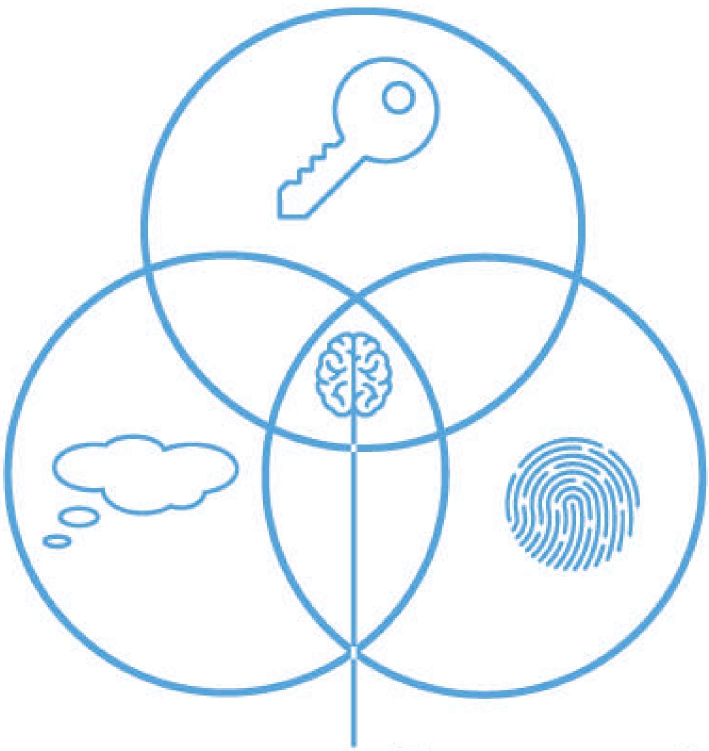
Bachelor Graduation Project



3-Factor Authentication

Using a Dry-Electrode in-ear EEG recorder

TU Delft

3 Factor Authentication Using a Dry-Electrode In-Ear electroencephalography recorder

A research on the feasibility of several classifications methods

Students:

Aanhane, P.J.	4644581
Molenkamp, S.H.	5163234
Vrijdag, J.A.J.M.	4719026

Supervisors:

dr. D.G. Muratore
dr. T. Costa
Cyril Weustink
Patricija Burgar

Faculty of Electrical Engineering, Mathematics and Computer Science
Technical University of Delft
Netherlands

Abstract

This bachelor end project thesis is describing ways to use dry-electrode electroencephalography (EEG) measurements for authentication. There are five proposed methods based on the literature study, which in this report are called: 'frequency tagging', 'pseudowords', 'familiar music', 'mental tasks', and 'emotions'. First, all methods were tested, however, this proved too much work to perform each of them thoroughly and well-bounded, so two were selected for further investigation.
For the method involving frequency tagging, experiments were done and the obtained data was used to extract features. This analysis showed that tagging can be discovered in the EEG data, but re-induction of tagged words was hard to achieve, which would make it difficult to create an authentication system reliant on it.
Analyzing results from experiments to test the pseudoword method was more promising. Some similarities between the features that were expected from papers and features from our results were found. These features were therefore used for classification, and resulted in small improvements in accuracy. This accuracy did however vary a lot between different data sets.

Preface

Working on the bachelor's final project was quite different from every quarter that we experienced before at the bachelor of Electrical Engineering. During courses, most work could be done whenever and wherever we liked, which meant a lot of freedom regarding planning.
The freedom that was lost in planning our personal lives, was however gained in how we would like to spend this project. At first this was daunting because we did not have the clearest idea of what was expected of us. After a while however, thanks to our supervisors, we realized that the most important thing to do was to do research well and that it was up to us to determine how exactly we wanted to achieve this.
In the end, we learned a lot during this experience. It did require a lot of work and there were disappointing moments when results were not as expected. This was however a small price to pay for the interesting research we were given a chance to do.
We would like to extend our gratitude to our supervisor Dante Muratore for the guidance through the course of the project, initially with setting our goals and requirements for the project and later with the direction in which we deferred from our original plan. We would also like to give special thanks to Cyril Weustink for helping us with the OpenBCI headset and for providing useful insights into the experimental setup. Finally, we would like to thank Tiago Costa and Patricija Burgar for helping to make this project possible. Lastly, we would like to thank our whole BAP group: Bonne Bogaert, Chris Bot, Kees Broek, Arthur de Groot, Luuk Mijjer, Mirthe Otter, Aurore de Spirlet, Jorn Teurlings, and Bart Zuidema.

Sam Aanhane
Simon Molenkamp
Joos Vrijdag

Contents

1
Introduction

The foundation for electroencephalography (EEG) was laid a century ago. Richard Caton and Adolf Beck established the presence of electrical brain activity in 1875 and 1890 respectively. Measuring electrical activity using electrodes that are placed on the skin results in a graph that is called an EEG. What is actually being measured is the activity of millions of cortical neurons that produce an electric field [1]. Hans Berger later introduced the EEG for human usage (1924) and is therefore, along with the former 2 scientists, regarded as the founder(s) of EEG. The main use of EEG was defined ten years later, when Fisher and Lowenbach demonstrated epileptiform spikes.
EEG evolved to clinical use, mainly to identify and detect epilepsy and other forms of seizure disorders [2]. It is also commonly used in psychological experiments, for instance in attention research due to the high temporal resolution of the measurements.

The classical way of recording EEG signals is with the use of wet electrodes distributed over the entire scalp. This is very useful in laboratory and clinical settings, but not suited for other situations during daily life.
A slightly more novel way of recording is using dry-electrodes, which are far easier to use. They are however still very invasive when worn over the entire scalp, due to discomfort and a far from discrete look.
The next step is therefore dry-electrode in-ear EEG recording. This has the dry-electrode advantage of being easy to setup, but also has less impact on the wearer, because they are worn like regular earphones. The current status of the development of these types of recorders is that they are less precise, especially when it comes to spatial resolution [3]. The goal of the project is to aid in the creation and applications of such a device. The different subgroups are listed below and can be seen in figure 1.1.

- **Stimulation:** This group is tasked with creating a circuit that can stimulate nerves around the ear, in order to prevent seizures.

- **Phantom:** This group is tasked with creating a fake ear that mimics the electrical properties of the human skin.

- **Signal Processing:** This group is tasked with processing EEG recording data in order to detect epilepsy seizures.

- **Authentication:** This group is tasked with creating a system which can identify a user based on an EEG recording.

The goal of this project is to devise a system that can identify users based on brainwave recordings. If a comfortable in-ear EEG recording device would exist, it could be used to Authenticate users of the device as a replacement for passwords and key-cards. Not only would such a system be convenient to use, it would also be extremely safe. In [4] and [5], a few security vulnerabilities are stated. Some of these vulnerabilities include poorly selected passwords, shoulder surfing, brute force and theft. An authentication system based on thought would be resilient against all of these attacks.

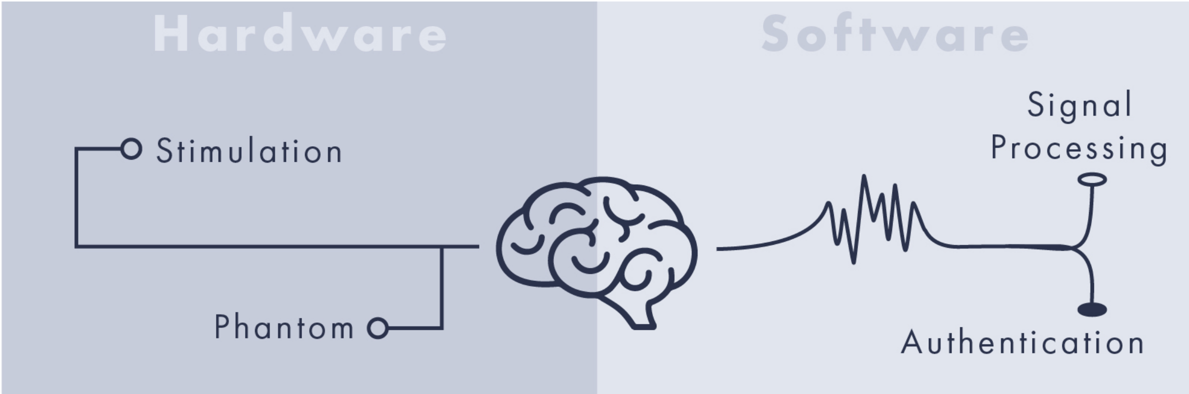


Figure 1.1: Overview of the four different subgroups working on the project

1.1. State-of-the-Art

There is already proof available that identifying a user using an EEG recording is possible in a lab setting [6] and in a real-world setting [7]. There are however still improvements to be made to the current state-of-the-art research. In a real-world setting, the authentication methods described [7] are about 90% accurate based on the situation. The algorithms needed to achieve such a high accuracy are also quite complex, which makes them hard to implement on a small device connected to the recorder. This is a disadvantage, because processing the signal on the device itself would make it nearly impossible to hack compared to sending all data to a phone, laptop or desktop.

1.2. Thesis synopsis

The goal of this project is therefore to find a way that makes dry-electrode in-ear EEG recording a secure and easy way to authenticate users. An extra benefit would be to use fairly simple algorithms, which might be possible to fit on a microcontroller. Using dry electrodes for EEG recordings results in a lot of noise, which needs to be removed. Artifacts, especially of the eyes, also pose a problem in order to use the data. That is why the first part, chapter 3, of this thesis is focused on preprocessing the collected data.
This data comes from recordings that were done using a dry-electrode scalp EEG recorder, because the in-ear variant was not yet available. The specific model used is the Ultracortex "Mark IV" EEG headset by OpenBCI [8]. In this report, five methods are proposed in chapter 4 to function as a means of authentication, which require experiment setups for which data was gathered. For two of the five methods, data was obtained for this report. The other three methods were not investigated due to time limitations.
From the data features are extracted, which are described in chapter 5. There are some general features that prove useful for almost all EEG signals such as the wavelet transform [9], but other features are dependent on the type of authentication method. Finally, these features are used for classification in chapter 6. The classification was both done with machine learning, as well as looking at the possibility of using filters and power comparison techniques.
An overview of the work done, combined with an overview of the content per chapter, can be found in Figure 1.2.

1.3. Equipment

In order to create a system that can be used to authenticate users using an EEG recording device, measurements are needed. This source of these measurements is an OpenBCI dry electrode scalp EEG recorder that can be connected to a computer through Bluetooth. The headset has 8 electrodes placed around the scalp on the locations given in figure 1.1 and table 1.2 resulting in 8 channels of signal data. The received signal is recorded using the corresponding OpenBCI software. The hardware consists of the following parts:

- Earclip Electrode
- Cyton + Daisy Biosensing boards (16-Channels)

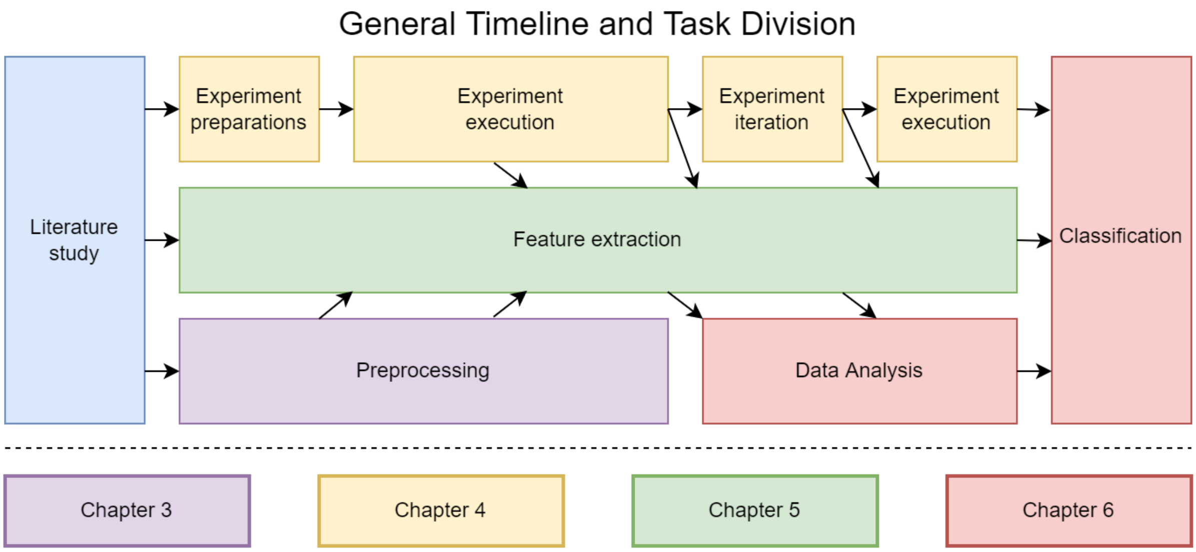


Figure 1.2: Global timeline on what was focused on during the project. Lines indicate the connection between consecutive steps of the process. The boxes in the bottom of the figure describe in which chapter each block is described.

- Dry EEG Comb Electrodes (Pack of 30)
- EMG/ECG Snap Electrode Cables
- Ultracortex "Mark IV" EEG Headset

In order to stimulate the brain, several experiments were created which could be done using the recording device and a computer. Subjects will be instructed about the experiment beforehand and receive the stimuli on the computer screen. The experiments were created using python and additionally stored all information regarding timestamps and labels, which could be used to classify the EEG data.

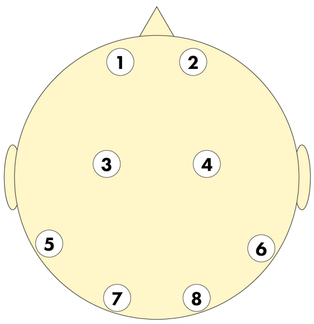


Table 1.1: Locations of the 8 electrodes of the EEG recording device.

Table 1.2: Overview of where electrodes record.

Electrode	Region
1	Frontal Lobe
2	Frontal Lobe
3	Frontal/Parietal Lobe
4	Frontal/Parietal Lobe
5	Temporal Lobe
6	Temporal Lobe
7	Parietal Lobe (Visual Cortex)
8	Parietal Lobe (Visual Cortex

2
Programme of Requirements

The final goal of this BAP is to create an algorithm that can perform binary classification based on electroencephalogram (EEG) data for authenticating purposes. This algorithm will serve as a proof of concept for future implementation in an in-ear EEG recording device. Therefore the requirements are chosen such that the classifier will be comparable with other login services e.g. password-based methods or biometric recognition.

The requirements are either mandatory or trade-off. Features and boundaries that are necessary for our algorithm to function properly are mandatory. When something could be achieved to improve the result but it is not necessarily needed, the requirement is a trade-off.

2.1. Functional Requirements

i The algorithm must remove artifacts and noise from the data. (mandatory)

ii The classifier must detect induced changes in the EEG data. (mandatory)

iii Based on the classifier, the algorithm must identify a unique user. (mandatory)

iv The algorithm must use only EEG data. (mandatory)

v The set-up needed to gather data for the algorithm consists only of an EEG recording device and software. (mandatory)

2.2. System Requirements

i The algorithm should run on intervals of 5 seconds of data. (mandatory)

ii The algorithm should take at most 3 seconds to run. (trade-off)

iii The algorithm should have a True Positive Rate of at least 80%. (mandatory)

2.3. Security Requirements

i The algorithm should have a False Positive Rate of at most 1%. (mandatory)

3

Preprocessing of EEG Data

An EEG is a useful test to measure the electrical activity inside the brain. However, due to the nature of this measurement technique, electric fields from different sources will always make unintended contributions to the gathered measurements. A distinction can be made which defines two different categories for these undesirable contributions:

- **Noise:** These are small and mostly random fluctuations in the electrical signal caused by e.g. physical properties of the conducting medium and bad placement of the electrodes.
- **Artifacts:** These are meaningful electrical signals which are picked up on EEG recordings but are not generated by the brain. The contribution of these artifacts can be several magnitudes higher than the actual EEG signal and therefore greatly reduce the signal-to-noise ratio.

This chapter will introduce the type of data this project will deal with, as well as describe the techniques used to filter the data in order to retrieve the meaningful EEG signals. The general pipeline to filter the data will also be presented.

3.1. The EEG signal

As mentioned in the introduction, the EEG data this project is concerned with comes from the OpenBCI recording headset. The headset has 8 electrodes providing 8 different measurement channels. The sample rate of the device is 250 Hz. Figure 3.1 shows an example of 8 channels with ideal EEG data.

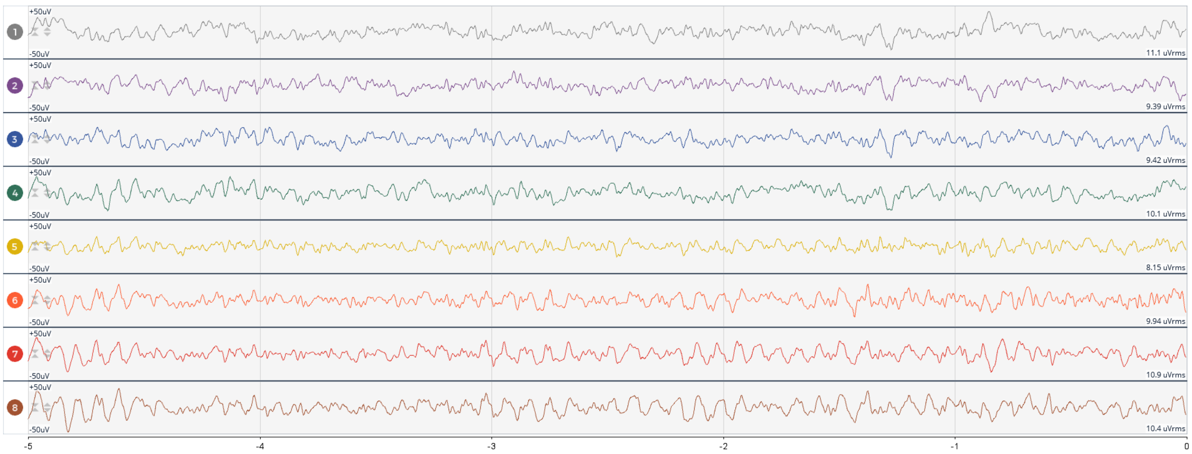


Figure 3.1: Example of an EEG recording with the OpenBCI recording device and software. The data was processed using the build-in filters to create this plot.

The amplitude of the signal is within a range of $\pm 50 \mu V$ and contains 5 major bands of frequencies:

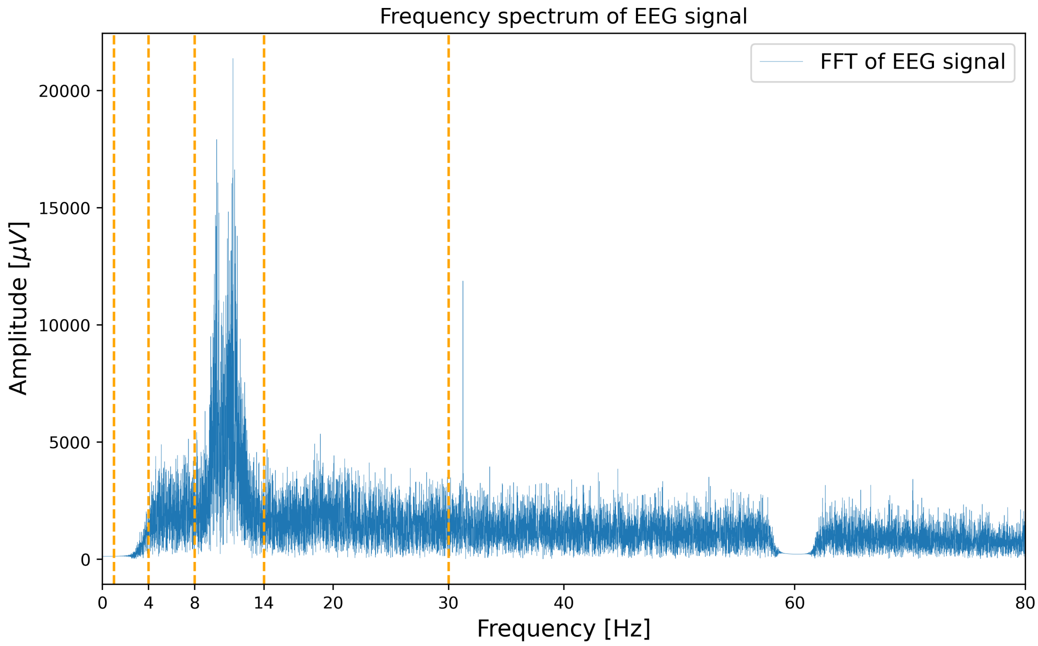


Figure 3.2: Fourier Transform of the example EEG signal 3.1. The dashed lines represent the frequency band boundaries.

Band	**Frequency**	**Brain states**
Delta (δ)	0.5-4Hz	Deepest meditation and dreamless sleep
Theta (θ)	4-8Hz	Deeply relaxed, inward focused (meditation, memory, intuition)
Alpha (α)	8-12Hz	Very relaxed, passive attention (calmness, alertness)
Beta (β)	12-35Hz	Anxiety dominant, active, external attention (decision making, problem solving)
Gamma (γ)	>35Hz	Concentration (simultaneous processing of information)

Table 3.1: Brain-wave Frequency-Bands, ([10], [11])

Figure 3.2 reflects the information in Table 3.1 with a peak in activity in the alpha band range. Namely, this was recorded with the test subject's eyes closed in a relaxed state.
The data collected from the experiment is far from ideal. It contains many distortions and is therefore difficult to analyze. Figure 3.3 shows 5 seconds of raw EEG data collected with the recording device. The frequency spectrum of the raw recording can be seen in figure 3.4. Most of the signal power can be found in two bands of approximately 5 Hz, namely 0-5 Hz and around 50 Hz, indicating noise that needs to be removed.

3.2. Sources of Noise and Artifacts

As mentioned at the start of this chapter, there are different sources of noise and artifacts which disturb the signal. Due to the nature of electric recordings, all nearby sources of electric activity will influence the measurements.

Powerline noise

As described in [12], powerline noise is a form of noise which cannot be avoided. It is characterized by a sinusoidal element of 50 Hz and its cause is the presence of electrical outlets in the recording room, which operate at the described frequency. The influence of powerline noise can be seen in figure 3.4. At the 50 Hz frequency bin, a large spike can be seen. The magnitude of this contribution is several times higher than the actual EEG signal, meaning that ii has to be removed from the recorded signal. Figure 3.5 shows the powerline noise waveform, isolated using a fourth-order bandpass filter. As can be seen from the figure, its maximum amplitude is 150 μV, which is about three times as high as the EEG signal.

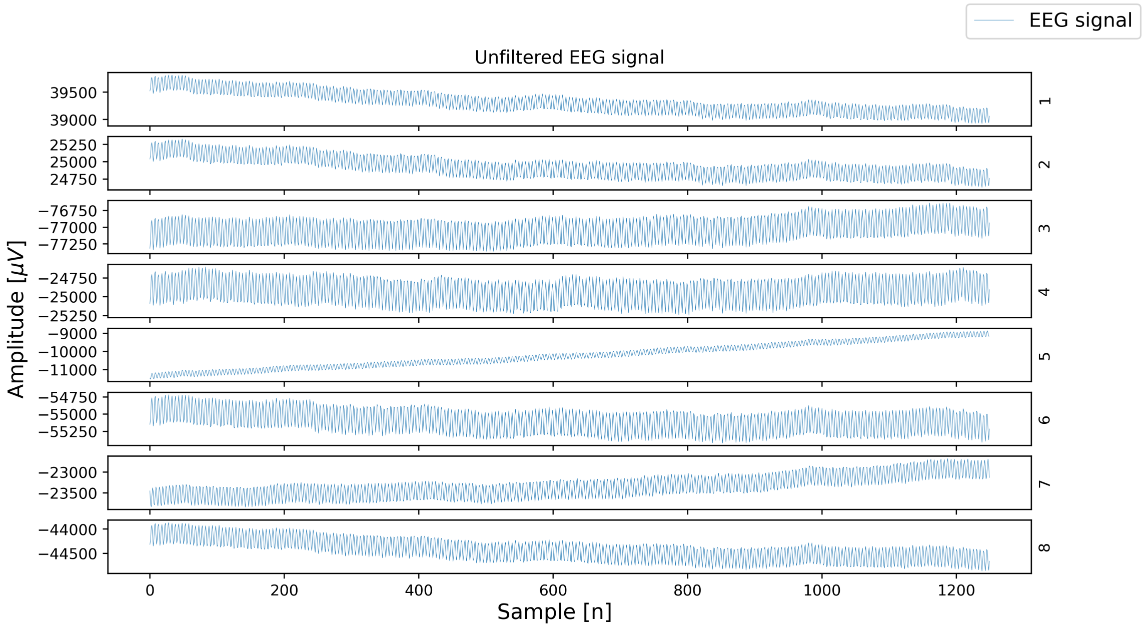


Figure 3.3: Plot showing the raw EEG data from the recording device.

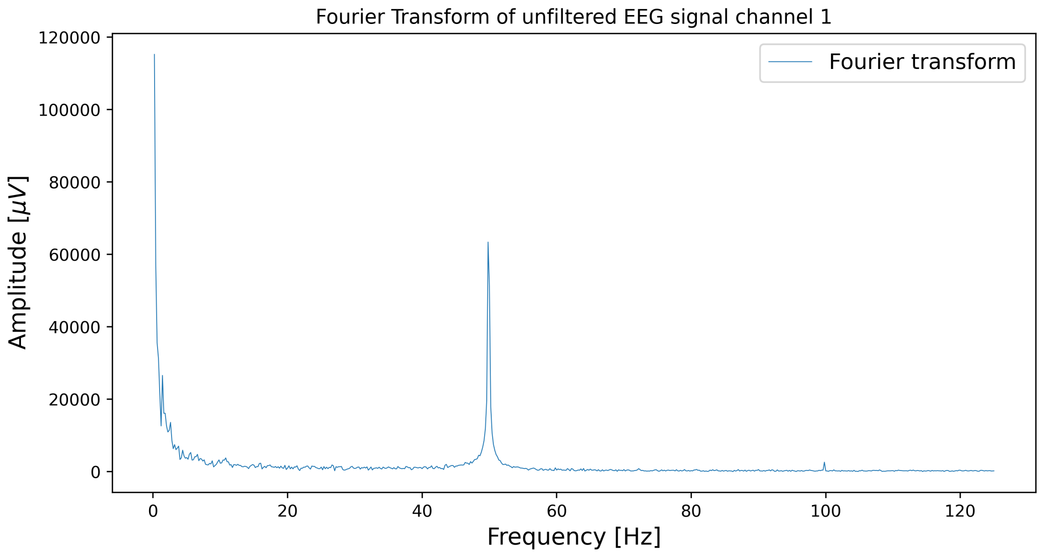


Figure 3.4: Fourier Transform of the raw EEG data. The DC bias (0 Hz) bin is excluded from this figure due to its large magnitude compared to the other bins.

Baseline noise

As stated in [12], poor contact of the electrodes and sweating of the patient under the electrodes may affect the electrode impedance which causes low frequency artifacts. Baseline drift may sometimes be caused by variations in temperature, bias in the instrumentation and bias in the amplifiers as well. This type of noise is undesired and needs to be removed before any further signal processing, so that proper analysis and display of the EEG signal is possible.

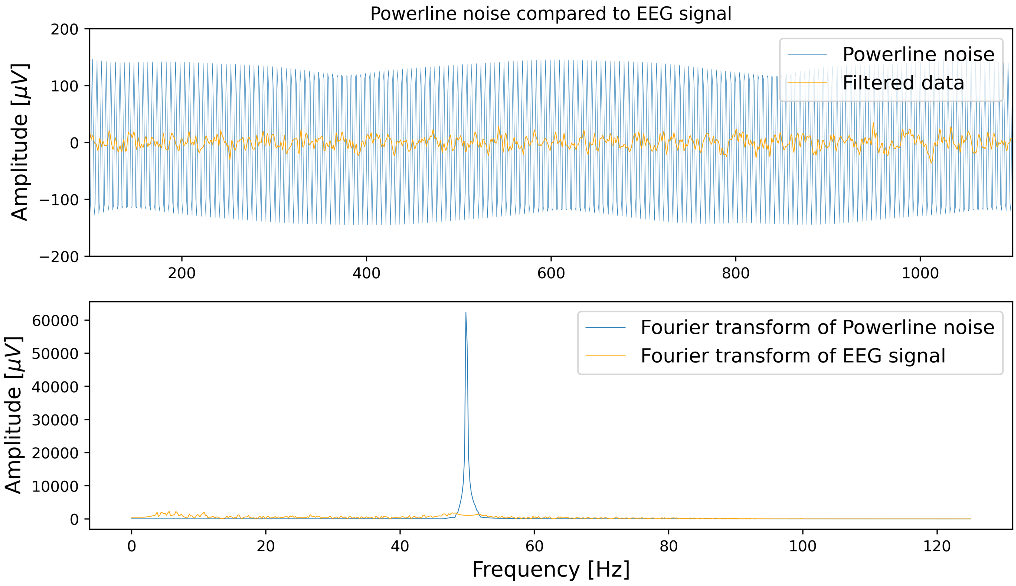


Figure 3.5: Plots showing the isolated powerline noise together with its frequency spectrum

Electromyogram (EMG) artifacts

EMG artifacts, as identified by [13], are disturbances of the EEG signal caused by movement. The movement can come from for example the jaws, facial muscles or sudden head movements. The artifacts are high in magnitude, especially when caused by general head movement, and occur on the frequency spectrum between 0 - 200 Hz. Figure 3.6 shows 4 seconds of data where the artifacts are present in channels 3, 4, 5, and 6. In this case, the artifacts were caused by jaw movement which is less noticeable at the frontal and rear electrodes, which is why they are omitted in the figure. Figure 3.9 shows the Fourier transform of these intervals of data. As can be seen from the figure, the artifacts cover most of the spectrum and are high in amplitude.

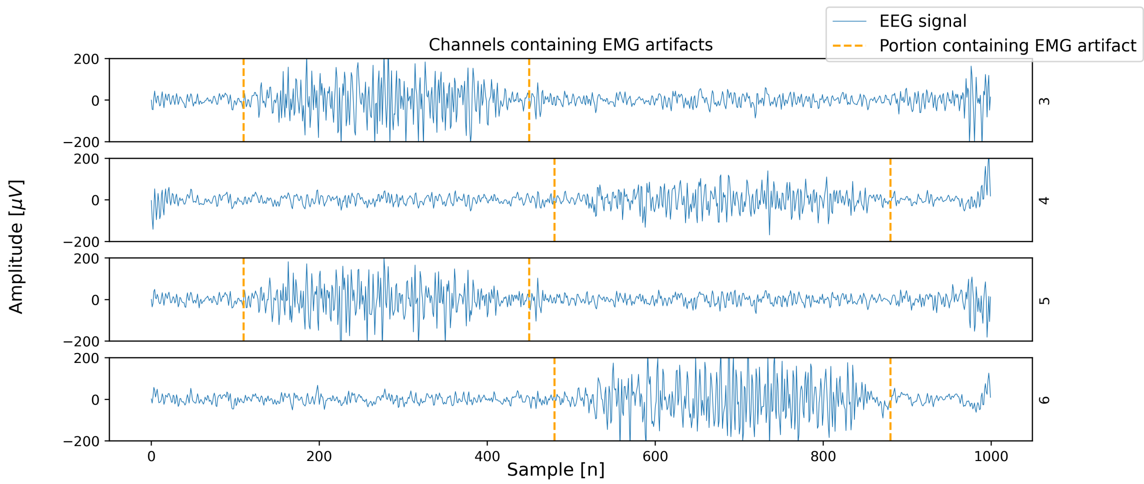


Figure 3.6: Four channels of EEG data containing EMG artifacts caused by jaw movement(channel 3 to 6, see Table 1.1), marked with orange.

Electro-oculogram (EOG) artifacts

EOG artifacts are disturbances of the EEG signal caused by blinking. The magnitude of the artifacts is approximately twice as high as the EEG signal. An example of these artifacts in the data can be seen in figure 3.8. The frontal channels mainly suffer from these artifacts, with the other channels showing no clear sign. The

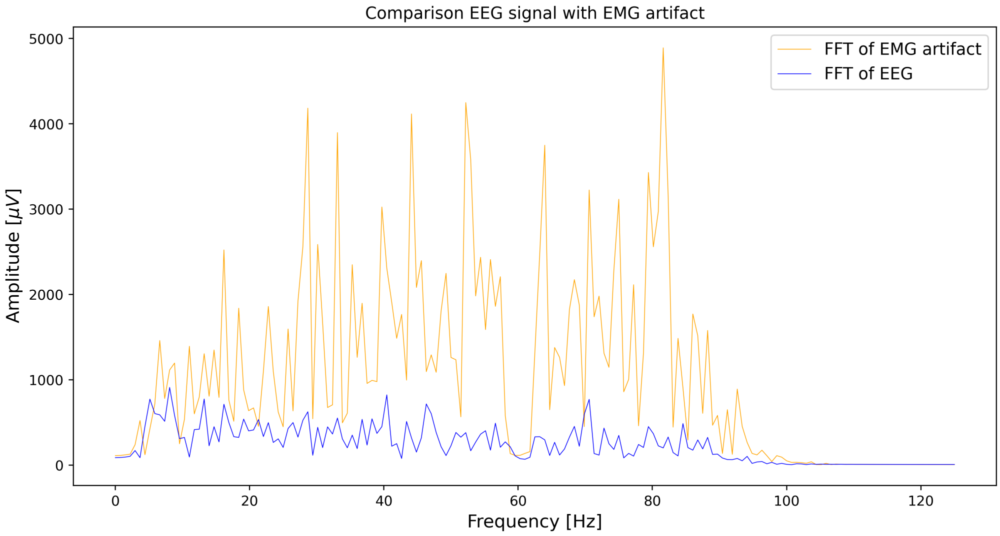


Figure 3.7: Fourier transform of the signal portions containing the artifacts.

frequency spectrum of the EOG artifacts is not very broad. It is limited by the speed at which a blink occurs. As stated [14], the average duration of a blink is about one-third of a second. Figure 3.9 shows the Fourier transform of a blinking artifact compared to a sample without one. This figure indeed shows that the artifacts mostly occur around 6 Hz.

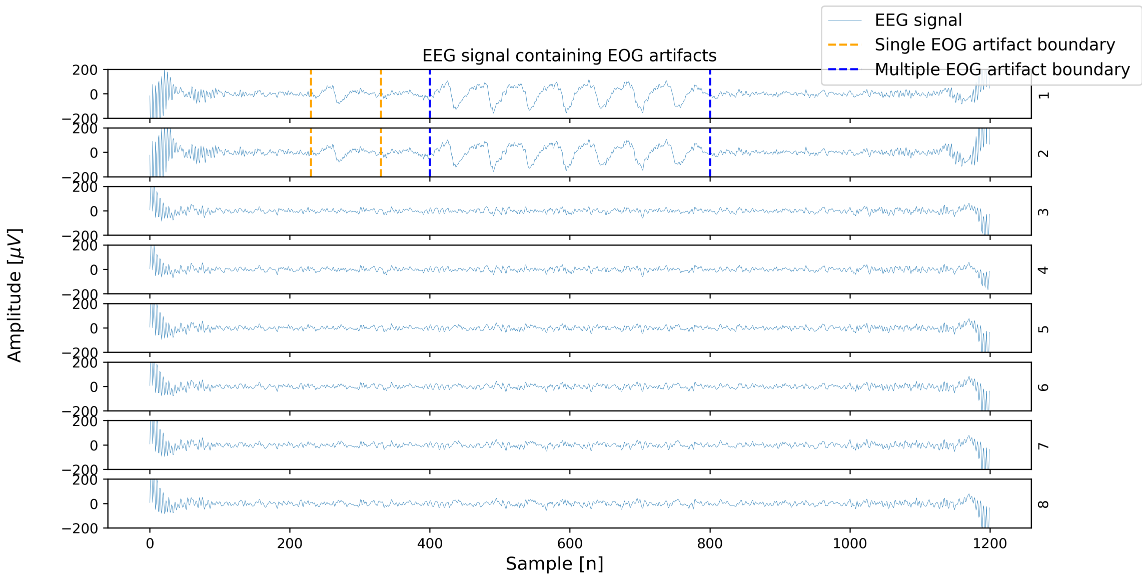


Figure 3.8: EEG signal which contains EOG artifacts. The portion in yellow contains a single artifact and the portion in blue contains 7 artifacts.

Electrocardiogram (ECG) artifacts

ECG artifacts are caused by the detection of electric potential around the heart being measurable across the surface of the scalp. The artifacts result in peaks in the EEG data which occur periodically between one and half a second, which originates from the fact that this is approximately the resting heartbeat of an adult. These artifacts were not identified or extracted during experiments, which is why no example is provided.

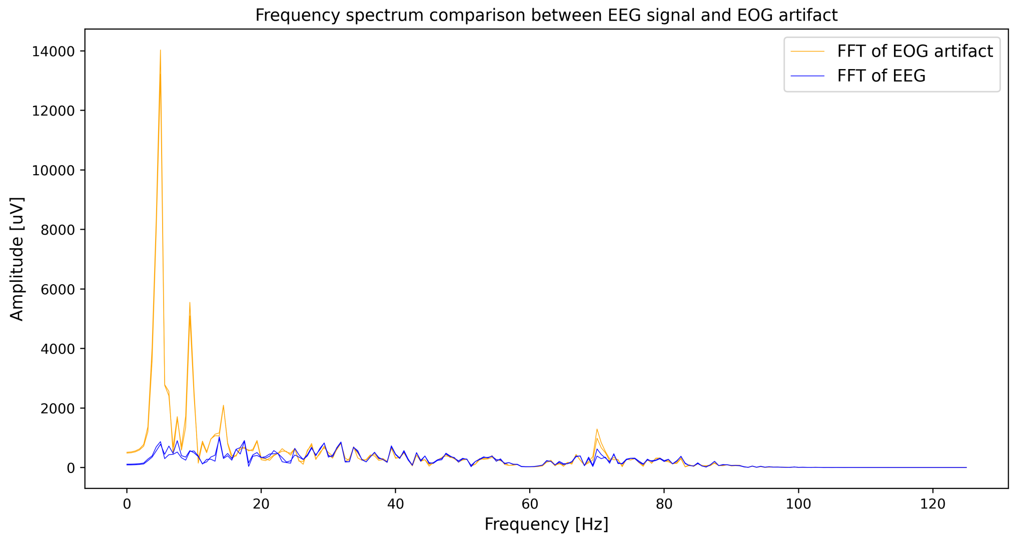


Figure 3.9: Fourier transform of the signal portion containing the artifacts.

3.3. Filter Techniques

Based on existing literature different strategies were created in order to remove the noise and artifacts. Some required a straightforward approach while others were more difficult to remove automatically. Most of the random noise could be removed at once using linear Infinite Impulse Response (IIR) or Finite Impulse Response (FIR) filters. The solution to remove the artifacts consisted of actively detecting and removing bad channels in intervals.

Baseline noise

In order to remove the baseline noise, a bandpass filter was used with a critical frequency range of 5 to 50 Hz. This frequency was chosen because little information can be extracted below this range. This is because the random 1/f noise is relatively high in that range.

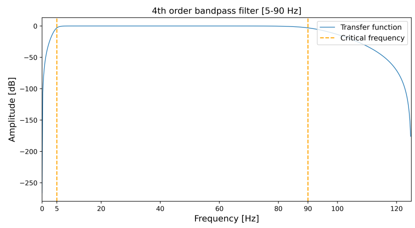


Figure 3.10: FIR Bandpass Filter to remove the baseline noise

Powerline noise

As mentioned in section 3.2, powerline noise mostly occurs at a frequency of 50 Hz. In order to remove this contribution, a fourth-order IIR notch filter around 50 Hz was used. This filter worked well and was easy to implement. The frequency response of this filter can be seen in 3.11

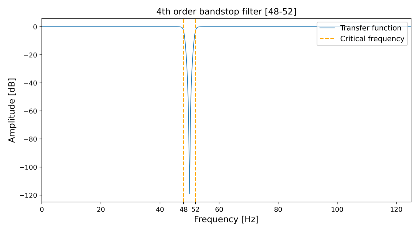


Figure 3.11: Bandstop Filter to remove the powerline noise

Artifacts

The artifacts are more difficult to remove. This is due to the fact that they cannot be predicted and occur in a broad frequency range, meaning they cannot easily be filtered. The problem can however be simplified if

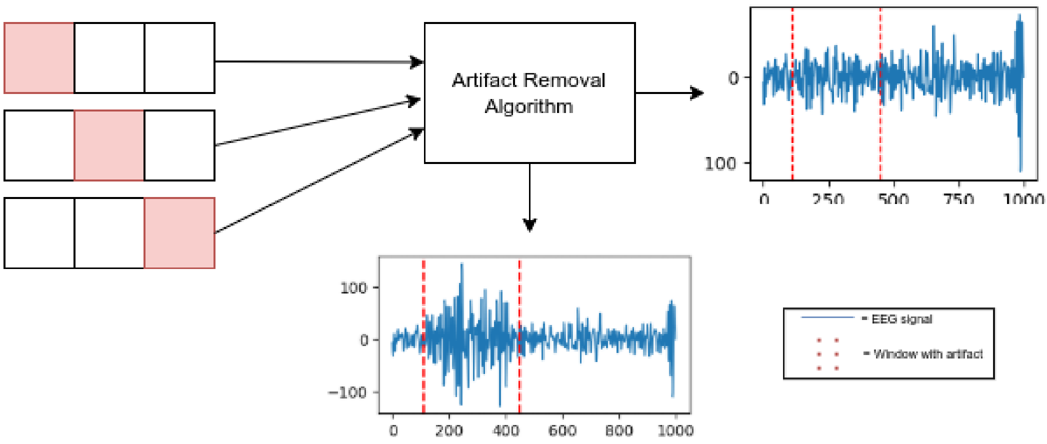


Figure 3.12

three facts about the possible artifacts are taken into consideration:

1. The amplitude of the different artifacts are orders of magnitude higher than the EEG signal and are of short duration.
2. Blinking artifacts predominately affect the two frontal channels, illustrated in figure 3.8.
3. Minor EMG artifacts can occur in every channel, but do not affect all channels, as illustrated in figure 3.7.
4. Major EMG artifacts affect every channel.

Based on these facts, a strategy can be devised in order to handle the artifacts. First, in contrary to the digital filters, the recording data is analyzed in intervals. Because artifacts are of short duration, they will mostly be contained to individual intervals and not spread over multiple intervals. The length of these intervals is elaborated on in chapter 5. Each of these intervals will still contain 8 channels. For each of these channels, the presence of artifacts can be detected based on their amplitude. The signal power can be calculated using the following formula:

$$P = \frac{1}{N} \cdot \sum_{n=0}^{N} x[n]^2 \tag{3.1}$$

Where $x[n]$ is the discrete signal value at sample n. When the calculated signal power crosses an experimentally determined threshold, the channel is marked for correction. Due to the fact that artifacts mostly affect only a few channels, the remaining channels can be used to correct the bad ones. Based on a method described in [15], the bad channels were replaced with the average of the good channels. This value is calculated using the following formula:

$$V_{cm}[n] = \frac{1}{K} \sum_{k=i}^{K} x_i[n] \tag{3.2}$$

Where K is the number of good channels and k the index of a particular good channel. This efficiently removed the artifacts without introducing a lot of false information. When a major EMG artifact occurs that corrupts all channels, the interval is rejected. This however usually only occurs during the beginning and at the end of a recording. When the measurements are done carefully, these bad intervals can be kept out of the experiment recording window.
The result of this strategy can be seen in figures 3.6 and 3.13. These are the four channels containing artifacts from figure 3.7. Based on visual inspection and the resulting frequency spectrum of the artifact interval shown in figure 3.14, the artifacts are successfully removed.

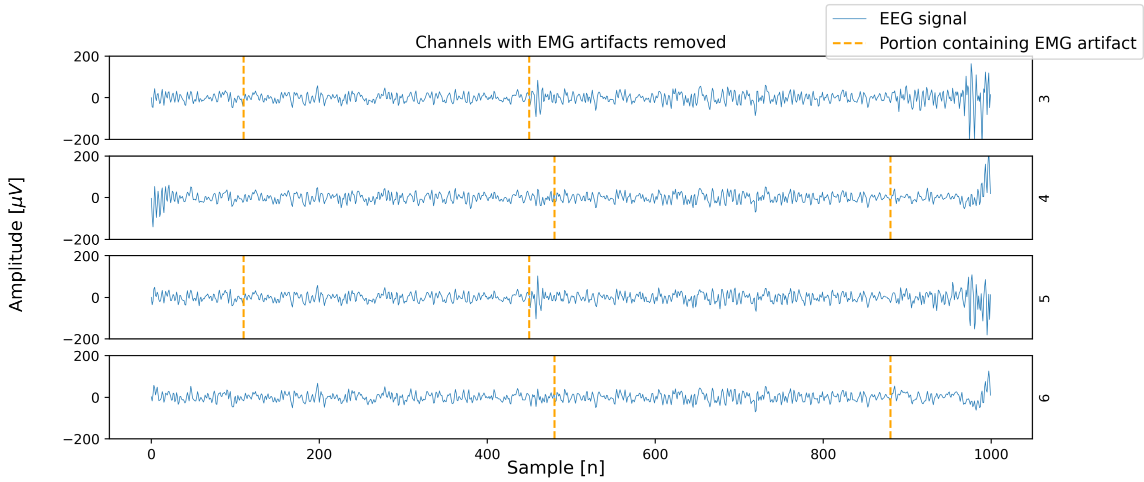


Figure 3.13: Four channels of EEG data after artifact removal. The old artifact locations are marked with orange.

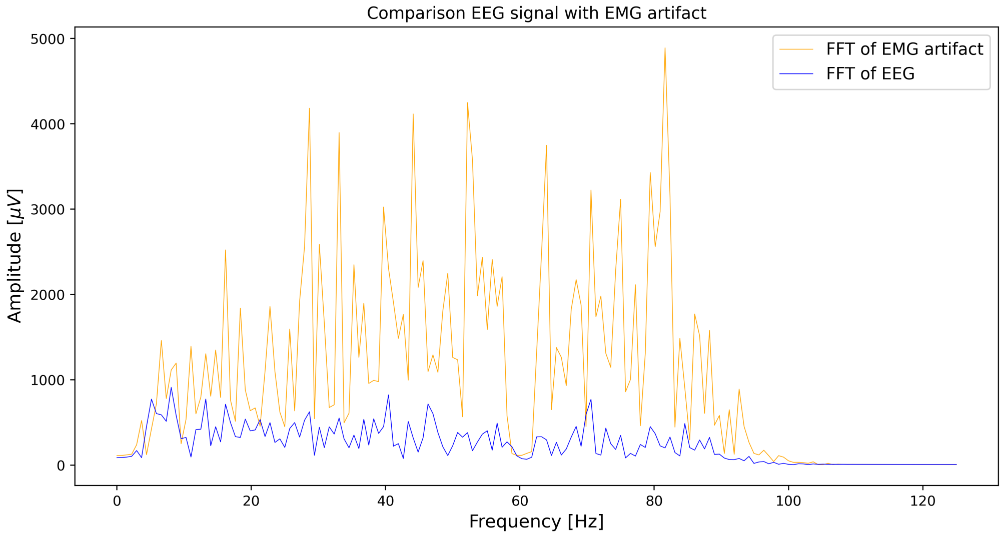


Figure 3.14: Fourier transformation of the interval in the first channel containing the artifact. The orange spectrum is prior to the removal algorithm and blue is the result after the removal.

4

Data Collection

In order to create a system that could be used to authenticate users using EEG recordings, it was necessary to gather data. Based on different stimuli, particular changes in an EEG signal could be detected which can, in turn, be used to identify a person. The first section in this chapter will present the general setup that was used to do measurements. After that, several options for stimulating the brain are discussed as well as their potential for achieving the desired results. Finally, not all experiments were executed. The ones that did will be discussed in more detail.

4.1. Stimulation methods

In order to classify a user, it is necessary to extract something unique for that person from the EEG recording. This is possible using data that is recorded during daily life [7], but it takes quite some time. Another disadvantage to this method is that a user has no control over when he wants to be authenticated and that the device both needs to be on as well as process data continually, which is not power efficient.
Through stimulation, the brain could create a signal which is unique to a particular user. This unique signal could be for example in response to hearing a particular song or seeing a particular word. If this can be detected, this change can be used to authenticate a user. Additionally, this method would be far quicker and power efficient. There are many possible ways the brain can be stimulated. For this report, five were examined in order to determine their feasibility. Using this information, a selection was made regarding which methods to prioritize. The description of these experiments are presented in the following sections.

4.1.1. Frequency tagging

A technique that is used in psychology studies to detect the recognition of stimuli is frequency tagging. It consists of the following stages:

1. **Tagging stage:** During the tagging stage, a subject is shown a word that acts as a password together with a flashing background. This flashing is done at a certain frequency and should evoke a detectable change in an EEG recording. These detectable changes are discussed in chapter 5.

2. **Recognition stage:** During the recognition stage, a subject is shown its password, but this time without the flashing background. As stated in [16], the EEG data should show a response comparable to when the tagging occurred.

In earlier research on frequency tagging, mostly scalp EEG recording devices were used. However, as stated in [3], Steady-State Visual Evoked Potentials (SSVEPs), which are a form of tagging, can be recognized by in-ear EEG recordings, so they are worth investigating.

For authenticating purposes, this phenomenon can be used to tag people with one or more passwords. When this password or a series of passwords is shown at a later time to an individual, and the particular EEG response is induced, this should only occur when a person is tagged. This means that the person is identified. For extra security, more tagged words with different frequencies can be added to create a longer password.

4.1.2. Pseudowords

Pseudowords are a series of letters that, when read or heard, sound like existing words but are in fact not. Words as such include "cigbet", "blonck" and "candinitially". According to literature, when these words are shown to subjects they elicit a different response in the brain compared to when a real word is shown. For example, as stated in [17], there is an increase in power in the gamma band around 30 Hertz [17].

One way this fact could be used to authenticate users is if it can be shown that the induced differences disappear when the pseudoword is taught to a user. This way, a pseudoword can be taught to a user and act as a password because only the user will experience it as a meaningful word. The meaning can come from a name or object very few people know or by training someone to give meaning to a pseudoword.

4.1.3. Familiar music

Another possibility would be to stimulate the brain through sound. When a person hears a song that is familiar to them, this creates a different response in the brain compared to an unknown song, as stated by [18]. This could already be detected within 350 ms.
This can be used for authentication purposes by letting people choose songs and measure their response to those songs. If these particular responses can be detected, a user can be identified.

4.1.4. Mental tasks

Mental tasks are already used as a way of creating a brain-computer interface (BCI) [19] for quite some time. When these mental tasks are performed, an EEG recorder can quite accurately classify among tasks. Each classified task is therefore two bits of information sent.
Signals sent by this type of BCI can be used to fill in a password to authenticate for the purposes of the research in this report. The bitrate of these methods is unfortunately very low, around 0.19 bit/s [20], which means it takes quite some time to authenticate a user this way.

4.1.5. Emotions

Emotions can be detected with a very high accuracy using EEG recordings, as stated in [21].
Detecting a certain emotion or series of emotions can therefore act as a way to authenticate a user. Emotions can be self-induced and have physiological effects, as mentioned in [22], so they can be elicited whenever the user would like to. Disadvantages are the unpleasantness people might experience during the induction of some emotions as well as the fact that the recording might have to last tenths of seconds in order to classify the different emotions.

4.2. Experiments

The plan was to set up experiments for each of the stimulation methods described in section 4.1. The first round of experiments used a general format. In an attempt to do a lot of experiments quickly, each stimulus or task was present on a screen for 5 seconds. However, the expected response to each of these stimuli was different. Some would result in changes that possibly could be detected throughout the entire recording. Other would result in induced differences that would only last a short duration. This proved problematic for the initial recording sessions.

This first round of experiments therefore did not provide the desired results. For none of the collected datasets, useful information could be extracted. After looking back at the literature study it was concluded that five seconds was long enough for each experiment in theory, but that some experiments only provided useful information within a second of the stimulation onset. Besides that, for most cases, the parameters of the experiment were not ideally chosen. For some of the methods collecting data also went too slow to generate datasets that were sufficiently large.

Extracting useful information from an EEG recording was harder than expected based on the literature study, so more tailoring of the experiments was required. The focus was therefore shifted from doing all experiments quickly to starting with one or two experiments and doing those well.
Frequency tagging was chosen for this because it is based on Steady-State Visual Evoked Potentials (SSVEP), which are relatively easy to detect in EEGs. It is also an experiment type that could be done executed efficiently in larger quantities, resulting in more data for machine learning.

The second experiment that was chosen to focus on, was to explore a method based on the difference in activity between pseudowords and existing words. This choice was made because the literature provided useful information on how exactly this could be detected, which did not require very complex algorithms. For this experiment, it was also relatively easy to collect sufficient amounts of data.

4.3. Frequency Tagging Experiment

In order to judge whether Frequency Tagging would be an effective method, the experiment was conducted in three stages:

- **Stage 1:** First it is important to verify that the reaction to the visual stimulus of the flickering background can be detected. This is investigated in the first stage.
- **Stage 2:** When there is proof that this can be measured, the next step is to tag subjects with their chosen or assigned passwords. It will then be investigated whether the changes can be induced without the visual stimulus.
- **Stage 3:** Finally an experiment will be conducted to check if the tagging effect remains after time has passed.

Stage 1: Response recognition
In this first experiment, the goal is to collect data in order to verify that the effects of Frequency Tagging can be seen in EEG recordings. The ideal situation to execute the experiments was determined based on [23], which states that the ideal stimulation method is to use a 15 Hz sinusoidal flash. However, due to practical limitations of computer screens regarding frame rate and brightness, a square wave stimulus was used, which is also possible according to the same paper. In terms of frequencies, it is time-consuming to test many of them. Therefore subjects were stimulated at frequencies of 15 Hz according to the paper as well as 6, 10 and 12 Hz. A visualization of an example of one of the experiments using a laptop screen can be seen in figure 4.1

Stage 2: Tagging
When there is sufficient evidence that responses to the flickering are measurable using the setup, a second experiments will take place to test the effects of frequency tagging. In a similar setup, a password is shown on the computer screen with a flickering background of a particular frequency. After a certain period, the tagging stops and the word disappears for a short time. After this interval has passed, the word reappears without the initial flickering. This can be used to see whether it is possible to detect changes in the EEG signal when tagged words appear.

Stage 3: Classify tagged words against untagged words - not carried out
For the final experiment, which is shown in Figure 4.1, it will be investigated whether it is possible to classify EEG data and differentiate tagged words from untagged words. The tagging will be done with six words, three of which are chosen by the subject and three of which are chosen at random. The effect of the tagging depends on how well the subject recognizes the word, so this separation in known and unknown words was made to see if subjects have an easier time recognizing known words. The words will be shown in random order with a flickering background at a certain frequency. Each word will be shown for 2 seconds with a 2-second interval in between words.

4.4. Pseudowords

Pseudowords can generate a different response in the brain compared to existing words. This difference should be noticeable after 320 ms after the word is shown. In order to determine whether this can be used to authenticate users, the experiment was conducted in two stages:

- **Stage 1:** First it must be investigated if the potentially evoked response from showing pseudowords to a subject is actually measurable by the dry electrode EEG recorder that is available.
- **Stage 2:** After that an experiment will be conducted to see if pseudowords can be taught to subjects and whether this alters the response.

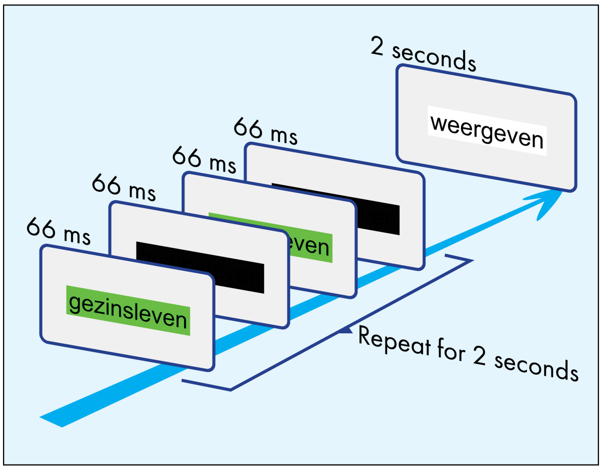


Figure 4.1: Shows screenshots for the frequency tagging experiment. First a random word is shown with flashing background for 2 seconds and then another random word without flashing. This is done for 40 words.

Stage 1: Difference words and pseudowords
The subject taking part in the experiment will have the OpenBCI EEG recorder while words are shown on a laptop screen. 40 words are shown, for 1 second each. The first 20 words will be real words the last 20 words are the pseudowords. The words selected are dutch, because that is the mother language of all participants and are randomly selected from two lists of around 55 real words and pseudowords for each experiment, which can be found in Table A.1. A visualization of the experiment is shown in figure 4.2.

Stage 2: Passwords - not carried out
When it is verified that with the used measurement set-up the changes in EEG that are expected can be detected, the next step is to test for authentication purposes. Subjects will get a list of 10 pseudowords prior to the experiment, which they are instructed to remember and more important give a meaning. After this is done, the experiment can take place. Participants will see 40 words on a laptop screen, 30 of which are real words and 10 which are words that were given prior to the experiment. The order of words and pseudowords is random.

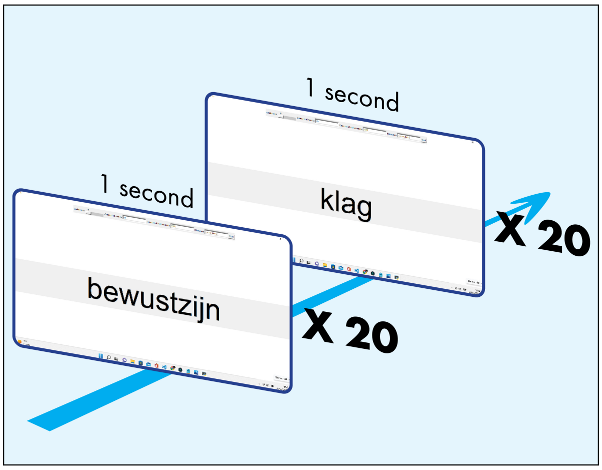


Figure 4.2: A series of 40 random words is shown. The first 20 are real words, the next 20 are pseudowords. Each word is shown for 1 second

5 Feature Extraction

Sampling brainwaves generates a large amount of data. The challenge that arises from this is interpretability. When looking at a raw EEG recording as in Figure 3.3, no features are apparent, and one can draw no conclusions from this plot. Even its frequency spectrum, as in Figure 3.4, is not illustrative. Moreover, after preprocessing, the signal might look something like Figure 3.14, and still no inferences can be made. This is where feature extraction comes in. There has been done extensive research on the operation of the human brain. Through this research, regions were discovered that contain increased activity during different kinds of stimuli, which is something that can be taken advantage of. Furthermore, EEG recordings themselves have also been widely investigated and useful features that become apparent through different experiments have also been discovered.
This chapter will elaborate on how features were extracted from the recorded EEG data that can provide useful insights into the performed experiments.

5.1. General EEG features

5.1.1. Brain-wave frequency bands

Brain waves are oscillating electrical voltages in the of small amplitude. There are five recognized brain wave frequency-driven sub-bands for human EEG signals. They are listed with the respective brain states they belong to in Table 3.1 .
It has been proven that significant differences in mean power per frequency band occur when measuring voltage at resting state versus when measuring whilst performing complex mental tasks [24]. This conclusion was also verified using the equipment at our disposal, see Figure 5.1.
Therefore, the choice was made to employ this distribution into frequency bands and then look at the average power in each band as a feature in the analysis of the performed experiments. These band powers were calculated using the Fourier transform of the signal, and subsequently dividing the spectrum into the mentioned bands, where the power in each band is equal to:

$$P_{band} = \frac{1}{N}\sum_{f}^{N} X(f)^2 \tag{5.1}$$

Where f is a frequency within the band and N it the total number of frequency bins.

5.1.2. Wavelet Transform

The wavelet transform is a transform that is used for time-frequency analysis. It is a transformation that splits data into different frequency components and examines each component with its resolution at that scale. The most important advantage of wavelet transform over the Fourier transform is that it exploits window sizes for optimal resolution. Namely, its window ranges from narrow for high frequencies to wide for low frequencies, thereby creating optimal time-frequency resolution [25].
The process of multilevel decomposition that the Discrete Wavelet Transform (DWT) employs performs a high-pass filter to produce the detail (D1) coefficients and a low-pass filter to produce the Approximate (A1) coefficients. The approximate sub-band diverges again, and this process is repeated recursively, following the

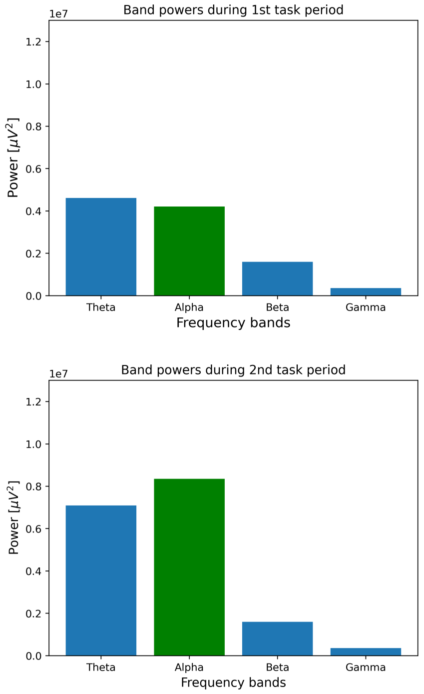


Figure 5.1: Band powers during two periods of focus.

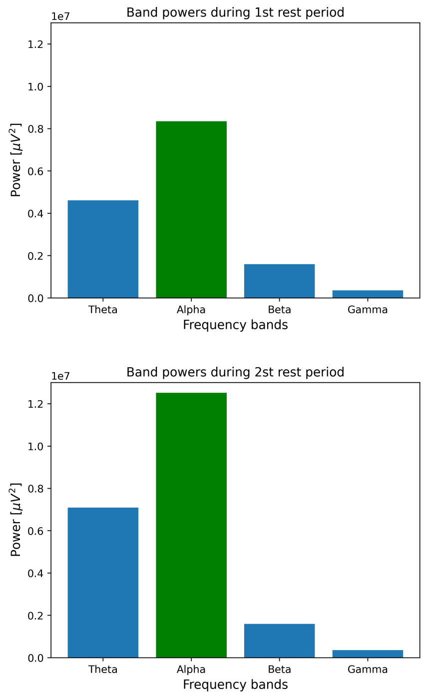


Figure 5.2: Band powers during two periods of relaxation.

structure in Figure 5.3.
In order to calculate the coefficients, the methodology of [26] was followed. This meant that out of the choices of different wavelet families and functions, the 2nd order Daubechies wavelet was chosen, and used to produce the coefficients.
From the wavelet transform, the coefficients as stated in Table 5.1 were further used. This selection was determined by following the methodology of [26], [27], and [28], who all commonly use the detail coefficients and the final approximation coefficient.

Sub-bands	**Frequency (Hz)**
D1	62.5-125
D2	31.25-62.5
D3	15.6-31.25
D4	7.8-15.6
A4	0-7.8

Table 5.1: Frequency bands used out of wavelet decomposition

Wavelet Statistics
Statistical features were extracted from the wavelet transform in order to reduce the size of the feature vectors. Namely, as a consequence of creating our own dataset, there was a limited amount of samples available. Therefore, in order to effectively train machine learning, the feature vectors needed to be reduced so underfitting could be prevented.
The features that were extracted are mean absolute value (MAV), average power (AVP), standard deviation (SD), variance, skewness and kurtosis as listed below.

$$\mathrm{MAV} = \frac{1}{k}\sum_{n=1}^{k}|x[n]| \tag{5.2}$$

$$\mathrm{AVP} = \frac{1}{k}\sum_{n=1}^{k}|x[n]|^2 \tag{5.3}$$

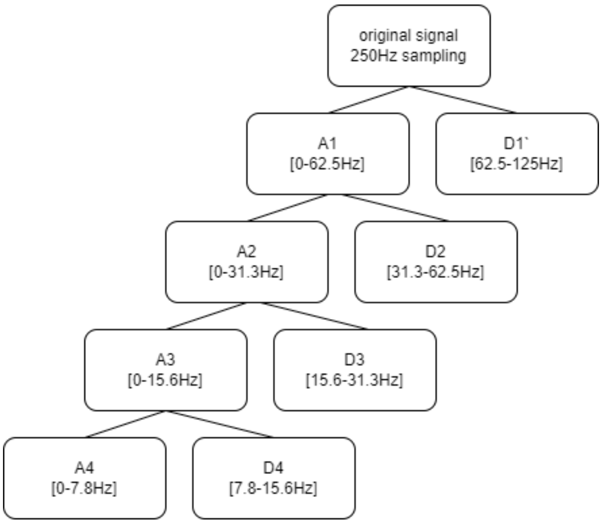


Figure 5.3: Wavelet Transform decomposition tree

$$\text{VAR} = \frac{1}{k-1}\sum_{n=1}^{k}|x[n]-\mu|^2 \tag{5.4}$$

$$\text{SD} = \sqrt{\frac{1}{k-1}\sum_{n=1}^{k}|x[n]-\mu|^2} \tag{5.5}$$

$$\text{skewness} = \frac{\sum_{n=1}^{k}(x[n]-\mu)^3}{k * SD^3} \tag{5.6}$$

$$\text{kurtosis} = \frac{1}{k}\sum_{n=1}^{k}\frac{(x[n]-\mu)^4}{SD^4} \tag{5.7}$$

These statistics thus lead to feature vectors per sub-band. This means 30 features per channel for 8 channels, which is 240 features per data segment compared to the raw signal which is 250 samples * 2 seconds * 8 channels which is 4000 data points for 2-second data segments. The statistics were henceforth passed on to the Data Analysis module.

5.2. Frequency tagging

Next to general features that are useful for analyzing EEG signals, from the literature study different features were found specifically for frequency tagging. The tagging itself is a form of Steady-State Visual Evoked Potentials (SSVEP), which manifests itself as an increase in frequency power around the tagging frequency.

5.2.1. Power around tagging frequency

As stated in [16], frequency tagging can be detected by an increase in power around the frequency of flashing. This power increases in a very narrow band around this frequency and can have a slight offset. Testing different bandwidths and offsets around the expected center frequency for classification can therefore be useful. This power increase is apparent the entire duration for which a flashing screen is shown, but only for a short time window (<1s) when reactivation without a flashing screen is attempted.

5.2.2. Peak-Average-Ratio (PAR)

To better assess the clarity of a peak in the signal, the PAR can be used. This ratio between the peak of a signal and the average power does not tell a lot about the size of the peak but is good at predicting whether it will be easy to pick up as a peak by a classification algorithm. It is useful for frequency tagging to have a value that tells how easily visible the peak around the flashing frequency is.

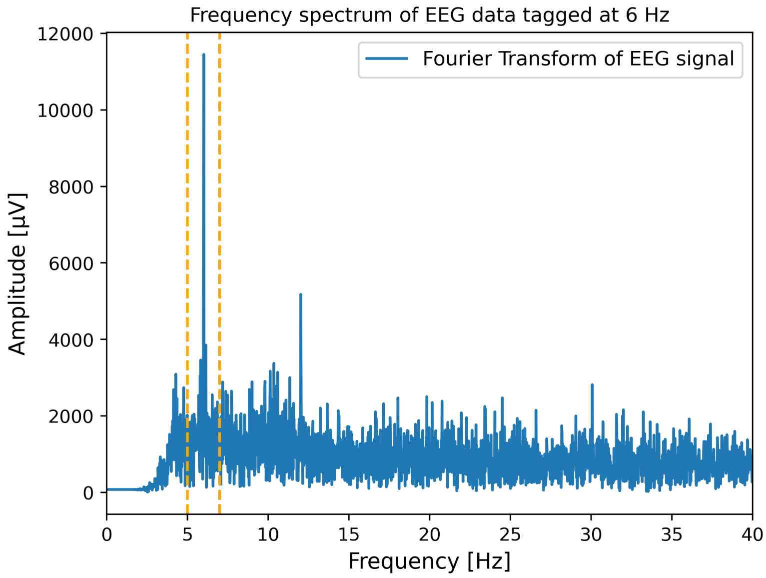


Figure 5.4: Frequency spectrum of tagging experiment with a flashlight. The tagging was done at a frequency of 6 Hz, which is the cause of the first spike in the spectrum.

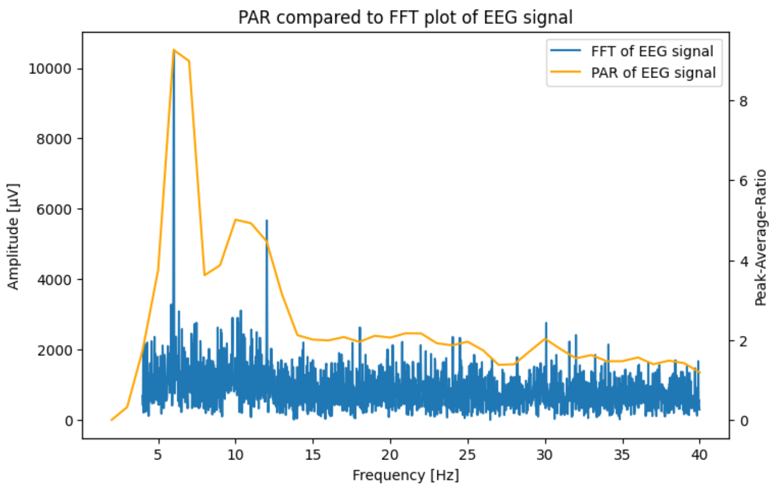


Figure 5.5: Frequency spectrum overlayed with the PAR values. The PAR values were calculated using bins of 2 Hz.

5.2.3. Cross-channel power ratio

SSVEP is also dependent on different brain regions. Although this will be less useful for a potential implementation of the algorithm on an in-ear EEG recorder because of the lower spatial resolution [29]. However, the data is available so it is interesting to consider for testing purposes. In [30] a topographical view of SSVEP is given. This shows that the signal will be more apparent around the visual cortex in the back of the head.

5.3. Pseudowords

From the literature it was also made clear that finding induced changes by pseudowords occurs around specific frequencies. In contrast to the frequency tagging, these changes are not always a power increase, but depend on the frequency, location and timing of the measured signal [17]. In the period of 320 to 520 ms after stimulus onset, the difference is clearest in the left hemisphere around 30 Hertz and in the right hemisphere around 40 Hertz. Around 30 Hertz the power drops below the baseline power for pseudowords, while the power around 40 Hertz rises above baseline power for normal words. These facts are summarized in

Figure 5.6

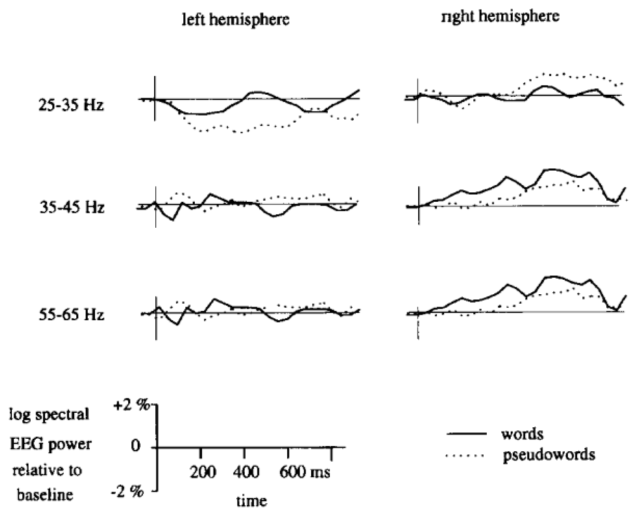


Figure 5.6: Image from [17] showing the differences in power in both hemispheres.

6
Data Analysis

In order to validate the experiment framework created, it is necessary to analyse the data manually. This way, the findings based on literary research regarding features can be compared to the gathered measurements. The experiment-specific features will therefore be determined under different conditions in order to judge the effectiveness of the experiments. The other features will not be evaluated in this respect, because it was not clear what their expected values would be based on the literary features. It would be difficult to find patterns for these features because they could arise from bad data. However, this does not rule out their effectiveness, and therefore they will still be used in the Machine Learning module at the final stage of this chapter.
To recall from chapter 5, the experiment-specific features are:

- **Frequency Tagging:** Increased power around tagging frequency.
- **Pseudowords:** Power differences in certain frequency bands and hemispheres, the most apparent of which is around 30 Hertz in the left hemisphere.

6.1. Frequency Tagging

To recall from chapter 4, the experiments based on the concept of frequency tagging were conducted in three stages. In the first stage, the goal was to verify that it is possible to see an increase in power around the frequency being tagged. For the second stage, it was investigated whether it is possible to re-induce a response to tagging without the visual stimulus. The goal of the third stage was to see whether it is possible to detect a response to a tagged word between untagged words. The final stage was not executed however for reasons explained in this section.

6.1.1. Stage 1

To confirm an increase in power caused by SSEVP around a predetermined frequency, two methods were explored in order to attempt to recreate this effect. One of these methods would consist of stimulating a subject through a computer screen. This method takes preference because it would simplify the remaining experiments. However, first a different method was tested based on the findings explained in [23]. This paper compares stimulation methods and concludes that while computer screens work, stimulation using LEDs creates a larger response. Therefore, as preliminary research, the measurement setup was first tested using a flickering flashlight at different frequencies.

These measurements were done at two different frequencies, 6 and 12 Hz in particular, and the frequency spectrum of these measurements can be seen in Figure 6.1 and Figure A.1 respectively. The orange boundaries in the first figure indicate the frequencies between 5 and 7 Hz. For the second figure, the orange boundaries indicate frequencies between 11 and 13 Hz. As can be seen from these figures, there is a large peak several times higher than the other frequency values within these boundaries at approximately the tagging frequency. These peaks are also present at multiples of the tagging frequency, albeit at a lower amplitude. These peaks in the frequency spectrum verify the ability to detect SSVEP tagging at particular frequencies in the case of frequency tagging using an LED light.

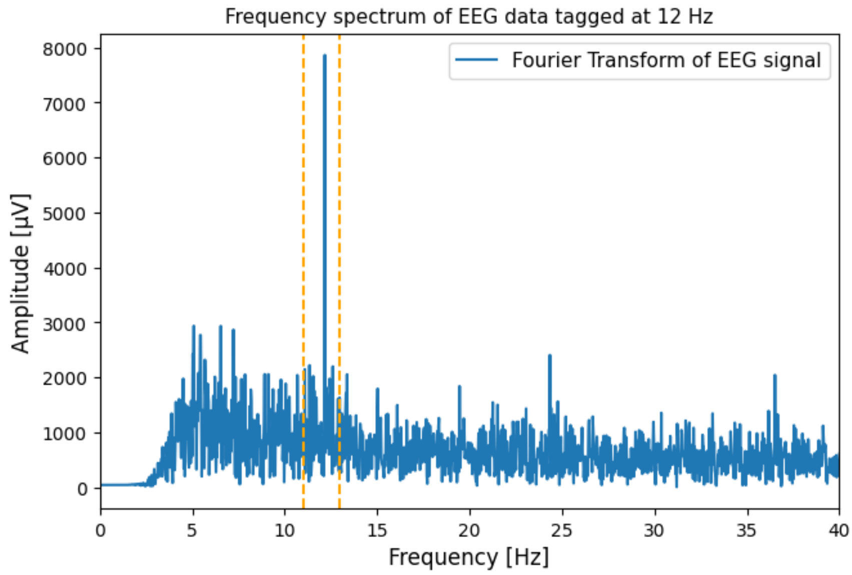


Figure 6.1: The frequency spectrum of a subject's EEG data stimulated using a flickering flashlight at 12 Hz.

After this result, the same experiment was conducted using a computer screen. This time, the tagging frequency was 10 and 15 Hz. The result of these measurements can be seen in Figure A.2 and Figure 6.2 respectively. In these plots, the orange boundaries again indicate the two integer frequencies around the tagging frequency. Within these boundaries, again a peak can be detected around the tagging frequency. These peaks relative to the remaining spectrum are lower compared to the LED case, but still detectable compared to the case when no stimulation is applied. The magnitude of these peaks is less important because they vary based on measurement duration, skin impedance and experiment subject. The detectability of these peaks indicate that the first stage of the experiment was succesful.

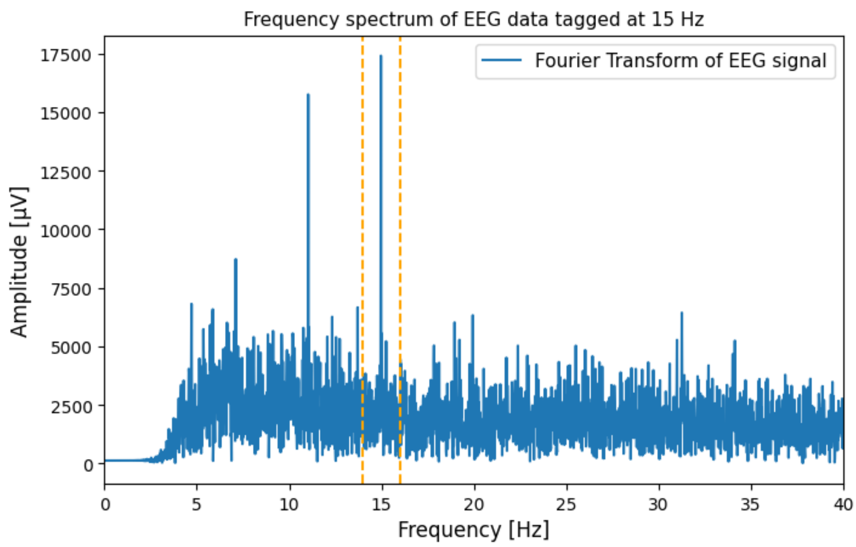


Figure 6.2: Frequency spectrum of a subject's EEG data stimulated using a computer screen at 15 Hz.

6.1.2. Stage 2

In order to assess the ability to re-induce the tagged frequency peaks as proposed, the second stage of the experiment was executed. For this stage, the measurement subject had to choose a word which would be tagged at a certain frequency. After the frequency tagging period, the word including the flickering would disappear for a certain period of time. After this interval, the word would reappear without the flickering background.

First, the effect of tagging duration was explored. In order to measure this a subject was stimulated for different time intervals. Afterwards, the frequency power at the tagged frequency was determined, as well as the Peak-Average-Ratio (PAR) at this frequency. Together, these measures describe the magnitude of the peak in relation to the other frequencies.

Figure 6.3 shows the result of these measurements. As can be seen from the figure, the magnitude of the tagged frequency power drops linearly based on the measurement interval. This is as expected. What is important to note is that the PAR stays approximately constant for tagging durations between 50 and 15 seconds. For shorter durations, the values start to diverge and decrease, indicating that the value of the peaks become comparable to the average of the data. This makes them difficult to detect. Based on this figure, any duration between 50 and and 15 seconds should give similar results. For the coming experiment, 50 seconds was chosen. This was done to increase the difference between the tagged frequency value and random spikes for other frequencies, simplifying potential visual inspection of the data.

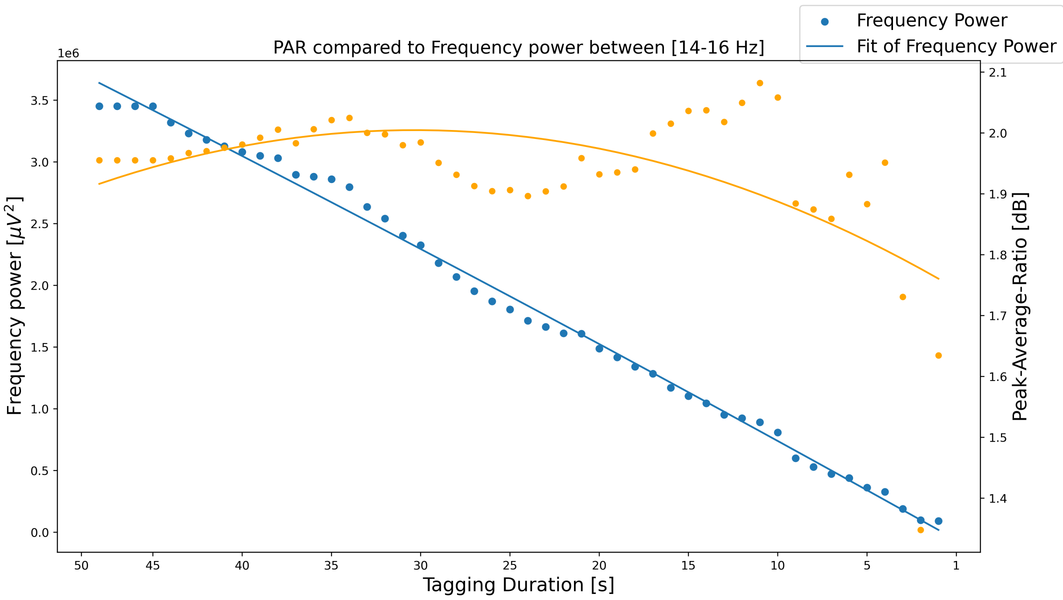


Figure 6.3: Relation between tagged power frequency and peak average ratio (PAR). This relation describes the visibility of the peak.

Now that a tagging duration has been decided, it can be investigated whether it is possible to re-induce the frequency peaks for tagged words. To do this, the stage 2 experiment described earlier was executed with a tagging duration of 50 seconds, based on Figure 6.3, at a tagging frequency of 15 Hz based on [23]. The parameter that was varied for this experiment was the time interval between the end of the tagging and the reapearing of the tagged word. By varying this value, it can investigated how long the effect of the tagged word lasts.

The results of this experiment can be found in Figure 6.4. The figure compares the power of the tagged frequency in channels 7 and 8 with the average power when the frequency is not tagged. This average was calculated using a calibration recording measured prior to the experiment. The relative power was calculated for the different waiting periods, as well as for different durations after the word reappears. This figure could show what was deduced from literature, which is that the power around the tagged frequency decays over time. It could also give insight into how long a frequency can stay "tagged".

Based on the relative power, the order can be seen in Table 6.1 This order appears quite random, but is does seem to imply that waiting for a longer interval is favourable compared to a shorter interval. Data for longer intervals than 10 seconds is not available, so it is difficult to draw conclusions. Additionally, the actual values

Table 6.1: Relative Power ordered by waiting periods

Relative power	waiting period
1	4 seconds
2	6 seconds
3	2 seconds
4	8 seconds
5	10 seconds

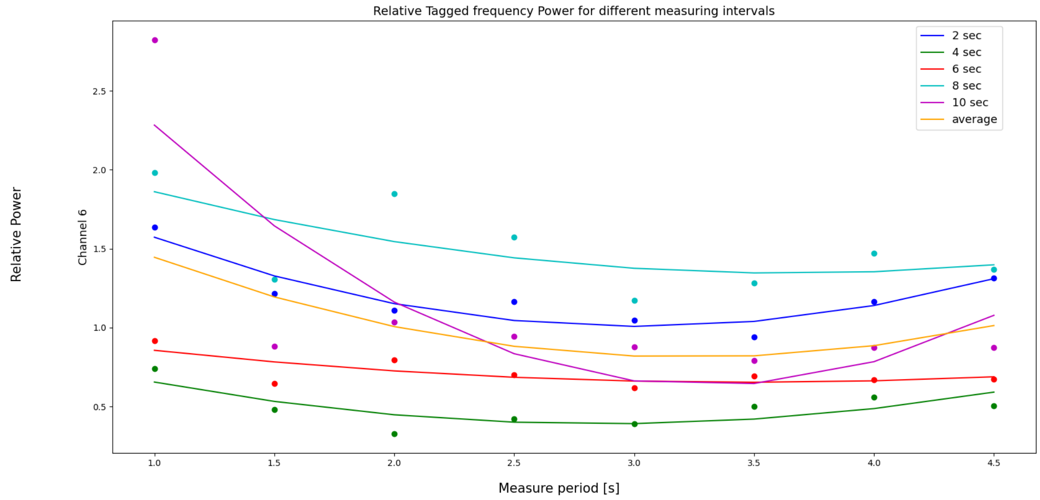


Figure 6.4: Relative power at the tagged frequency for different waiting intervals between tagging and reappearance

of the relative power are not taken into account yet.
In order to prove that a power increase can be detected, the single values are not enough because the increase in relative power can be caused by the randomness. An actual increase means that the value should decay over time, because this implies that a source caused an increase in relative power initially which is no longer there for the remainder of the measurement period. Looking at the figure, the relative power stays approximately constant for waiting intervals 2, 4, and 6. This seems to suggest that no increase in power is detected after the word reappears. For waiting periods 8 and 9 it is more difficult to draw a conclusion. For these intervals there does seem to be a downward trend. It is however possible that this trend is caused by outliers. The purple frequency power for the first second especially seems too high compared to the other values.

More measurements for longer intervals are required in order to better explain the effects of frequency tagging, but these were not gathered at the time of writing. The results were however concise enough to justify abandoning frequency tagging as method of authentication due to the difficulty in creating a system reliant on it.

6.2. Pseudowords

As explained in chapter 4, the main goal of the pseudowords experiment was to investigate whether it can be determined from EEG data if a subject is looking at existing or non-existing words. In order to do this, as mentioned in chapter 5, the frequency power in the left and right hemispheres are considered. For example, when looking at pseudowords, less power around 30 Hz is expected as shown in 5.6.

6.2.1. Stage 1

In an attempt to verify the hemisphere power difference, the described experiment was executed and the results were analysed. Two subjects were shown three times 20 real words and 20 pseudowords. In figure 6.5 the experiment trials are shown. In these figures, a heatmap can be seen showing different frequency bands on the left for the different electrodes on the bottom. The two rows both show a single subject who underwent

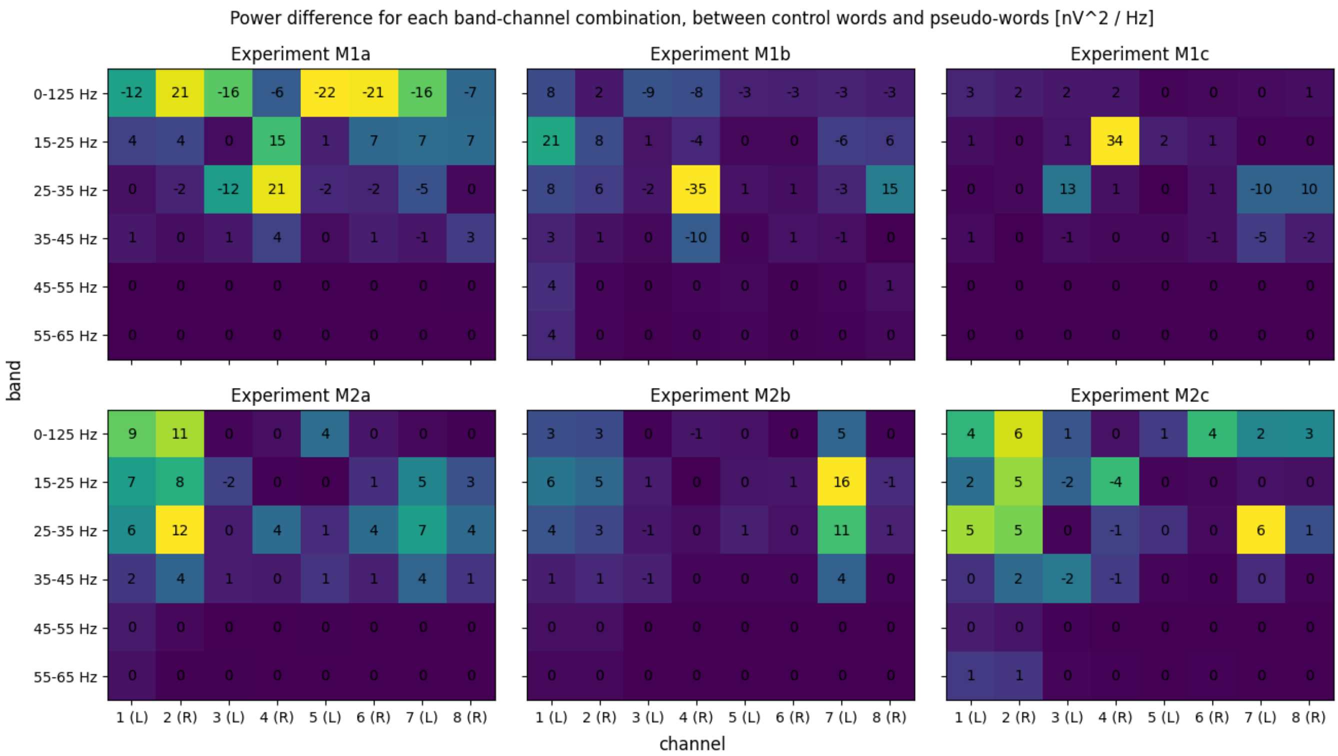


Figure 6.5: This figure shows an analysis for 6 experiments conducted on 2 persons (M1 and M2). Each experiment consisted of looking at 20 normal words and 20 pseudowords for 1 second per word. Both these 20 second segments were compared for different frequency bands at different channels of the EEG recorder. The difference in power between the two 20 second segments is shown in the figure. The hemisphere which the band measures is indicated with an L for left or R for right. Channel 1 and 2 are at the front of the head while 7 and 8 are at the back, the rest is in between those (exact locations can be found in figure 1.1).

three trials. The values indicate the difference in frequency power in nV^2, in the different frequency bands, between regular words and pseudowords.

For different bands and frequencies the power difference can be seen between a 20-second period of existing words followed by a 20-second period of pseudowords. This figure shows that the largest difference in power occurs at the right side of the head around 30 Hertz, which was one of the regions that was expected. There are however a few ifs and buts:

- The 30 Hz band power increase on the right side is not seen for every experiment.
- There are some cases were the power difference is also large at the back of the head (channels 7 and 8), which were not expected.
- In one case the 30 Hz band power difference is negative instead of positive.
- The exact location and frequency in which the power difference is greatest varies a bit between experiments.

From this analysis was concluded that it was possible to find at least some of the features mentioned in [17], but they are not as clear as expected. It is therefore concluded that the 30 Hertz band is useful, but not enough on its own for classification. Other features, such as the other bands shown in the plot, will therefore also be used for classification.

6.3. Classification using Machine Learning

Machine Learning can in many cases find patterns or divisions that are not humanly detectable. The principle of Machine Learning is to use exhaustive algorithms to try and classify data that is obscure/dense because computers/machines are much faster than humans at computations.

Therefore, Machine Learning was also implemented to attempt classifying the data. Multiple popular machine learning algorithms were tested:

- **Decision Tree**: Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.
- **K Nearest Neighbour(KNN)**: Supervised Machine Learning that uses proximity to make classifications or predictions about the grouping of an individual data point. For every sample, it finds the k number of samples closest to the query and then votes for the most frequent label in those samples.
- **Logistic Regression**: Supervised Machine Learning used to predict the probability of a target variable using statistical sigmoid function.
- **Support Vector Machine(SVM)**: Supervised Machine Learning, which uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. It translates the real problem into high dimensional feature space and then finds a non-linear boundary in the real space by finding a linear boundary in the high dimension space.

From literature, SVM was identified as the most promising algorithm for the classification of EEG signals [31], [32], [28], [33]. Namely, all these papers achieve highest or near highest accuracy (and other performance measures) using an SVM. Moreover, these studies are especially relevant because they all investigate different information extracted from EEG data, namely both mental tasks [32], Steady State Visual Evoked Potential(SSVEP) [31]- in particular frequency tagging- and epilepsy [28].
However, papers that provide comparative analysis between Machine Learning algorithms for EEG purposes illustrate how the other three (decision tree, knn, logistic regression) are also worth looking into [27].

6.3.1. Prinicipal Component Analysis (PCA)

The dimensionality of the data is very high. For each sample which needs to be classified, many features are available, though not all of these are relevant. Some features contribute little when trying to classify a sample. Therefore, the dimensionality of the data needs to be reduced. Principal Component Analysis (PCA) can be used to achieve this.
PCA is a statistical procedure that allows you to summarize the information content of large chunks of data in a smaller set of summary indices, that can be more easily visualised and reduce the dimensions of your data. PCA attempts to find the optimal rotation of your axes as to enable dimensionality reduction with minimal loss of variance (statistical measure for information). As can be seen in Figure 6.6, visually, PCA rotates the axes of the data in such a way that less axes (in this case 1 instead of 2) are needed to explain the information, and thus dimensionality reduction is achieved. Mathematically speaking, it does this through the following steps:

- Standardizing the data: Through subtraction of the mean and dividing by the standard deviation every sample (x_i), the data obtains a new mean of 0 and is scaled down by its standard deviation Equation 6.1.

$$z = \frac{x_i - \mu}{\sigma} \tag{6.1}$$

- Calculating covariance matrix: The covariance matrix is a matrix with the same number of rows and columns as the original dataset. It explains how the features differ from each other by calculating the covariance between pairwise means.
- Calculating eigenvalues & eigenvectors: From the covariance matrix the eigenvectors and eigenvalues are calculated. The eigenvectors point in the direction of most explained variance. The magnitude of the corresponding eigenvalues tells us how much of the variance is explained by that component.
- Reduce dimensionality: The first Principal Component is thus determined by the eigenvector corresponding to the eigenvalue with greatest magnitude. The principal components are ranked in this manner. One can then choose to discard principal components that only explain a small part of the

variance. In this project, the choice was made to keep 99.9% of the variance. This was possible using only the first 10 Principal Components.

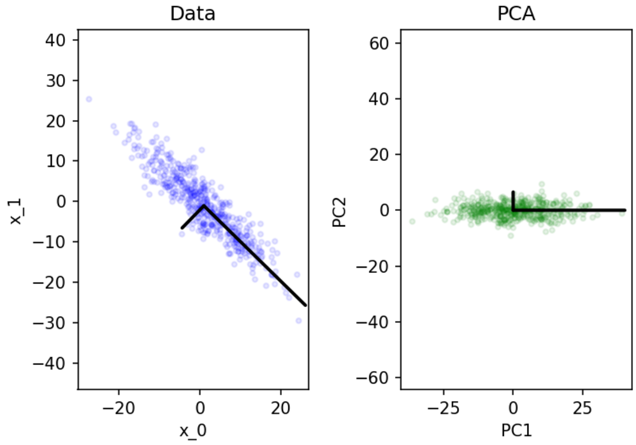


Figure 6.6: Visualisation of PCA for 2d data

After PCA has been applied, the feature vectors become abstract. It is no longer clear what features contribute the most information, because the features are transformed into new features. This transformation of the features is however applied to every sample, so the data can be classified using the abstract features.

6.3.2. Validation and testing

In order to get reliable and unbiased results out of the Machine Learning algorithms, the data was split into:

Set	Split
Training	70%
Testing	30%

The rule of thumb, according to renowned Machine Learning expert Andrew Ng, is to split your data into 4:1, thus giving 80% training data and 20% test data. However, this rule of thumb applies to Machine Learning in general, where datasets used are rather large (in the range of 1000 to 1000000 samples).
For this project, only data collected through experiments was available, which meant a dataset of about 200 samples. Therefore the choice was made to save a greater part (30%) of the data for the test set, in order to still achieve a robust performance measurement. The 70/30 split is still very commonly used when splitting data in solely test and training. A validation set was not needed, since cross-validation was utilised, see Figure 6.7.

6.3.3. Cross-validation and Hyperparameter tuning

5-fold Cross-Validation, as in Figure 6.7, was used in order to tune the hyperparameters used for the training of the algorithms. Hyperparameters are parameters that are used to control the learing process in machine learning. They can be considered settings to a machine learning model. Examples of hyperparameters entail the amount of decision nodes in a decision tree or the amount of groups in a K Nearest Neighbour algorithm The concept of cross-validation entails dividing the dataset in five data chunks, and training the data leaving one of those chunks out each time. Then, the average of these 5 iterations is taken to determine performance of the model with the chosen hyperparameters. This way, the concept of overfitting is nullified.

The method of tuning the hyperparameters was based on sweeping a parameter grid using scikit-learn its gridsearchCV function [34], [35]. The parameter grid was first constructed on logarithmic scale, which included values [0.1, 1, 10, 100, 1000]. The grid was then sweeped, meaning every value was chosen once as a hyperparameter to for example an SVM its regularization parameter, which is a parameter that describes the degree of importance that is given to misclassifications by the model.

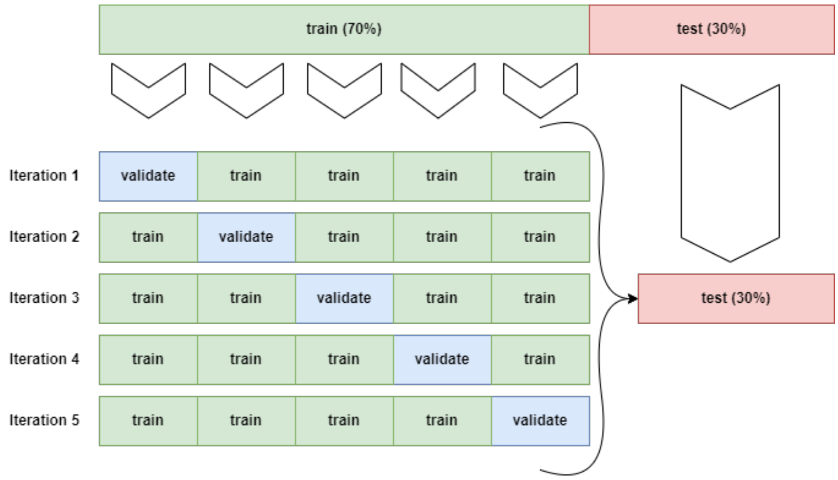


Figure 6.7: Visualisation of 5-fold Cross-Validation

From the result of this sweep, the parameters were further tuned. For the regularization parameter for example, the outcome of 0.1 was optimal. Then the grid sweep would contain more precise values, e.g. [0.01, 0.05, 0.1, 0.2, 0.4, 0.5]. For all numerical parameters this method was applied. As an example, the cross-validation method for the K Nearest Neighbours algorithm is shown in Figure 6.8.

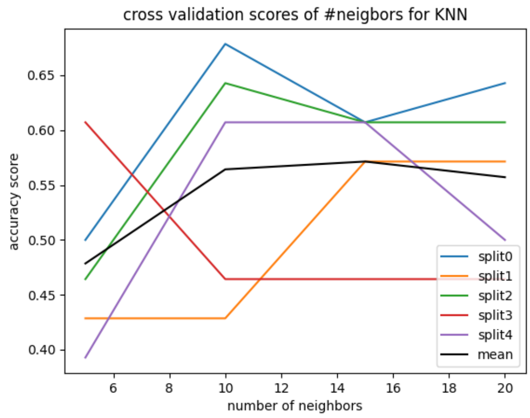


Figure 6.8: Visualisation of tuning parameter for KNN

As for the option parameters, such as the kernel used for SVM, all options were tried, and the one with best performance was then chosen. To illustrate the process of cross-validation to tune the kernel variabel for SVM, Figure 6.9.

The actual function gridsearch CV does this comparison on a higher level than these visualisations. Namely, it exploits every combination of all provided parameters' values of all tunable parameters to find the optimal one. This means that this function is actually operating in higher dimensional space than what is visualised in Figure 6.8 and Figure 6.9.

6.3.4. Comparisons drawn

After the hyperparameter tuning, the full training set (70%) of the data was then retrained on the best parameters found. The trained model that resulted from this was used to evaluate the performance of the algorithm on the test set.
A comparison can be drawn between the different Machine Learning algorithms using a confusion matrix, which is shown in Figure 6.10. The confusion matrix is a visualisation from which all possible statistical in-

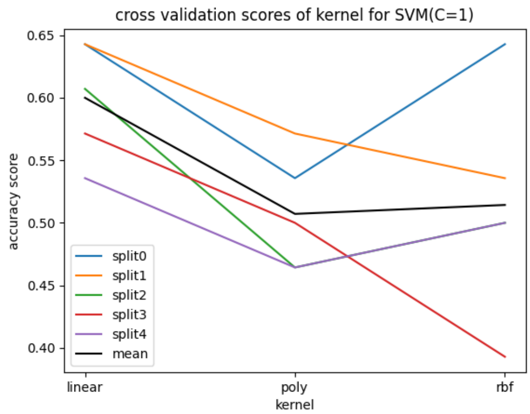


Figure 6.9: Visualisation of tuning parameter for SVM

algorithm	TPR	FAR	acc
Decision Tree (random state=44)	0.448	FAR = 0.552	acc = 0.517
KNN (n neighbors=4, weights='distance')	0.621	0.379	0.5
SVM (C=0.01, kernel='poly', random state=44)	1.0	0.0	0.483
Logistic Regression (C=3, random state=44)	0.345	0.655	0.517

Table 6.2: Performance metrics for different algorithms (on features of data) and PCA

formation can be inferred. This matrix contains information on the following statistics:

- **True positives:** An outcome where the model correctly predicts the positive class where the positive class is, in the case of the pseudowords experiment, a non-existing word.
- **False positives:** An outcome where the model incorrectly predicts the positive class.
- **True negatives:** An outcome where the model correctly predicts the negative class where the negative class is, in the case of the pseudowords experiment, an existing word.
- **False negatives:** An outcome where the model incorrectly predicts the negative class.

Based on the true positives, false positives, true negatives and false negatives, a comparison of algorithms is made through a comparison of their:

- True positive rate (TPR)
- False acceptance rate (FAR)
- Accuracy

These metrics were chosen due to their relevance in an authentication system. For such a system it is important that an unkown user is not mistakenly identified. However, it is less important that a known user is not recognised, due to their ability to simply try again to authenticate themselves.

In Table A.3, the Machine Learning was unsuccessful in finding any reliable results, as anywhere below 60% for binary classification means near randomness. All possibilities were exhausted, including thus preprocessed data directly into the machine learning (Table A.3), features as described in chapter 5, with normalization(Table A.2, and with PCA (Table 6.2). However, none proved to give a reliable performance. An elaboration on why it was unsuccessful/fruitless is given in subsection 6.1.2

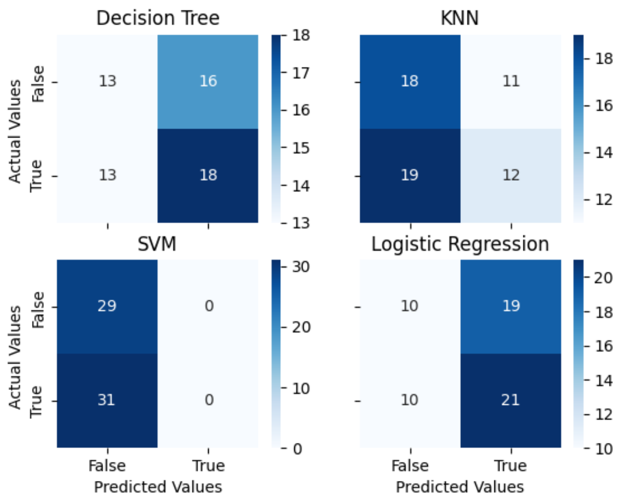


Figure 6.10: Performance on features of data and pca performed

Table 6.3: Accuracy's for pseudoword SVM (C=10, kernel='poly', random state=44) classification of different subsets from measured data and using different features. The wavelet statistics consisted of to many features to be useful for small data sets. For session M1a and M1b, the words were shown in order. For the other sessions words and pseudowords were mixed in the experiment

Session	10 Hertz Band Powers	EEG Bands Powers	Wavelet Statistics	Combined
M1a	64%	72%	-	72%
M1b	52%	49%	-	50%
M2a	38%	44%	-	44%
M2b	48%	48%	-	48%
M3a	48%	43%	-	42%
All	55%	56%	54%	56%

6.3.5. Classification Pseudowords using SVM

As stated earlier, based on literature it was concluded that the best machine learning algorithm for classification of EEG data is a Support Vector Machine. This algorithm needs to be trained and tested using data obtained over different measuring sessions and two different methods. In some cases the words were shown in order, which meant 20 existing words were shown first and 20 pseudowords were shown second. In the other situation, the words were shown in random order.

Each of the conducted experiments resulted in 40 data segments for classification. The initial grid search confirmed the fact that an SVM will outperform other algorithms (Figure 6.11, Table 6.5). Therefore, the SVM classification algorithm was pursued further.

Table 6.3 shows the accuracies of when the machine learning model was trained and tested with data from these different experiments, sorted by person and word order. The kernel of the SVM was linearly based which was chosen using the mentioned hyperparameter tuning technique. A linear kernel means that a linear plane was found which separated the data classes. The classifier was trained using the 70/30 data split mentioned. The SVM was trained and tested 200 times resulting in an average accuracy listed in Table 6.3 for the different experiment trials conducted.

Table 6.4: Amount of samples per experiment session shown in table 6.3.

session	M1a	M1b	M2a	M2b	M3a	total
samples	120	120	80	200	120	640

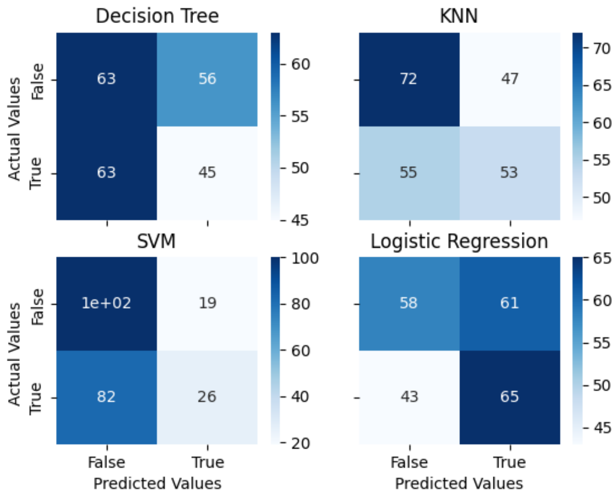


Figure 6.11: Performance on features of data and pca performed (pseudowords)

algorithm	TPR	FAR	acc
Decision Tree (min samples leaf=4, random state=44)	0.529	FAR = 0.471	acc = 0.476
KNN (n neighbors=3, weights='distance')	0.605	0.395	0.551
SVM (C=10, kernel='poly', random state=44)	0.840	0.160	0.555
Logistic Regression (C=5, random state=44)	0.487	0.513	0.542

Table 6.5: Performance metrics for different algorithms (on features of data) and PCA for pseudowords

7
Discussion of Results

At the end of this project, two potential techniques which could be used to implement an authentication system were tested. These techniques consisted of frequency tagging and pseudowords. Many papers have been written about these techniques, but most were reliant on wet electrodes. Therefore, experiments were conducted in an attempt to research their feasibility in an authentication system reliant on dry-electrodes.

7.1. Experiment Results: Frequency Tagging

The first step in investigating frequency tagging as a technique to implement for an authentication system was showing that words could be tagged using a particular frequency. This was definitely possible for dry-electrode recording devices. Regarding the method of stimulation, both a strong visual stimulus using a flashlight, shown in Figure 6.1, as well as a weaker one using a computer screen, shown in Figure 6.2, can be detected. For the stimulation duration, Figure 6.3 shows that the stimulation duration does not need to be long. Starting from approximately 15 seconds the peak is about twice as high as the average frequency power, meaning that the tagged frequency can easily be detected.

Regarding the reinduction of increased frequency power for tagged words, the results were less concise. Figure 6.4 shows the average relative frequency power at tagged frequencies for different waiting intervals between the tagging and reappearing of words. The results indicate that after a short waiting interval, no increase in power can be detected. For longer waiting intervals, there does seem to appear an increase in power at the tagged frequency, which would indicate that reinduction is possible, but it is difficult to be certain and rule out chance in this case. As mentioned, more measurements would be required. Both to get a better average in order to rule out the possibility of chance, as well as to gather more information on how long a word can stay tagged in case reinduction is possible. Based on the current inability to show that tagged words show up in the spectrum, further research was not performed in order to dedicate more time to an alternative method.

7.2. Experiment Results: Pseudowords

In order to test whether it is possible to tell the difference between existing words and pseudowords using an EEG recording, measurements were done which were then analysed using both manual analysis as well as using machine learning.
For the manual analysis, the difference in frequency power was calculated for multiple frequency bands of 10 Hz in width. These calculations show that for some recordings, the expected difference in the 30 Hz band is indeed noticeable. However, this is not the case for every experiment trial. It was also expected that this increase was noticeable across the entire hemisphere but this did not show up in the analysis. Additionaly, sometimes there was an increase in power. These results seem to suggest that the measurements do not reflect the power differences expected from literature. This could mean that dry-electrode recording devices are not able to pick up on these differences. It is more likely however that due to intra- and interexperimental differences, which means differences between measurement subjects as well as differences between recording sessions, the data cannot be combined as well as initially expected. This means conclusions are drawn based on sample sizes which are too small, allowing for noise and randomness to heavily influence the results.

Using machine learning, 1 second (1 word) samples were used. Still, the small data size influences the results even more severely. It can be seen that the performance of the machine learning was above statistically random guessing. However, as stated in Table 6.3 the combination of these different measurements is unsuccessful.

7.3. Comparison between methods

Accuracy
Pseudowords can be more accurately classified (Table 6.3, Table 6.2). On accuracy, the pseudo-word method is therefore more promising at this time.

Durations
The SSVEP that is caused by the tagging, is very visible starting from 10 seconds onwards. The reactivation of this SSVEP by seeing the tagged word again is visible only within a second after stimulus onset, so if this were measurable it would be swift by definition. Because it is not measurable, this makes it hard to say something about the practical duration of this way of authentication.
For the classification of pseudowords, one second data segments were used. Although the accuracy was quite low, this means that quick authentication is possible. Just as with the frequency tagging method, the effect only lasts less than a second, so extending the duration of the data segments would not help improve accuracy.
If both methods would work in the given experimental setup, they would differ very little in duration. Nevertheless, for the experiments described in this thesis the data segments used for pseudowords are shorter.

Complexities
Because the detection of frequency tagging relies on the measuring of a power increase around a single frequency, this could be detected with a fairly simple algorithm. It could even by implemented as a circuit using filter, to obtain the power in this frequency.
For pseudowords, a more elaborate algorithm is used. To obtain good results around 30 features were needed to train and test the SVM that was used. This still resulted in a runtime of far less than a second after the algorithm was trained. Both methods have algorithms that are therefore simple enough, but the one used for frequency tagging it could be far less complicated.

Overall
On two out of the three criteria listed above, pseudowords is the better option. When taking into consideration the significance of each criterium, pseudowords currently is the most promising option of the two.

8 Conclusion and Future Work

8.1. Conclusion

During the initial stage of this project, evidence was found that authentication using EEG signals is possible. Based on papers, several methods were discovered that could have the potential to function as a novel way of authentication, each with its own advantages and disadvantages. Based on what was read, the conclusion was drawn that most of these methods could influence EEG signals in such a way that they can be detected using conventional analysis or machine learning. This resulted in the plan to test as soon and as much as possible to investigate all of these methods.

Unfortunately, this was much more complex than expected. The initial measurement had a large amount of noise and it was not straightforward at all to detect the induced changes. It turned out that the experiments needed to be very tightly constrained as opposed to the initial approach. The focus on creating a complete authentication system was therefore shifted towards trying to find a way to reliably influence one's brain in a way that could be detected. This would then serve as a proof of concept that an authentication system could be possible.

The final focus was on the method of frequency tagging as well as the difference between existing words and pseudowords. Both methods were tested in stages. For frequency tagging, the actual tagging of words could be shown in the recordings but re-induction of the effect was not possible, at least based on current measurements. Regarding existing word and pseudoword differences, both an analysis based on literature as well as machine learning techniques were explored. The technique described in the literature, which relied on determining frequency power differences in the left hemisphere of the brain, was unable to make a clear distinction. Machine learning techniques were able to classify the words better than random guesses, but this difference was barely significant. The data interval length of one second simply proved to be too short. In the end, based on the current analysis of the data, both techniques would not suffice as a method to base an authentication system on.

8.2. Requirement evaluation

Of the goals listed in the programme of requirements found in chapter 2, six were met. Below a short conclusion for each requirement can be found.

- **2.1.i:** Artifacts are detected and discarded, and filters are used to remove noise.
- **2.1.ii:** The algorithm was able to detect changes in EEG data for longer duration SSVEP stimulation in the case of frequency tagging.
- **2.1.iii:** Identification based on frequency tagging or pseudowords was not possible.
- **2.1.iv:** Only EEG data was used as an input to the algorithm.
- **2.1.v:** The set-up only required a headset with the algorithm. Experiments were run on a computer and only required software to function.

- **2.2.i:** Five seconds of time intervals proved too short for the detection of induced EEG signals. In the case of the frequency tagging stimulus, which was best detectable, a minimum of approximately 15 seconds was required.
- **2.2.ii:** The algorithm ran in less than three seconds. The preprocessing of the signal could be done within 1 second. Classifying data happened almost instantly once the model was trained.
- **2.2.iii:** A True Positive rate of 80% was reached, though the accuracy of the model was low. This is shown in Table 6.5
- **2.3.i:** A False Positive rate of 1% was not reached. For the final model, the FAR was 16%. This is shown in Table 6.5

8.3. Future Work

Three of the five methods that were possible according to the literature study are still not investigated, and there are probably many more ways authentication can be based on EEG measurements. This would be very interesting to find out, but this section will mainly focus on how the two proposed solutions that we have extensively tested can be better researched.

Frequency tagging
Frequency tagging should be possible to see in much shorter time frames than was achieved by us. This is partially due to imperfections in our filtering, features, and classification, but also due to the large amounts of noise present in the signal. It would therefore be interesting to test this method with wet electrodes in lab settings. This is even more unlike the final application (in-ear dry-electrode in real life) than our way of testing (scalp dry-electrode in semi-lab setting), but would serve as a proof of concept to further extend on. When the frequency tagging is proven possible in this setting, which it should according to literature, testing how long the tagging effect lasts is an important next step to find out if it would be really possible to set up a password this way. Repeating experiments more would also have been useful because with more data, more complex forms of machine learning like neural networks could be used. Finally, it would be very interesting to see if this classification can be done using filters or a system based on a phase-locking loop, so that it can be easily implemented on the recording device.

Pseudowords
Experiments involving pseudowords were carried out quite late in the project, after frequency tagging was completely ruled out. This meant less time could be spent on the right features, which were more opaque to find than for frequency tagging. It was clear these features would occur within certain frequency bands, but the exact effect was based on time and location of the recording. It was also not possible due to time limitations to test if a pseudoword could be given meaning in order to use it as a password for people. Just like with the frequency tagging experiments, a more elaborate measurement setup would have helped in order to create a proof of concept for applications using in-ear EEG recorders.
Although the average accuracy of classification across all our data was just slightly above chance level, one session with 120 samples achieved a much higher accuracy. It would be interesting to find out what caused this data set to be more useful. This cause could then be replicated or even further improved on to gain higher accuracy overall.

Bibliography

[1] J. A. Urigüen and B. Garcia-Zapirain, “Eeg artifact removal—state-of-the-art and guidelines,” Journal of neural engineering, vol. 12, no. 3, p. 031001, 2015.

[2] S. J. M. Smith, “Eeg in the diagnosis, classification, and management of patients with epilepsy,” Journal of Neurology, Neurosurgery & Psychiatry, vol. 76, no. suppl 2, pp. ii2–ii7, 2005.

[3] S. L. Kappel, M. L. Rank, H. O. Toft, M. Andersen, and P. Kidmose, “Dry-contact electrode ear-eeg,” IEEE Transactions on Biomedical Engineering, vol. 66, no. 1, pp. 150–158, 2019.

[4] X. Huang, Y. Xiang, A. Chonka, J. Zhou, and R. H. Deng, “A generic framework for three-factor authentication: Preserving security and privacy in distributed systems,” IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 8, pp. 1390–1397, 2011.

[5] S. Subangan and V. Senthooran, “Secure authentication mechanism for resistance to password attacks,” in 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), vol. 250, pp. 1–7, 2019.

[6] N. Merrill, M. T. Curran, S. Gandhi, and J. Chuang, “One-step, three-factor passthought authentication with custom-fit, in-ear eeg,” Frontiers in Neuroscience, vol. 13, 2019.

[7] T. Nakamura, V. Goverdovsky, and D. P. Mandic, “In-ear eeg biometrics for feasible and readily collectable real-world person authentication,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 3, pp. 648–661, 2018.

[8] “Ultracortex "mark iv" eeg headset.”

[9] N. Bajaj, “Wavelets for eeg analysis,” in Wavelet Theory (S. Mohammady, ed.), ch. 5, Rijeka: IntechOpen, 2021.

[10] M. Abo-Zahhad, S. Ahmed, and S. N. Seha, “A new eeg acquisition protocol for biometric identification using eye blinking signals,” International Journal of Intelligent Systems and Applications (IJISA), vol. 07, pp. 48–54, 05 2015.

[11] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, “Chapter 2 - technological basics of eeg recording and operation of apparatus,” in Introduction to EEG- and Speech-Based Emotion Recognition (P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, eds.), pp. 19–50, Academic Press, 2016.

[12] A. G. Reddy and S. Narava, “Article: Artifact removal from eeg signals,” International Journal of Computer Applications, vol. 77, pp. 17–19, September 2013. Full text available.

[13] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “EMG contamination of EEG: spectral and topographical characteristics,” Clin Neurophysiol, vol. 114, pp. 1580–1593, Sept. 2003.

[14] K.-A. Kwon, R. J. Shipley, M. Edirisinghe, D. G. Ezra, G. Rose, S. M. Best, and R. E. Cameron, “High-speed camera characterization of voluntary eye blinking kinematics,” Journal of the Royal Society, Interface, vol. 10, pp. 20130227–20130227, Jun 2013. 23760297[pmid].

[15] A. Lewis, H. Schriefers, M. Bastiaansen, and J.-M. Schoffelen, “Assessing the utility of frequency tagging for tracking memory-based reactivation of word representations,” Scientific Reports, vol. 8, no. 7897, pp. 1–12, 2018.

[16] M. Wimber, A. Maaß, T. Staudigl, A. Richardson-Klavehn, and S. Hanslmayr, “Rapid memory reactivation revealed by oscillatory entrainment,” Current Biology, vol. 22, no. 16, pp. 1482–1486, 2012.

[17] W. Lutzenberger, F. Pulvermüller, and N. Birbaumer, “Words and pseudowords elicit distinct patterns of 30-hz eeg responses in humans,” Neuroscience Letters, vol. 176, no. 1, pp. 115–118, 1994.

[18] A. Zhigal, M. Yoneya, S. Zhao, and M. Chait, "Probing cortical excitability using rapid frequency tagging," NeuroImage, vol. 195, pp. 59–66, Oct 2019.

[19] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Increase information transfer rates in bci by csp extension to multi-class," in Advances in Neural Information Processing Systems (S. Thrun, L. Saul, and B. Schölkopf, eds.), vol. 16, MIT Press, 2003.

[20] J. Kronegg, G. Chanel, S. Voloshynovskiy, and T. Pun, "Eeg-based synchronized brain-computer interfaces: A model for optimizing the number of mental tasks," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 15, no. 1, pp. 50–58, 2007.

[21] S. Valenzi, T. Islam, P. Jurica, and A. Cichocki, "Individual classification of emotions using EEG," Journal of Biomedical Science and Engineering, vol. 07No.08, p. 17, 2014.

[22] E. Siedlecka and T. F. Denson, "Experimental methods for inducing basic emotions: A qualitative review," Emotion Review, vol. 11, no. 1, pp. 87–97, 2019.

[23] M. Jukiewicz and A. Cysewska-Sobusiak, "Stimuli design for ssvep-based brain computer-interface," International Journal of Electronics and Telecommunications, vol. 62, no. 22, p. 109–113, 2016.

[24] H. U. Amin, A. S. Malik, N. Badruddin, and W.-T. Chooi, "Eeg mean power and complexity analysis during complex mental task," in 2013 ICME International Conference on Complex Medical Engineering, pp. 648–651, 2013.

[25] G. Chen, "Automatic eeg seizure detection using dual-tree complex wavelet-fourier features," Expert Systems with Applications: An International Journal, vol. 41, pp. 2391–2394, 04 2014.

[26] A. Saday and A. Ozkan, "Classification of epileptic eeg signals using dwt-based feature extraction and machine learning methods," International Journal of Applied Mathematics Electronics and Computers, vol. 9, no. 4, pp. 122 – 129, 2021.

[27] A. al Qerem, F. Kharbat, S. Nashwan, S. Ashraf, and khairi blaou, "General model for best feature extraction of eeg using discrete wavelet transform wavelet family and differential evolution," International Journal of Distributed Sensor Networks, vol. 16, no. 3, p. 1550147720911009, 2020.

[28] S. Satapathy, A. Jagadev, and S. Dehuri, "Weighted majority voting based ensemble of classifiers using different machine learning techniques for classification of eeg signal to detect epileptic seizure," Informatica, vol. 41, pp. 99–110, 03 2017.

[29] P. Kidmose, D. Looney, M. Ungstrup, M. L. Rank, and D. P. Mandic, "A study of evoked potentials from ear-eeg," IEEE Transactions on Biomedical Engineering, vol. 60, no. 10, pp. 2824–2830, 2013.

[30] F.-B. Vialatte, M. Maurice, J. Dauwels, and A. Cichocki, "Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives," Progress in Neurobiology, vol. 90, no. 4, pp. 418–438, 2010.

[31] G. Acampora, P. Trinchese, and A. Vitiello, "Classifying eeg signals in single-channel ssvep-based bcis through support vector machine," in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2305–2310, 2020.

[32] L. Guo, Y. Wu, L. Zhao, T. Cao, W. Yan, and X. Shen, "Classification of mental task from eeg signals using immune feature weighted support vector machines," IEEE Transactions on Magnetics, vol. 47, no. 5, pp. 866–869, 2011.

[33] M. T. Sadiq, X. Yu, Z. Yuan, Z. Fan, A. U. Rehman, G. Li, and G. Xiao, "Motor imagery eeg signals classification based on mode amplitude and frequency components using empirical wavelet transform," IEEE Access, vol. 7, pp. 127678–127692, 2019.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[35] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122, 2013.

A

Appendix

A.1. Code

The code for all experiments, all analysis and all machine learning can be found on the GitHub repository in the Authentication branch: https://github.com/Achtuur/BCIBAP

A.2. Data Collection

A list of the words used in the pseudoword experiment can be found in table A.1.

A.3. Data Analysis

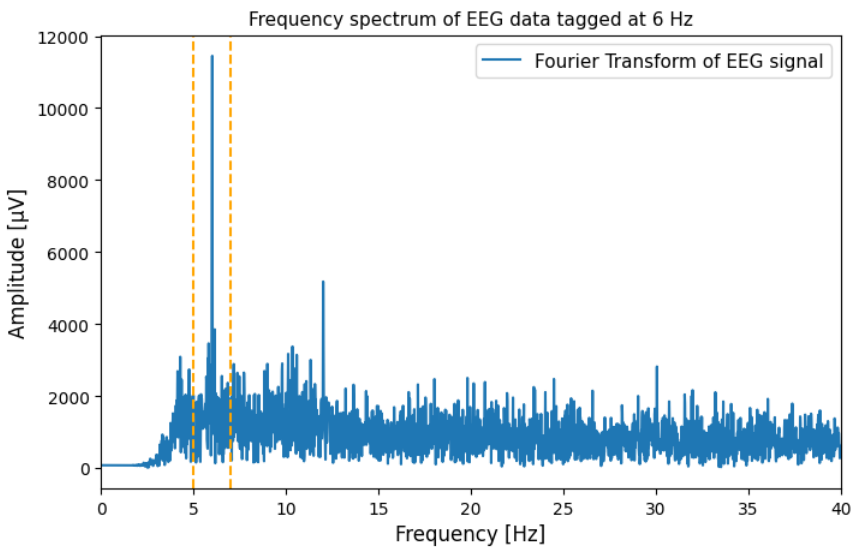


Figure A.1: Frequency spectrum of a subject's EEG data stimulated using a flickering flashlight at 6 Hz.

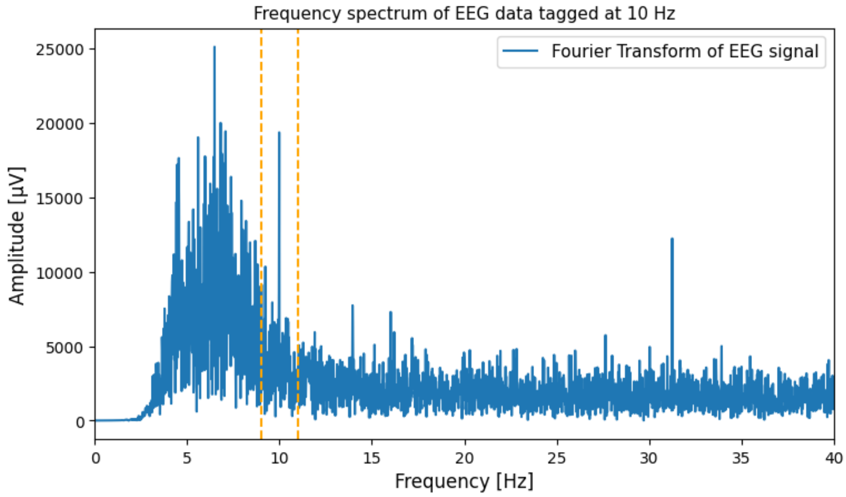


Figure A.2: Frequency spectrum of a subject's EEG data stimulated using a screen at 10 Hz.

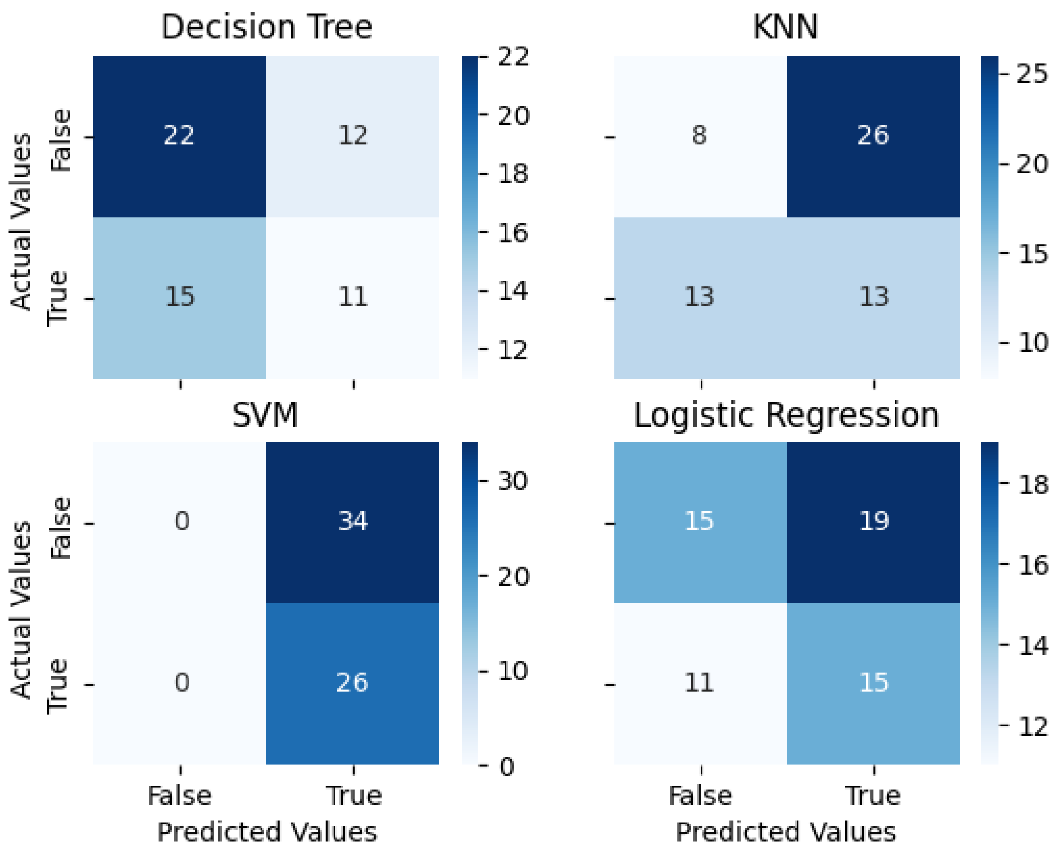


Figure A.3: Performance on features of data normalized

algorithm	TPR	FAR	acc
Decision Tree (min samples leaf=3, random state=44)	0.647	FAR = 0.353	acc = 0.55
KNN (n neighbors=15, weights='distance')	0.235	0.735	0.35
SVM (C=0.1, kernel='poly', random state=44)	0.0	1.0	0.433
Logistic Regression (C=3, random state=44)	0.441	0.559	0.5

Table A.2: Performance metrics for different algorithms (on features of data) normalized

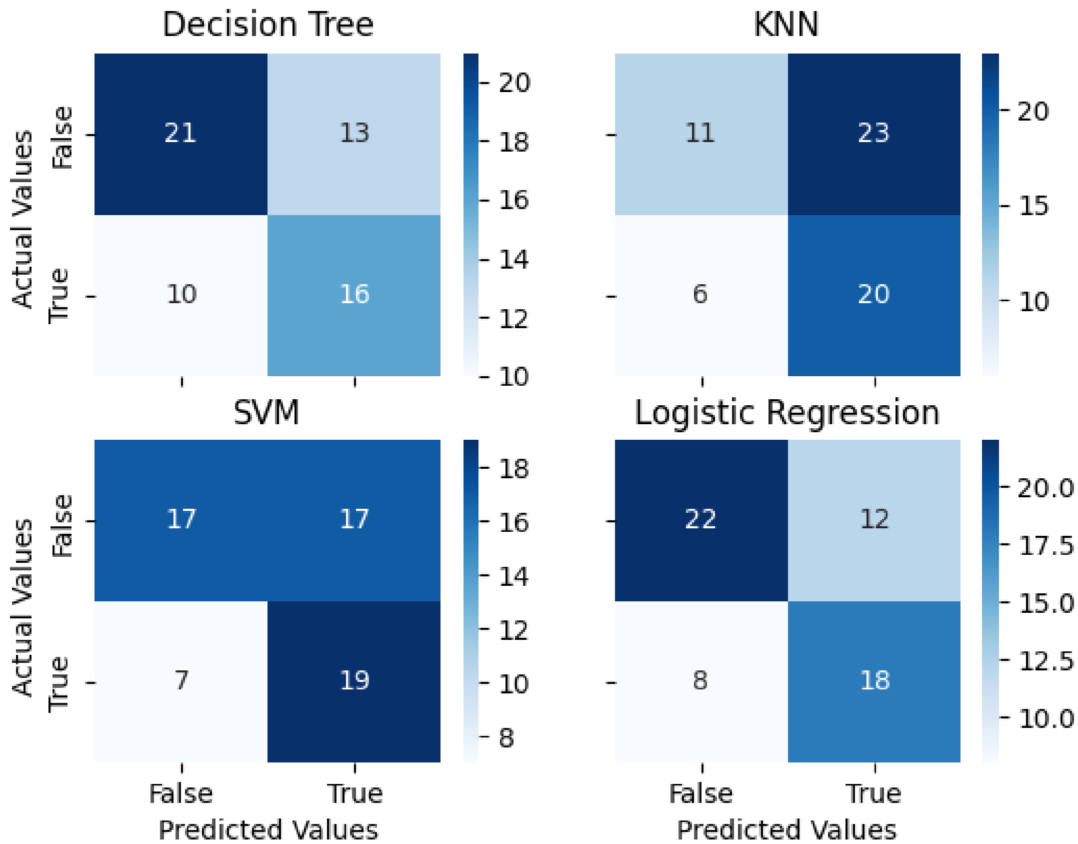


Figure A.4: Performance of respective tuned ML algorithms on preprocessed data without feature extraction

algorithm	TPR	FAR	acc
Decision Tree	0.5	0.5	0.567
KNN (n neighbors=15, weights='distance')	0.265	0.735	0.433
SVM (C=0.01, kernel='linear', random state=44)	0.5	0.5	0.6
Logistic Regression (C=5, random state=44)	0.647	0.353	0.666

Table A.3: Performance metrics for different algorithms (on raw data)

Table A.1: Dutch Words and pseudowords used for experiments

words	pseudowords
achteruit	aflag
belanghebbenden	beden
beschouwing	bekachtig
bestuderen	belaren
bewustzijn	bioleren
bolleboos	bobde
borst	carade
complicaties	compel
daarover	diten
daling	dren
dassen	drume
dekenkist	elig
dijken	fameur
draait	far
excuus	formen
gevoelig	gedig
gezinsleven	grafeit
grafische	haseer
handhaven	hemen
hein	jatie
hij	kajade
incidenten	keuk
indelen	klag
inkoop	klon
interviewer	leigen
inventief	mantie
inwoners	modelt
jullie	neerde
leuks	neren
literatuur	oerd
major	ombleef
medewerkers	ond
misverstand	onmen
muis	onost
onschadelijk	ontst
pakketten	opgespel
pakt	opvoeg
paleis	orteerd
reageert	parf
rollen	proten
roze	razzelen
schaden	samer
sluipen	slukken
standbeeld	spioen
stijf	stanker
stokt	studen
stop	submen
synthesen	terscheid
telegraaf	tortaal
veelal	tramen
vochten	vaarde
weergeven	verlon
wei	vreed
zichtbare	zweldel
ziek	
zoals	
zuipen	