinrich H. Bülthoff, and Lewis L. Chuang (2016). “Steering Demands Diminish the Early-P3, Late-P3 and RON Components of the Event-Related Potential of Task-Irrelevant Environmental Sounds”. In: *Frontiers in Human Neuroscience* 10.March, pp. 1–15. ISSN: 1662-5161. DOI: 10.3389/fnhum.2016.00073.
Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
Shi, Guanya et al. (2018). “Neural lander: Stable drone landing control using learned dynamics”. In: *arXiv preprint arXiv:1811.08027*.
Simons, Daniel J and Christopher F Chabris (1999). “Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events”. In: *Perception* 28.9. PMID: 10694957, pp. 1059–1074. DOI: 10.1068/p281059.
Sutton, Richard S, Andrew G Barto, et al. (1998). *Introduction to reinforcement learning*. Vol. 2. 4. MIT press Cambridge.
Trevor, Hastie, Tibshirani Robert, and Friedman Jerome (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Vol. 2, p. 800. ISBN: 978-0-387-84857-0. DOI: 10.1007/b94608. arXiv: arXiv:1011.1669v3.
Van der Heiden, Remo M.A., Christian P. Janssen, Stella F. Donker, Lotte E.S. Hardeman, Keri Mans, and J. Leon Kenemans (2018). “Susceptibility to audio signals during autonomous driving”. In: *PLoS ONE* 13.8, pp. 1–17. ISSN: 19326203. DOI: 10.1371/journal.pone.0201963.
Wang, Jie, Zuren Feng, Na Lu, and Jing Luo (2018). “Toward optimal feature and time segment selection by divergence method for EEG signals classification”. In: *Computers in Biology and Medicine* 97.January, pp. 161–170. ISSN: 18790534. DOI: 10.1016/j.compbiomed.2018.04.022.
Wester, A. E., K. B.E. Böcker, E. R. Volkerts, J. C. Verster, and J. L. Kenemans (2008). “Event-related potentials and secondary task performance during simulated driving”. In: *Accident Analysis and Prevention* 40.1, pp. 1–7. ISSN: 00014575. DOI: 10.1016/j.aap.2007.02.014.
Yoshida, Yutaka, Hayato Ohwada, Fumio Mizoguchi, and Hirotoshi Iwasaki (2014). “Classifying cognitive load and driving situation with machine learning”. In: *International Journal of Machine Learning and Computing* 4.3, p. 210.
Young, Kymberly D., Vadim Zotev, Raquel Phillips, Masaya Misaki, Wayne C. Drevets, and Jerzy Bodurka (2018). “Amygdala real-time functional magnetic resonance imaging neurofeedback for major depressive disorder: A review”. In: *Psychiatry*

and Clinical Neurosciences 72.7, pp. 466–481. ISSN: 14401819. DOI: 10.1111/pcn.12665.