A Comparative Study of Physiological Feature Selection Methods for Emotion Recognition

Andreas De Lille

Supervisors: Prof. dr. ir. Joni Dambre, Dr. ir. Pieter van Mierlo
Counsellor: Ir. Thibault Verhoeven

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Computer Science Engineering

Department of Electronics and Information Systems
Chair: Prof. dr. ir. Rik Van de Walle
Faculty of Engineering and Architecture
Academic year 2015-2016

UNIVERSITEIT GENT

Contents

Nomenclature

ANOVA Analysis of Variance

BCI Brain Computer Interface

CT Computed Tomography

DASM Differential Asymmetry

DCAU Differential Caudality

DE Differential Entropy

DEAP Dataset for Emotion Analysis using Physiological Signals

EEG Electroencephalography

ELM Extreme Learning Machine

ERP Event-Related Potential

fMRI Functional Magnetic Resonance

GELM Graph regularized Extreme Learning Machine

GSR Galvanic Skin Response

KNN k-nearest neightbors

LDA Linear Disciminant Analysis

LDS Linear Dynamic System

LR Logistic Regression

MEG magnetoencephalography

MLP Multilayer Perceptron

MRMR Minimal Redundancy Maximal Relevance

NIRS Near Infrared Spectroscopy

OCR Optical Character Recognition

OOB Out of Bag

PCA Principal Component Analysis

Positron Emission Tomography nomeqref 1.0

PSD Power Spectral Density

RASM Rational Asymmetry

RCAU Rational Caudality

RF Random Forests

RSP Respiration Belt

SAM self-assessment manikins

SBS Sequential Backward Selection

SC Skin conductivity

SFS Sequential Forward Selection

SNR Sound to Noise Ratio

1

Introduction

This chapters introduces the masterthesis. It starts by introducing the basic concepts of emotion recognition based on physiological signals, machine learning. Then it explains the problem statement and the goal of the thesis, followed by an overview of the next chapters.

1.1 Emotion recognition

Human-to-machine communication, where humans communicate with machines or computer agents, is becoming more and more common[1]. Fully understanding human communication is a complex problem. In addition to verbal communication, non-verbal communication is also used to exchange information[2]. To better understand human-to-machine communication, more insight in the non-verbal communication is needed. Emotion recognition is becoming an increasingly important field as a result[3].

Emotion recognition is the proces of recognizing a subject's emotional state. In psychology a clear distinction between physiological behavior and the conscious experience of an emotion, called expression[4] is made. Expression consists of many parts, including facial expression, body language and voice concern[5]. Unlike expression, the physiological aspect of an emotion, e.g. heart rate, skin conductance and pupil dilation, is much harder to control. This makes emotion recognition based on physiological signals more robust to social masking[6]. Social masking is the process where an individual masks or hides their emotions to conform to social pressure. To really know one's emotions, it seems, one has to research the physiological aspect of the emotion.

Before emotions can be recognized, an objective class model describing different emotions is needed. A simple way of achieving this is using several discrete emotions, e.g. anger, joy, sad and pleasure. A more convienent model to classify emotions is the bipolar arousal-valence model[4, 3], which places emotions in a two dimensional space. The main advantage of using a continuous multidimensional model, is that all emotions are modelled in its space, even when no particular discrete label can be used to define the current feeling. Figure 1.1 shows the mapping of different emotions for this model.

The valence-arousal model consists of two dimensions. Arousal indicates how active a person is and ranges from inactive/bored to active/excited. The valence indicates if the emotion is perceived as positive or negative. Even though arousal and valence describe emotions quite well,

a third dimension, dominance, can also be added. This third dimension indicates how strong the emotional feeling was and ranges from a weak feeling to an empowered, overwhelming feeling. The dominance component can aid to filter out samples of strong feelings, since feelings with low dominance are less likely to show significant effects.

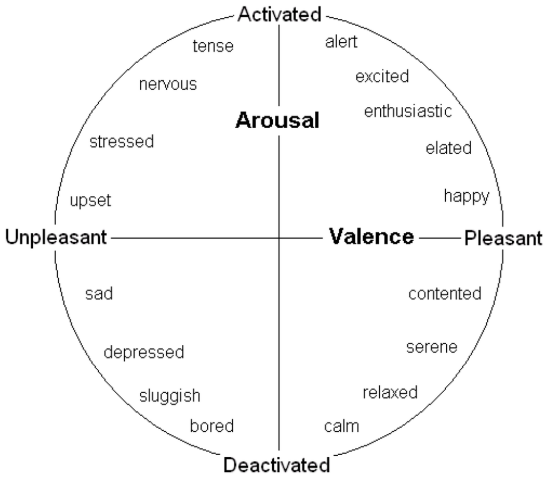

Figure 1.1: The arousal - valence model maps emotions in a two dimensional plane[7]

1.1.1 Physiological signals

Now that the classification model is defined, the different physiological signals will be explained. As mentioned before, these signals are used to do automatic emotion recognition. Physiological signals can be divided in two subgroups: brain activity and other signals, e.g. heart rate, respiration rate, etc. Different technologies exist to record brain activity. The most convenient method is electroencephalography (EEG), since it is a non-invasive method. Non-invasive methods, in contrast to invasive methods require no surgery. In case of EEG, they simply measure electrical activity using electrodes placed on the scalp.

Electrical activity in the brain is generated when an incoming signal arrives in a neuron. This triggers some sodium ions to move inside the cell, which in turn, causes a voltage rise[4]. When this increase in voltage reaches a threshold, an action potential is triggered in the form of a wave of electrical discharge that travels to neighbouring neurons. When this reaction occurs simultaneously in a lot of neurons, the change in electrical potential becomes significant enough, it is measured by the EEG surface electrodes. EEG can thus only capture synchronized activity of many, many neurons[4]. This explains why EEG signals have low spatial resolution capabilities. EEG measurements consist of electrical potentials of different channels, measured over time, like shown in Figure 1.2.

Figure 1.2: EEG measurements is a trace electrical potentials of different channels over time.[8]

Signals originating from the cortex, close to the skull, are easier to measure, while signals originating deeper in the brain cannot be observed directly. Even for signals originating close to the cortex, EEG is far from precise as the bone between the the cortex and electrodes distorts the signal. Additionally, other artifacts like eye and muscle movement add a lot of noise to the signal. This explains why EEG signals are very noisy by nature. Noise removal techniques are therefor advised[9]. Note that even though EEG data contains a lot of noise and has a low spatial resolution, it still provides significant insight into the electrical activity of the cortex while offering excellent temporal resolution[10].

To ensure that experiments are replicable, standards for locations of electrodes have been developed. One of these systems is the 10/20 system, an internationally recognized method to describe the location of scalp electrodes[11]. The numbers 10 and 20 refer to the distances between the electrodes, which are either 10% or 20% of the total front-back or left-right distance of the scalp, this is depicted in Figure 1.3.

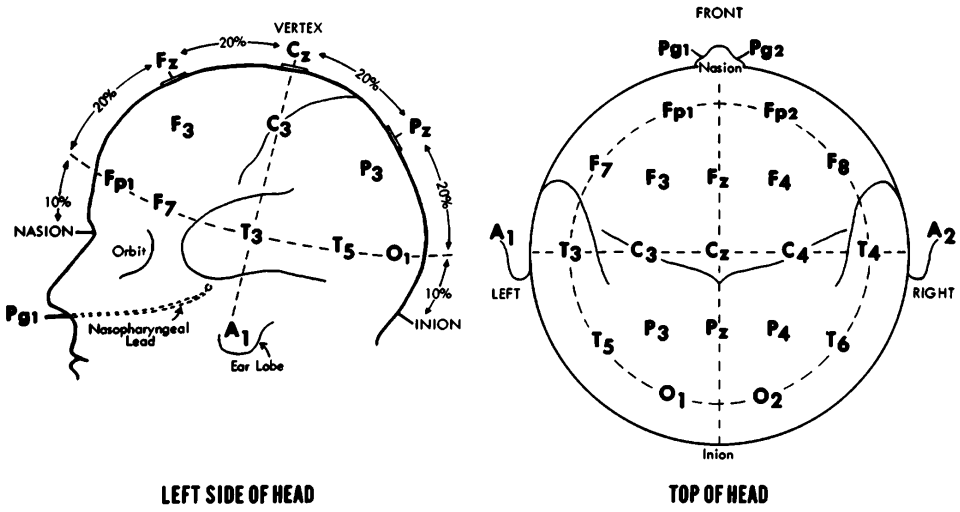

Figure 1.3: The electrode placement of a 23 channel system[12].

Each site is identified with a letter that determines the lobe and a number that determines the hemisphere location.

- **F:** Frontal
- **T:** Temporal
- **C:** Central
- **P:** Parietal
- **O:** Occipital

Note that no central lobe exists; the C letter is only used for identification purposes. The letter z indicates that the electrode is placed on the central line. Even numbers are used for the right hemisphere, while odd numbers are used for the left hemisphere. Note that the 10/20 system does not require a fixed number of channels. Some experiments may use a different set of channels, but they all follow the same naming convention. In this work, a 32 channel EEG cap is used. The corresponding electrode locations are shown in Figure 1.4

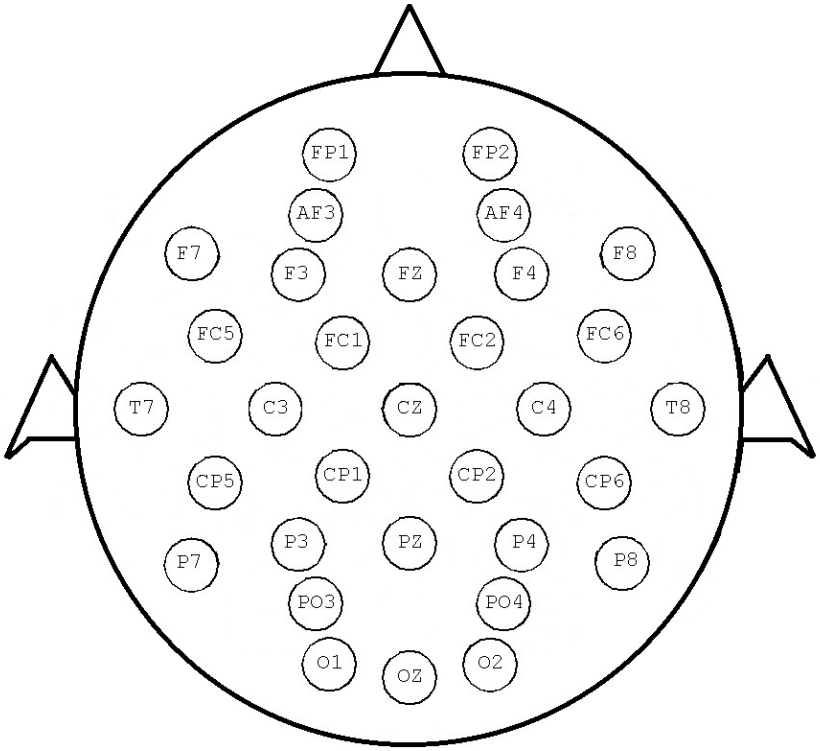

Figure 1.4: Placement of the 32 electrodes in this thesis.

In the frequency domain, brain waves are usually split up into different bands[2, 13], with a different medical interpretation for each band. These wavebands are:

1. **Alpha:** 8-13Hz, indicate how relaxed and/or inactive the brain is.
2. **Beta:** 13-30HZ, indicate a more active and focused state of mind.
3. **Gamma:** 30-50Hz, relate to simultaneous processing of information from different brain areas.
4. **Delta:** 0-4hz, these waves are generated during dreamless sleep and meditation.
5. **Theta:** 4-8Hz, occurs during dreaming.

Even though EEG is used in this thesis, alternative methods to measure brain activity exist. What follows is an overview of some of these techniques.

- Magnetoencephalography (MEG) use magnetic fields to measure brain activity[14]. Since MEG is more prone to noise from external magnetic signals, i.e. the earth's magnetic field and electromagnetic communication, a magnetic shielded room is required, making this method very expensive and not mobile.
- Functional magnetic resonance (fMRI) [15]: works by detecting changes in blood oxygenation and blood flow. An active area of the brain consumes more oxygen and has an increased blood flow.

- Computed tomography (CT) [16]: uses X-rays to create an image of the brain.
- Positron emission tomography (PET) [17]: this methods uses trace amounts of short-lived radioactive material. When this material undergoes decay, a positron is emitted that is picked up by a detector.
- Near infrared spectroscopy (NIRS) [18]: an optical technique to measure blood oxygenation in the brain. This technique works by shining light in the near infrared part of the spectrum through the skull and measuring how much remerging light is attenuated.

In addition to brain activity, other physiological signals are also used in this work. The most known signal is the heart rate, which measures the number of contractions per minute. Respiration rate gives to number of breaths a human takes in one minute[19]. Another physiological signal is the galvanic skin response. The galvanic skin response measures the electrical characteristics of the skin[20, 19]. In addition to the electrical characteristics, the temperature of the skin can also be measured. A plethysmograph is another physiological signal, that measures changes in volume within an organ[19]. A plethysmograph can be used to measure a subject's blood pressure and heart rate.

1.2 Machine learning

Machine learning is the missing link between the physiological signals and the emotion recognition. It is, in short, an input output model, that takes physiological signals and maps them to an emotional state. Machine learning is a very broad domain. As a result, this discussion will be limited to an introduction of the basic machine learning concepts with the focus on the application of machine learning and machine learning techniques used in this thesis.

A possible definition for machine learning is: "the science of getting computers to act without being explicitly programmed"[21]. To do so, machine learning uses pattern recognition to find patterns or structure in the data. A simple example of machine learning is the Optical Character Recognition (OCR), where a computer recognises characters in pictures[22]. An example of OCR is shown in Figure 1.5.

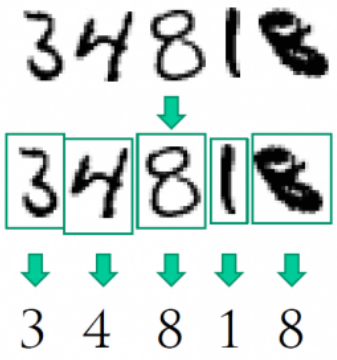

Figure 1.5: In optical character recognition, a computer uses machine learning to find characters in an image[23].

To further explain how machine learning works, have a look at the following example. Suppose one has a price list of houses that are for sale combined with their total area, shown in Table 1.1. Logic sense dictates us that a bigger house will have a higher price than a smaller house. The total area is a characteristic of the house that helps us in determining the price. In the context of machine learning, the characteristic 'total area', will be called a feature as the asking price of a house is correlated to the total area.

Area of the house (m^2)	**Price (x 1000 EUR)**
70	312
73	429
76	174
79	410
82	334
⋮	⋮

Table 1.1: total area of different houses and their corresponding asking prices.

One possible way of predicting the asking price of a house is machine learning. Machine learning works in several steps, first you train the machine learning algorithm with a list of asking prices and the corresponding area of the house. This process is called training or fitting and gives the machine learning component an idea to what price corresponds to a house with a certain area. Once trained, the algorithm's output will look like Figure 1.6. The black dots represent the data points from Table1.1. The blue line represents the predicted price for different area. The predicted price is simply defined by the total area of the house multiplied by some weight.

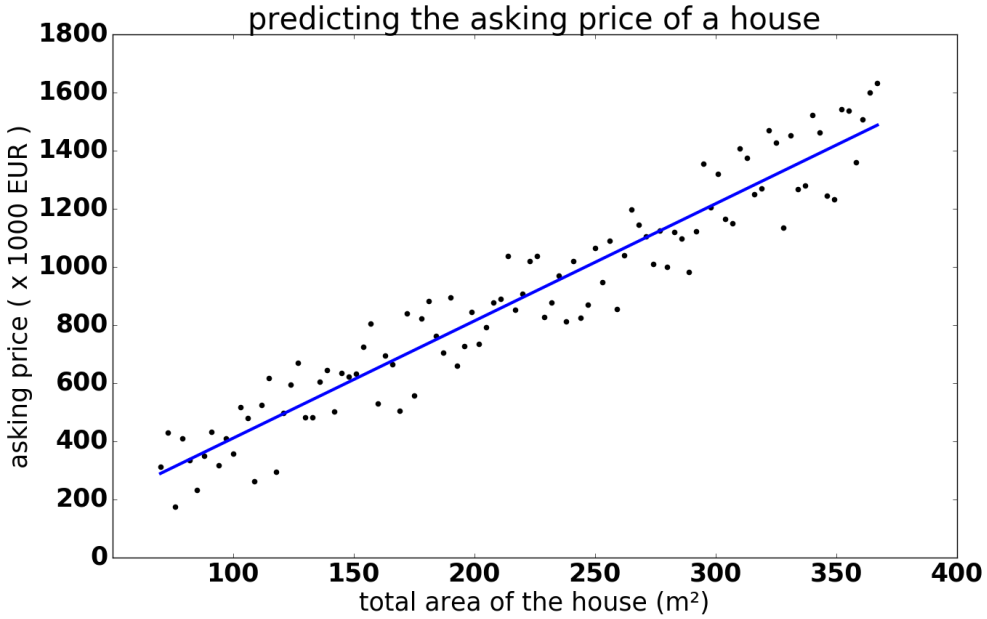

Figure 1.6: The price of a house is determined by its total area.

Even though, the blue line looks reasonable, there is sometimes a big difference between the predicted value and the actual value. This is due to the fact that the area of the house is

only one feature that determines the price. Other features, like the number of bedrooms or the location of the house, were not taken into consideration. Adding additional features, gives more insight into the data, e.g. a house with 5 bedrooms is more expensive than a house with only 3 bedrooms. Having more features is thus likely to improve the performance of the machine learning algorithm.

There are many machine learning algorithms. One way to group these algorithms is to look at the produced output. In the asking price examples above, the output is a price, which is (more or less) a continuous value. Machine learning problems that require the output of a continuous value, are called regression problems[24]. In the OCR example above, a picture of a character is classified as a character. This means that OCR is a classification problem, as there are only a limited number of characters in an alphabet[24].

Another way to group algorithms is based on their training data[24]. In the asking price examples above, the training data consists of labelled results. Labelled training data corresponds to data where the correct output (in this case the asking price) is given for each input (the area). This type of machine learning is referred to as supervised machine learning[24]. The alternative is unsupervised machine learning[24]. Unsupervised learning often results in finding groups of similar data points (clustering), without knowing the actual labels. Note that the combination of supervised and unsupervised data, known as semi-supervised learning, is also possible[25]. Imagine a dataset with 5000 webpages that need to be grouped into 10 distinct categories, e.g. science, nature, cooking, ... Only 100 of the 5000 pages in the train set are labelled. An approach to solve this problem could be to first cluster the pages in similar groups using unsupervised learning. As soon as a group contains a single labelled page, all pages in the group can be labelled accordingly. This is possible because clustering returns groups of similar samples. Semi supervised learning has the advantage that one can also use unlabelled data, which is often easier and cheaper to obtain, unlike labelled data which is usually quite rare; if there was a fast and easy way to label the data then there would not be a need for machine learning.

1.2.1 Over- and underfitting

Over and underfitting is a common problem in many machine learning projects[24]. Suppose the example in Figure 1.7, where one tries to find a good function to fit the given data points. Looking at the three proposed functions, one can easily see that the middle figure corresponds to the most logical generator function¡[1] of the red points.

[1] A generator function is a theoretical concept to describe the 'actual' function that generated the outputs.

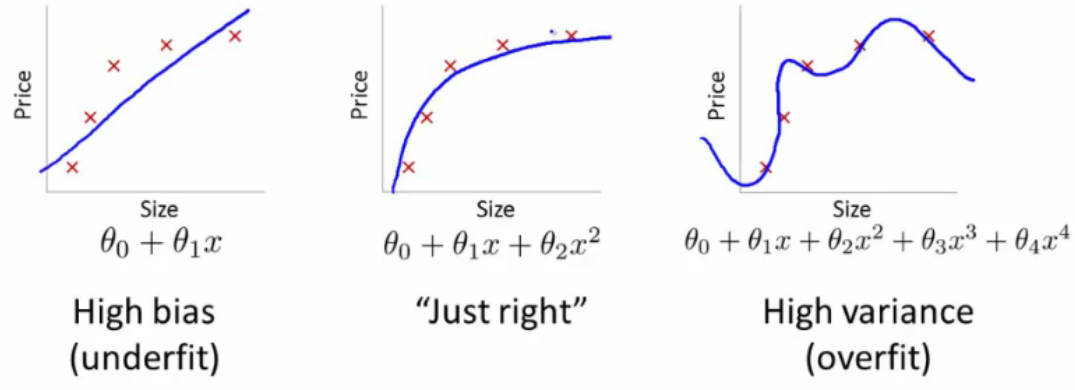

Figure 1.7: Overfitting versus underfitting[26].

The figure on the left corresponds to an underfit, where the proposed function is not able to capture sufficient detail of the points. The function is not complex enough to approach the generator function. As a result the best fit will always contain a relatively big error.

The function on the right corresponds to an overfit. The function fits or 'goes through' each point exactly, which will a very low error for these data points. However, one can see that the behaviour of the hypothesis function in between data points is not what one would expect. This is the result of using a too complex function to fit the data. As a results, all sample points are matched exactly, which results in a very low error for these points. However, the problems arise when the algorithm is test on unseen points. The algorithm will have a much higher error for those points.

Part of a good machine learning algorithm is finding the right tradeoff between overfitting and underfitting. In case of the aforementioned overfitting, it would be better to lower the performance of the algorithm on the sample datapoints, to gain performance on the 'unseen' datapoints. Different techniques exist to estimate how good an algorithm performs on unseen points. Often a part of the sample points is put in a test set that is neglected during training. After the algorithm is designed, the performance on the test set will indicate how well the algorithm generalises. It is only the performance of the test set that gives a fair estimation of the performance of a machine learning algorithm.

1.2.2 Feature selection

Feature selection is a technique that aims at selecting the features that perform well, while trying to remove irrelevant features[27]. The advantages of having a smaller features set are twofold. First having fewer features, will lower the risk of overfitting[27]. Second, knowing which features are important makes it possible for humans to interpret the machine learning model. In this thesis, knowing which features are relevant might help in gaining insight in the processing of emotion by the brain.

There exist several approaches to do feature selection. The first one is to simple use a statistical metric and remove all features with low correlation to the output. Another approach is to look

at the weights of a model. When a machine learning model gives a large weight to a feature, then that feature is considered more important than a feature with an assigned weight close to zero. Embedded methods also exist, they rely on the build in feature selection mechanisms of some machine learning algorithms. A more thorough overview of the different feature selection techniques is given in Section 2.5.

1.3 Problem statement

A lot of different physiological features are reported in the literature. Unfortunately, the literature does not fully agree on a specific set of features nor does it agree on what EEG channels and/or frequency bands are most important for emotion recognition. The features that are reported in different studies are often quite different, as you can see in Table 1.2 below.

study	**features used**
[28]	Alpha and beta power
[29]	PSD and asymmetry features
[30]	PSD
[31]	discrete wavelet transform of alpha, beta and gamma band
[4]	alpha/beta ratio, Fpz beta and alpha band power
[32]	PSD, RCAU, DCAU, DASM, RASM, DE

Table 1.2: Six different papers on emotion recognition, six different feature sets.

Another related problem with physiological signals is that they are very personal by nature. Features that work well for one person might not work well for another person[19]. Finding a set of features that works well for all persons is hard, but it might make the system more robust against personal differences.

The last problem is that is hard to compare the performance of different physiological feature studies, as they do not share the same dataset.

1.4 Goal of the thesis

The first goal is finding relevant features for emotion recognition in a person specific setting. This is already quite challenging as there are fuzzy boundaries and individual variation of emotion[33]. To do so, the output of different feature selection methods is compared. In a successful scenario, good features are found. These features could be used by a machine learning algorithm to accurately predict the emotions of one person. Some attention will also be spend on comparing non-EEG and EEG features to see which whether it is useful or not to include EEG and/or non-EEG signals in the emotion recognition.

The second goal is finding features for emotion recognition in a cross-person setting. In this setting features should generalise well across different persons, thus the algorithm should be able to recognize emotions from unseen persons. The comparison for non-EEG and EEG features will also be done here. Emotion recognition is harder in a cross-subject setting, since physiological signals are very personal[19].

Both goals are tackled by comparing a large range of different feature selection methods combined with a huge feature set. Additionally, the accuracy on the DEAP, a dataset designed to compare different emotion recognition studies[19] will be reported. This will ensure that the results obtained in this thesis can serve as a benchmark for future research. This is important as performance of emotion recognition algorithms based on physiological signals often varies a lot for different datasets[6].

The contents of this thesis are as follows. The next chapter gives an overview of the dataset, features and feature selection methods that are used in this thesis. It also gives an overview of similar state of the art emotion recognition studies.

!!(TODO)
Chapter 3 will give an overview of the obtained results in this thesis Chapter 4 states the conclusion Chapter 5 gives an overview of future research that is possible.

2

Methods

This chapter starts by explaining the used dataset and features. Next some state of the art methods are briefly discussed, to give an idea of similar research. After that the contributions of this thesis are explained. Next different feature selection methods are explained. The last part of this section explains the used approach in this thesis.

2.1 Dataset

One of the most used datasets in the context of emotion recognition is the Dataset for Emotion Analysis using Physiological Signals (DEAP)[19]. This dataset consists of several parts, the first part is a rating of 120 music videos by 14 - 16[1] persons. Each video is rated for valence, arousal and dominance on a scale ranging from 1 to 9 using self-assessment manikins (see later). This part of the dataset is not used during this thesis, because it contains no physiological signals.

The next part of the dataset is the physiological experiment that contains emotional reactions of 32 subjects. The emotional reactions were triggered using music video excerpts. Each subject watched 40 one-minute videos, while several physiological signals were recorded. These physiological signals consist of 32 channel, 512Hz EEG signals combined with peripheral physiological signals like respiration rate, skin temperature, etc. More concretely, this dataset contains following signals:

[1] 2 persons did not complete all the necessary ratings.

Channel	Name	Category	Channel	Name	Category
1	Fp1	EEG	**21**	F8	EEG
2	AF3	EEG	**22**	FC6	EEG
3	F3	EEG	**23**	FC2	EEG
4	F7	EEG	**24**	Cz	EEG
5	FC5	EEG	**25**	C4	EEG
6	FC1	EEG	**26**	T8	EEG
7	C3	EEG	**27**	CP6	EEG
8	T7	EEG	**28**	CP2	EEG
9	CP5	EEG	**29**	P4	EEG
10	CP1	EEG	**30**	P8	EEG
11	P3	EEG	**31**	PO4	EEG
12	P7	EEG	**32**	O2	EEG
13	PO3	EEG	**33**	hEOG	non-EEG
14	O1	EEG	**34**	vEOG	non-EEG
15	Oz	EEG	**35**	zEMG	non-EEG
16	Pz	EEG	**36**	tEMG	non-EEG
17	Fp2	EEG	**37**	GSR	non-EEG
18	AF4	EEG	**38**	respiration belt	non-EEG
19	Fz	EEG	**39**	plethysmograph	non-EEG
20	F4	EEG	**40**	skin temperature	non-EEG

Table 2.1: The available signals in the DEAP dataset.

A preprocessed version of the physiological experiment database is also available. In this version, the EEG recordings were downsampled to 128Hz and noise and EOG artifact removal was performed. A bandpass filter was applied to filter out frequencies below 4Hz and above 40-45Hz. This was done to remove noise, since most muscle and eye artifacts have a frequency around 1.2Hz and artifacts caused by nearby power lines, have a frequency around 50Hz[4]. This thesis uses the preprocessed version of the DEAP since it is the most practical version to use.

Additionally facial video for 22 of the 32 subjects was recorded, so research of facial expressions is also possible with this dataset. All videos are rated on 4 scales: arousal, valence, dominance and liking. The liking component indicates how much each person liked each video excerpt. It is important not to confuse the liking component with the valence component, as it inquires information about the participants' tastes, not their feelings. For instance, a person can like a video that triggers angry or sad emotions[2]. The liking rates are mentioned here for completeness and are not used in this work as they are not part of the emotion space.

For assessment of these scales self-assessment manikins (SAM) were used[19]. SAM visualizes the valence, arousal and dominance scales with pictures. Each picture corresponds to a discrete value. The user can click anywhere in between the different figures, which makes the scales continuous. All dimensions are given by a continuous value between 1 and 9.

The used SAM figures are shown in Figure 2.1. The first row gives the valence scale, ranging

[2]However strong correlations between the liking and valence ratings were observed[19].

from sad to happy. The second row shows the arousal scale, ranging from bored to excited. The last row represents the different dominance levels. The left figure represents a submissive experience, while the right figure corresponds with a dominant experience.

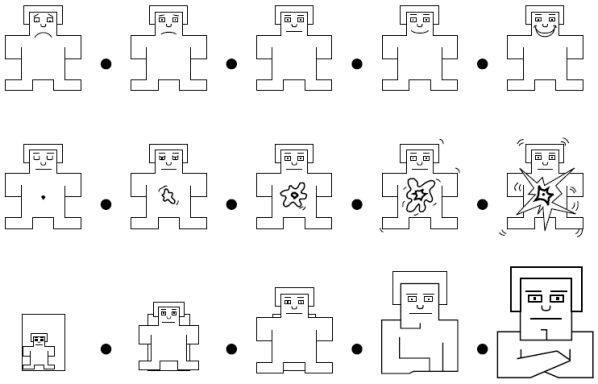

Figure 2.1: The images used for the SAM[19].

2.2 Features

Machine learning algorithms require good features to perform well[3]. In the context of this thesis, good features should be correlated with the subject's emotional state. Two categories of features are observed in this work: EEG features and non-EEG features. Both categories are covered in the following sections.

2.2.1 EEG-features

EEG features are extracted from the electroencephalography measurements from the subject's scalp. From these signals a lot of different signals can be extracted. The power spectral density (PSD) of a signal gives the distribution of the signal's energy in the frequency domain. By calculating the spectral density for different frequency bands of the signal, one can determine how much power of each frequency band is in the signal.

Differential entropy (DE) is defined as follows [32]

$$DE_{channel} = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{(x-\mu)^2}{2\sigma^2}) log(\frac{1}{\sqrt{2\pi\sigma^2}}) exp(\frac{(x-\mu)^2}{2\sigma^2}) dx$$

It is proven that the differential entropy of a certain band is equivalent to the logarithmic power spectral density for a fixed length EEG sequence[35]. This simplifies the calculations significantly.

$$DE_{channel} = log(PSD_{channel})$$

[3]There are some exceptions, for instance some types neural networks are capable of 'designing' their own features[34]. But these algorithms were not used in this thesis.

The most used feature for valence recognition is the frontal asymmetry of the alpha power[10]. The right hemisphere is generally speaking, more active during negative emotion than the left hemisphere. The left hemisphere is in turn more active during positive emotions[3, 36, 32]. The asymmetry can be calculated in different ways. First, one can calculate the differential asymmetry (DASM) , where the left alpha power is subtracted from the right alpha power.

$$DASM = DE_{left} - DE_{right}$$

Another way to measure the asymmetry is by division. The Rational Asymmetry (RASM) does exactly this and is given by:

$$RASM = \frac{DE_{left}}{DE_{right}}$$

With DE_{left} and DE_{right} being the left and right differential entropy respectively. Another reported feature in literature is the caudality, or the asymmetry in fronto-posterior direction[37]. Caudality measures the difference in power between the front and the back of the scalp. This can again be calculated in two ways. The first method is the differential Caudality (DCAU) , defined as:

$$DCAU = DE_{front} - DE_{post}$$

The second method to determine the Caudality is the Rational Caudality (RCAU) , which is defined as:

$$RCAU = \frac{DE_{front}}{DE_{post}}$$

With DE_{front} and DE_{post} being the frontal and posterior power respectively. Arousal is usually determined, by looking at the different frequency bands[4]. Each frequency and has their own medical interpretation, see 1.1.1. Alpha power corresponds to a more relaxed brain, while Beta power corresponds to a more active brain. The alpha / beta ratio therefore seems a good indicator for the arousal state of a person.

The Alpha/ Beta ratio is limited to comparing two frequency bands. Other frequently used features are fractions of PSD. These fractions indicate what proportion of power a certain frequency band has. They are defined for a channel, given by:

$$frac_{band,channel} = \frac{power_{band,channel}}{power_{total,channel}}$$

These fractions give insight in the distributions of wavebands at different channel locations.

2.2.2 non-EEG features

The aforementioned EEG features are just one class of physiological features, the DEAP dataset contains several other physiological measurements[19]. For each of these measurements the average, standard deviation, variation, median, minimum and maximum are calculated.

The Galvanic Skin Response uses two electrodes on the middle and index finger of the subjects left hand to measure the skin resistance. It has been reported that the mean value of the GSR is related to the level of arousal[20, 19].

The respiration belt (RSP), indicates the user's respiration rate. Slow respiration is linked to relaxation (low arousal). Fast and irregular respiration patterns corresponds to anger or fear, both emotions have low valence and high arousal[19].

A plethysmograph is a measurement of the volume of blood in the subject's left thumb. This can be used to determine the blood pressure of a subject. Blood pressure offers valuable insight into the emotional state of a person. For instance, stress is known to increase blood pressure[19].

The heart rate is not directly available in the DEAP dataset. fortunately, it can be extracted from the plethysmograph, by looking at local minima and maxima[19]. This is visible when looking at the plethysmograph's output, shown in Figure 2.2.

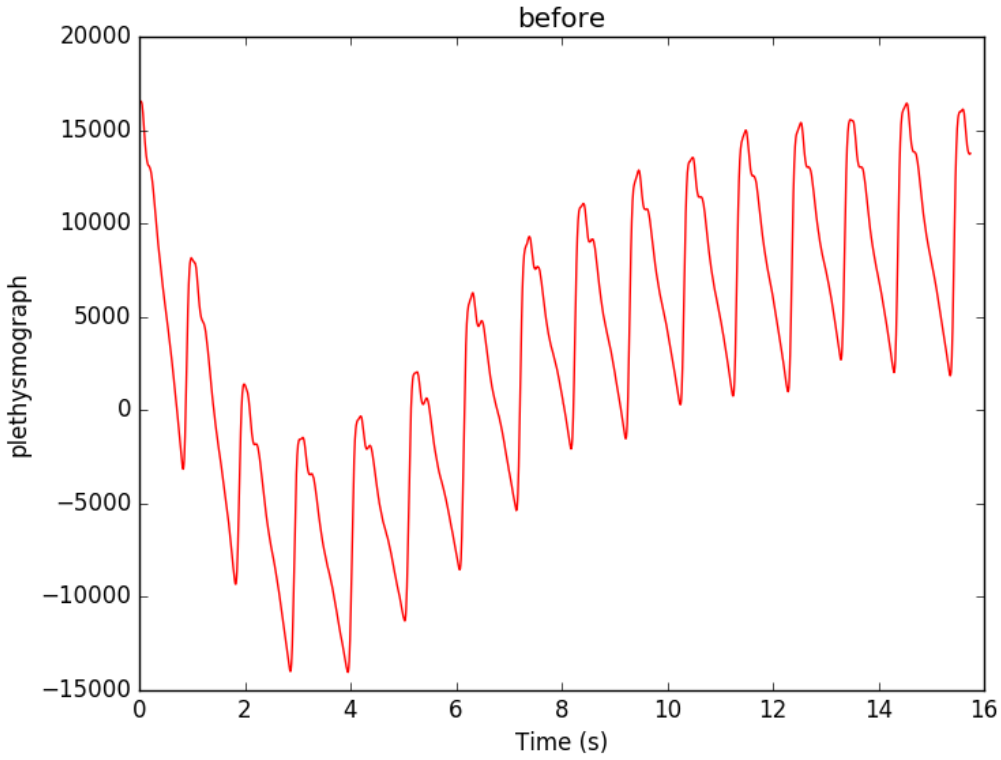

Figure 2.2: The plethysmograph before smoothing.

The heart rate extraction is done in two steps. In the first step the plethysmograph's output is smoothed to filter out high frequency components. This is done to avoid that noise is selected as a local optima. In the second step the local extrema are located, as shown in Figure 2.3

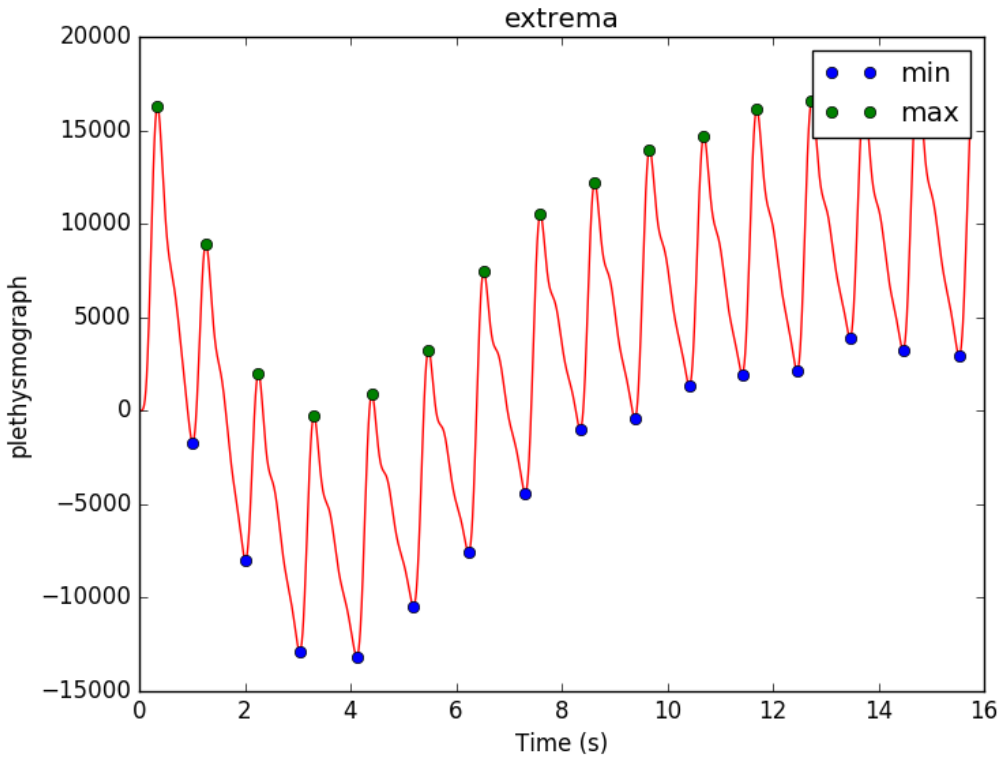

Figure 2.3: The local optima in the plethysmograph.

The combination of a local minimum and maximum correspond to a heart beat[19]. Therefore, the time between two consecutive local minima or maxima correspond to the time between two heart beats, known as the interbeat interval. Getting the average heart rate from the interbeat interval is straight forward. Lastly, the skin temperature of the subject is also available.

2.2.3 Overview

The following table gives an overview of the different features and their amount. The 32 EEG channels can be found in Table 2.1. There are 6 frequency bands: Alpha, Beta, Gamma, Delta, Theta and All. All refers to taking the total power of a channel. Note that the fractions only have 5 different frequency bands, as the percentage of all power would always be 100%.

Name	Type	no. Channels	no. Frequency bands	Total
PSD	eeg	32	6	192
DE	eeg	32	6	192
DASM	eeg	13	6	78
RASM	eeg	13	6	78
DCAU	eeg	11	6	66
RCAU	eeg	11	6	66
Frac	eeg	32	5	160
Alpha / Beta	eeg	32	1	32
EEG Total				864

Name	Type	no. Statistics
HR	non-eeg	6
Plethysmograph	non-eeg	6
GSR	non-eeg	6
ST	non-eeg	6
RSP	non-eeg	6
non-EEG Total		30

Overall Total	**894**

Table 2.2: An overview of the different features that were compared in this thesis.

The feature set has a size of 894, which is huge considering that there are only 40 samples for each person. Using this many features in combination with the low sample count, will quickly result in overfitting[24]. To solve this problem, one can either increase the number of samples or decrease the number of features. Increasing the number of samples is hard. Since EEG data is very personal [19], several recordings of the same subjects are required.

Reducing the feature set in size is another possibility. Two methods exists, dimension reduction and feature selection. The difference between dimensionality reduction and feature selection is that dimensionality reduction methods consider all information in the feature space. Feature selection methods, on the other hand, take a subset of the information[6].

This problem is even more severe in cross-subject emotion recognition system. Here, it is not possible to simply take a limited subset of features. Physiological signals are very personal by nature [19]. Selecting features that work for one person, might therefore not work well on different persons.

2.3 State of the art

This section will give an overview of similar studies and their conclusions. Some of these studies also did some research on cross-subject emotion recognition. Emotion recognition is still in its infancy[33] and subject independent features are hard to find [19]. Therefore, research is aimed more towards person specific emotion recognition systems.

2.3.1 DEAP method

The first method of emotion recognition is the DEAP method, described in the DEAP paper[19], the paper that introduces the DEAP dataset used in this thesis. The research found that Valence shows the strongest correlations with the EEG signals. Additionally the study found correlations in all frequency bands, with an increase in power for the lower range wavebands for an increase in valence. These effects occur in the occipital regions of the brain, above the visual cortices. This might indicate that the subject is focussing on a pleasurable sound. A central decrease in beta power was observed together with a occipital and right temporal increase in power for positive emotions. The research conclude that these observed correlations concur with other neurological studies. The absolute value of the correlations are seldom bigger than 0.1 for a cross person setting. This indicates that cross person emotion recognition is a non trivial problem. The absolute values of the person specific correlations were around 0.5.

The DEAP paper also propose their own classification method for person specific emotion classification. They start by performing feature selection using the Fisher's linear discriminant for feature selection. The Fisher's linear discriminant is defined as:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2}$$

With μ and σ being the mean and standard deviation of feature f. The Fisher's discriminant was calculated for each feature, before a threshold of 0.3 was applied to filter out irrelevant features. The used classifier was a Naive Bayes classifier, which assumes independence of features. The Naive Bayes classifier is a simple classifier that uses the following equation:

$$G(f_1, ..., f_n) = argmax_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

With F being the set of features and C the classes. $p(F_i = f_i | C = c)$ is estimated by assuming Gaussian distributions of features and modelling these from the training set.

2.3.2 Stable emotion recognition over time

EEG patterns are not only subject dependent, they are also dependent on the subjects mood and thus might vary in time[32]. This work starts by researching different EEG features: PSD, DE, DASM, RASM, DCAU, RCAU[4]. The different features are tested on the DEAP dataset. Afterwards, they develop a new dataset, where subjects have repeated trial sessions with some time in between. This dataset is then used to measure the performance of their time independent, subject specific, emotion recognition system.

Their machine learning set-up is as follows, first they perform feature extraction of the aforementioned features. Then feature smoothing is done using a Linear Dynamic system (LDS) , that can be expressed by:

$$x_t = z_t + w_t$$
$$z_t = Az_{t-1} + v_t$$

[4]Note that these features are explained in more detail in Section 2.2.

x_t denotes the observed variables or features, while z_t denotes the hidden emotion variables. A is a transformation matrix and w_t is Gaussian noise. The need for a linear dynamic system is supported by the assumption that emotion change gradually over time. The LDS filters out components that are not associated with emotional states.

The list of features at this point is too big and may contain uncorrelated features that might lead to performance degradation of the classifier. Two methods for this are compared, principal component analysis (PCA) and minimal redundancy maximal relevance (MRMR).

PCA uses an orthogonal transformation to create a lower dimensional feature space starting from the original higher dimensional feature space. It does so by minimizing the loss of information, i.e. the principal component should have the largest possible variance. PCA is explained later in Section 2.5.3.

PCA cannot preserve original domain information like channel and frequency, therefore the paper also uses the MRMR method. MRMR uses mutual information in combination with maximal dependency criterion and minimal redundancy. The algorithm starts by searching features satisfying:

$$maxD(S,c), D = \frac{1}{|S|} \sum_{x_d \in S} I(x_d; c)$$

Where S is the feature subset to select. When two features are highly correlated, the maximal dependency is not likely to change when one of the correlated features is removed. This is expressed by the minimal redundancy condition.

$$minR(S), R = \frac{1}{|S|^2} \sum_{x_{di}, x_{dj} \in S} I(x_{di}, x_{dj})$$

The two conditions are then combined to from the Maximal Relevance Minimum Redundancy, which can be expressed as:

$$max\varphi(D,R), \varphi = D - R$$

Note that incremental search methods exists and are often used in practice. After performing the dimensionality reduction, the samples from the DEAP data set are classified in high / low valence and high/low arousal, giving a total of four classes. All values close to the separation border are removed from the training data, as they might confuse the classifier.

For the classification, three conventional and one newly developed pattern classifiers were compared. k-nearest neighbors (KNN) , logistic regression (LR), Support Vector Machines (SVM) and Graph regularized Extreme Learning Machine (GELM) . Extreme Learning Machine (ELM) is a single layer feed forward neural network[38]. GELM is based on the idea that similar shapes should have similar properties and obtains better results for face recognition [39] and as the paper concludes, also for emotion classification.

The study found then performed a study on the different features and concluded that DE features are the most suitable EEG features, followed by the asymmetry features (RASM, DASM, DCAU and RCAU). The LDS smoothing was also found to be the better feature smoothing method.

2.3.3 EEG-based emotion recognition in music listening

This study[33] uses EEG features to recognize 4 different discrete emotions (joy, anger, sadness, pleasure) induced by music. They compared four different feature sets on 6 different wavebands: RASM and DASM of 12 channelpairs, raw PSD of the 24 channels and PSD of 30 channels (including 6 midline channels). The compared set of wavebands consists of: alpha, beta, gamma, delta, tetha and all wavebands. These features were fed to two different classifiers, one Multilayer perceptron (MLP) and an SVM.

Their main results were that the DASM features worked better than the RASM features and even better than using the corresponding 24 PSD features. They also did research to person independent EEG features and found that their accuracy remained consistent. Note that while these results sound promising, they were unfortunately not performed on the DEAP dataset. Performance of emotion recognition algorithms is known to vary a lot between datasets[6].

2.3.4 Comparing selected methods for feature extraction and classification

In this comparitive study four distrinct emotions (joy, anger,sadness and pleasure) were classified [6]. The emotions were triggered by songs that were selected for each subject. The subjects were instructed to select songs themselves, that trigger memories. These memories should in turn, trigger the desired emotions. The four emotional states were mapped in the valence-arousal model. The used features were typical statistical values of physiological signals (Skin Conductivity (SC), Electrocardiogram (ECG), Electromygraphy (EMG) and Respiration rate (RSP).

Several feature selection techniques were compared. The first one is the analysis of Variance (ANOVA) where the best D features were taken. Sequential forward selection (SFS) , where the algorithm starts with an empty feature set and then introduces a new feature in each iteration. Sequential backward selection (SBS) is an alternative, where a feature is removed in each iteration. These feature selection methods were also compared to two dimensionality reduction methods: PCA and Fisher projection.

The newly formed feature space was then fed to three different classifiers: K-nearest neighbors, Multilayer perceptron and Linear discriminant function. The results indicated that it is easier to classify arousal than valence. This might indicate that non-EEG features might be features for arousal classification, as this work only contains non-EEG features. SFS in combination with fisher seems to give the best classification performance, closely followed by LDF and ANOVA, a less computationally intensive method.

The paper also concludes that joy was characterized by a faster heart rate, while sadness was identified by low SC and EMG signals. There was also a higher breathing rate for negative valence emotions. They reported limited similarities for the selected features between subjects.

2.3.5 Advanced RF feature selection

One advanced method for feature selection is the two-step method using random forest[27]. There are two possible motivations for feature selection. The first motivation is to do interpretation, find out which features are important and use them for research. In the context of

this work, feature interpretation could help neuroscientist find out which parts of the brain are affected by emotion. The second motivation is to improve machine learning techniques. Having fewer features will not only speed up training and prediction times, it also reduces the complexity. Reducing complexity often has a good influence on the generalisation property of a machine learning algorithm[24, 27]. Additionally in the context of EEG data gathering, using fewer electrodes means less preprocessing time; mounting 32 electrodes to the brain of a subject is a time consuming task.

The selection procedure itself consists of two steps. In the first step, data is fitted to a random forest and the importance values for each feature are determined. The importance values are then averaged and the standard deviation of the importances over all trees is calculated. All features are then ranked based on their importance ranking. Next features with small importance values are cancelled.

Then, depending on the motivation of feature selection, one of two possible second steps is performed. For feature interpretation the second step starts by fitting a random forest using a single feature. The OOB is then averaged over multiple runs. The runs are needed because a random forest has an element of randomness; fitting the same data twice to a random forest, will not give you the same random forest. To get an accurate estimate of the performance of a random forest, fitting the data several times is required. The average OOB score and its standard deviation is then used to determine an initial OOB score.

$$OOB_{init} = AVG(OOB) - STD(OOB)$$

The standard deviation is used to avoid noisy results. This means that a result is only regarded as better, when it is better from a statistical point of view. Next features are added iteratively, when a larger features set has a better average OOB score (taking the standard deviation into account), the feature set is replaced by the larger feature set. Note that the whole set of features is always considered; it is not possible to leave a feature out and include the next feature.

The other possible second step is used for prediction, here the algorithm starts similarly, by determining an initial average OOB score and standard deviation. The idea behind the standard deviation is the same as with the interpretation step, noise removal and stability.

$$OOB_{init} = AVG(OOB) - STD(OOB)$$

The next part is different, in each iteration a feature is introduced. When the average OOB score and standard deviation of the feature are better, the feature is added to the feature set, otherwise it is neglected. This is a greedy forward selection algorithm, once a feature is selected, it remains selected. The difference between the interpretation and prediction step is that here single features are added to the feature set, while step two-interpretation always takes the whole feature set, meaning all features before the last feature, as a replacement. The prediction version of step two is able to select a distinct set of features out of the results from step one.

In the end the paper notes several observations, the step two-prediction method provides better OOB scores using fewer features. Additionally they mention that highly correlated features might confuse the algorithm, as correlated features have lower importances.

2.4 Contribution of this thesis

It is clear from the aforementioned papers, that some research has already been done for emotion recognition. The contribution of this thesis is threefold, first compare a bigger ranger of feature selection methods on a bigger set of physiological features. Second, compare EEG features to non-EEG features to see how much information can be retrieved from the EEG signals compared to the non-EEG signals, which are usually easier to obtain. Third, perform the feature selection methods in a cross-subject setting to see which feature generalise well across subjects. It is also important to note that the feature selection will be performed on the DEAP dataset, so that it can serve as a benchmark. This is important as performance of emotion recognition algorithms based on physiological signals often varies a lot for different datasets[6].

2.5 Feature selection methods

Feature selection is the process of selecting good features from a set of features. The need for this is twofold: first reducing the number of features, is a protection mechanism against overfitting[27]. This is important when a smaller dataset is used. Second, reducing the number of features can speedup the learning process of a learning algorithm as fewer parameters need to be optimized. Additionally, in the context of research, looking at which features are important might give more insight in how emotion is processed by the brain. For example, knowing what features are relevant can help neuroscientists understand the working of the brain better. There is also a practical use of feature selection, limiting the physiological signals to fewer channels, can help the set-up time. Mounting an EEG cap to a subject is a time consuming process. Using fewer electrodes can make the system more convenient to use as it would save time.

Several approaches for feature selection exists: filter methods, wrapper methods and embedded methods. What follows is an explanation of how each approach works, combined with the used methods of each approach in this thesis.

2.5.1 Filter methods

Filter feature selection methods use an independent metric or statistical test to filter out features with low importance. The most simple example of to simple look at the correlation between each feature and the output. Afterwards, all features with low correlation can be removed.

Pearson correlation

The Pearson correlation coefficient measures the linear relationship between two variables. The output is a value r, that lies between -1 and 1, corresponding to perfect negative correlation and perfect positive correlation respectively. A correlation value of 0 means that there is no correlation.

More formally[40], the Pearson product-moment coefficient of correlation, r between variables X_i and Y_i of datasets X and Y is defined as:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

with

$$SS_{xy} = \sum_i (X_i - \tilde{X})(Y_i - \tilde{Y})$$

and

$$SS_{xx} = \sum_i (X_i - \tilde{X})^2$$
$$SS_{yy} = \sum_i (y_i - \tilde{Y})^2$$

The Pearson correlation coefficient is fast and simple to calculate, but has some major shortcomings. First off, it can only see linear relationships and will not see the correlation between a value x and x^2.

In the context of this thesis, whether the correlation is positive or negative is not important; a learning algorithm needs features that have a significant correlation. As a result the absolute value of the r value is used as this allows for a more convenient comparison of correlations.

Normalized mutual information

Mutual information is a more robust option for correlation estimation. The mutual information, MI, of two variables X and Y is defined as [41]:

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log(\frac{p(x,y)}{(x)p(y)}$$

Using the mutual information directly for feature ranking might be inconvenient because its results does not lie in a fixed range. Fortunately, normalized variants of the mutual information score exists. The normalized mutual information, NMI, of variables X and Y is given by:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{(H(X)H(Y))}}$$

With $H(X)$ and $H(Y)$ being the Shannon entropy of variable X and variable Y, defined as:

$$H(X) = \sum_{i \in X} p_i log(\frac{1}{p_i}) = -\sum_i p_i log(p_i)$$
$$H(Y) = \sum_{i \in Y} p_i log(\frac{1}{p_i}) = -\sum_i p_i log(p_i)$$

Distance correlation

Distance correlation solves some shortcomings the Pearson correlation has. The Pearson correlation coefficient might give a correlation of zero for dependent variables, as shown in Figure 2.4.

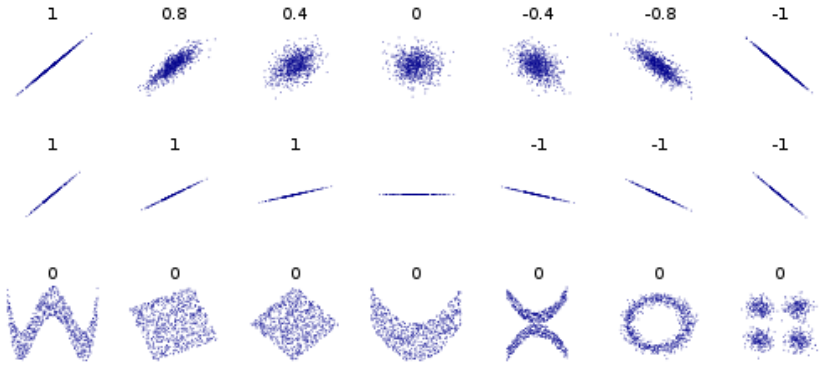

Figure 2.4: Pearson correlation coefficients for different sets of (x,y) points. Note that many coefficients are zero, while there clearly is some correlation. Source: Wikipedia

The distance covariance, sometimes referred as the Brownian covariance, addresses this problem[42]. Its main idea is that a good measurement for dependence is the 'distance' between the join distribution f_{XY} and the product of the marginal distributions f_X and f_Y weighted by a weight function W. This gives the following theoretical function:

$$dCorr_{X,Y} = W(||f_{XY} - f_X f_y||)$$

The result is that the distance correlation metric gives very different results, as you can see when comparing the distance correlation outputs in Figure2.5 with the Pearson correlation outputs in Figure 2.4.

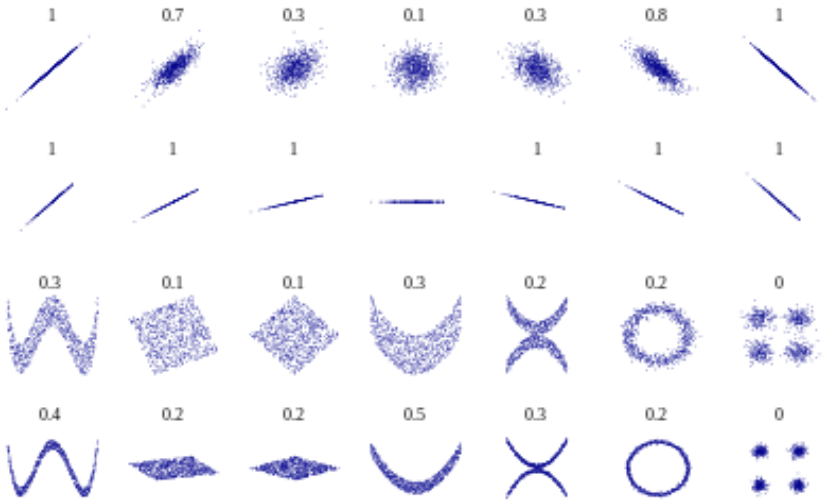

Figure 2.5: Distance correlation coefficients for different sets of (x,y) points. Note the difference with the Pearson correlation coefficients in Figure 2.4. - Source: Wikipedia

Without going further into the theory, the distance correlation between two variables X and Y, each with n data points can be calculated as follows.

First compute all pairwise Euclidean distances for both variables.

$$[D_x]_{j,k} = ||X_j - X_k||$$
$$[D_Y]_{j,k} = ||Y_j - Y_k||$$
$$j, k = 1, 2, ..., n$$

The result is two n by n distance matrices D_x and D_y. Next, both matrices are centered:

$$S_x = C_n D_x C_n$$
$$S_y = C_n D_y C_n$$

Finally, the covariance is computed.

$$\nu^2(X, Y) = \frac{1}{n^2} \sum_l \sum_k [S_x]_{k,l} [S_y]_{k,l}$$

This is the distance covariance, which is not normalized. The distance correlation is the normalized version of the distance covariance, $dCorr$, which is defined by:

$$dCorr(X, Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X) dVar(Y)}}$$

With $dCov(X, Y)$ being the aforementioned distance covariance, $dVar(X)$ and $dVar(Y)$ are the distance standard deviations. The distance correlation has the disadvantage that is much slower than mutual information or Pearson correlation, but in return, the distance correlation is able to detect more complex relationships between two variables.

Analysis of variance

Analysis of variance (ANOVA) is a statistical test to analyse differences between groups. The idea is that the total variance, found in the samples consists of two parts. The first part is the variance within a single group, the second part is the variance between groups.

Suppose you want to test the influence of caffeine on the reaction speed[5]. To do so, you take two groups of each 10 persons. The first group has to drink a large cup of coffee, the second group is the control group that only drinks water. Next the reaction times of all persons in both groups are measured. From these results it is possible to calculate the total variance as well as the variance within each group and the variance between the groups.

If the variance within each group is much larger than the variance in between the groups, one concludes that the groups are similar. The reaction time is thus dependent on the person and not on the caffeine. However should the variance between the groups be much bigger than the variance within each group, than one concludes that the variance in reaction time is caused by the caffeine and not by personal difference.

2.5.2 Wrapper methods

These methods select features by applying an arbitrary machine learning technique and looking at the coefficients of the features. The idea is that features with high coefficients have more influence on the end results than features with a lower coefficients and are therefore more important. Again absolute values are used, since a perfectly negative correlated variable is as useful as a perfectly positively correlated variable.

Linear regression

A first method is simple linear regression, where a linear combination of features is searched that produce a good estimate of the output value. Linear regression can achieve good results when the data does not contain a lot of noise and the features are (relatively) independent. When the set of features contains correlated features, the model becomes unstable. As a result, small changes in input data might lead to huge differences in output coefficients. for example assume the 'real output' is given by $Y = X_1 + X_2$ and the dataset contains output in the form of $Y = X_1 + X_2 + \epsilon$ with ϵ being some random noise. Further more assume that X_1 and X_2 are linearly correlated, meaning that $X_1 \approx X_2$. The suspected output of the model should be $Y = X_1 + X_2$, but since noise is added the algorithm might end up with arbitrary combinations of X_1 and X_2, e.g. $Y = -X_1 + 3X_2$. the result will rate one feature much higher than another one, while in reality they are of equal importance. This is due to the noise. While maximizing the performance, the algorithm will minimize the influence of noise on the output, which results in unstable behaviour.

SVM

A Support vector machine (SVM) is a well known and proven method for machine learning. It has been used in several emotion recognition studies. An SVM works in essence by creating a

[5]This example was based on the following video: https://www.youtube.com/watch?v=ITf4vHhyGpc

hyperplane that separates two classes. Shown in Figure 2.6 is a simple line separating the red from the blue balls.

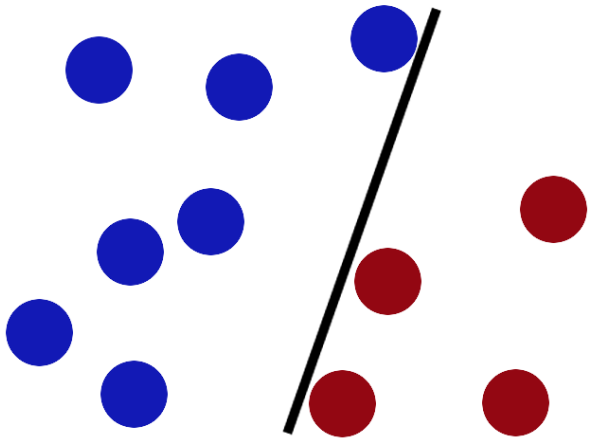

Figure 2.6: One possible separation border.

This is one possible solution, but note that an SVM will always search for a decision boundary that maximizes the boundary between the two classes, shown in Figure 2.7.

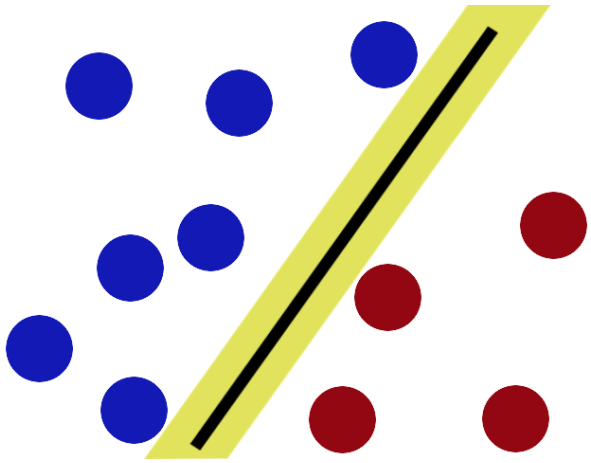

Figure 2.7: A separation with maximal boundary.

This all works well, as the balls are separable using a single straight line. This is not always the case though. Shown in Figure 2.8 is a scenario where it is not possible to separate the red balls from the blue ones using a single straight line.

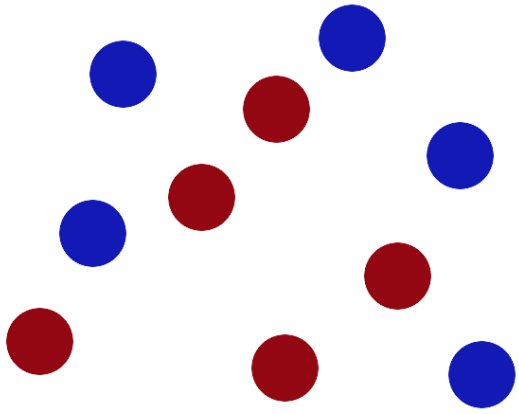

Figure 2.8: There exists no possible line that can separate the red balls from the blue ones.

A solution for this is to transform the input space to the feature space, where it is possible to separate the balls using a hyperplane, this is shown in Figure 2.9. Different transformations are possible. Each transformation corresponds to a different kernel, the component of an SVM that handles the transformation.

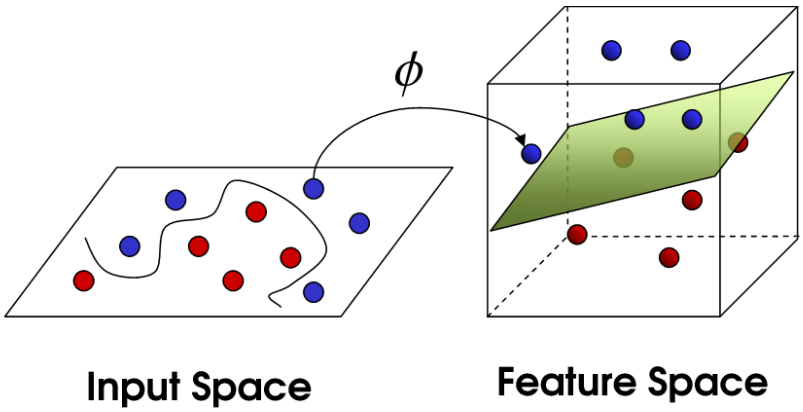

Figure 2.9: Transformation to a new features space where the balls can be separated by a hyperplane.

Back in the original feature space the separation boundary might look like Figure 2.10.

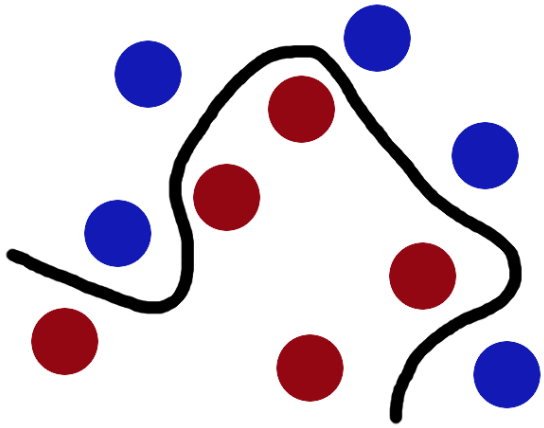

Figure 2.10: Separation boundary in the original feature space.

Linear discriminant analysis

Linear Discriminant Analysis (LDA), is a machine learning technique often used in combination with CSP[43, 44, 45]. LDA looks for a projection of the data where the data is linearly separable, as shown in Figure 2.11. Looking at the coefficients of the LDA model, one can again determine the importance of the different features.

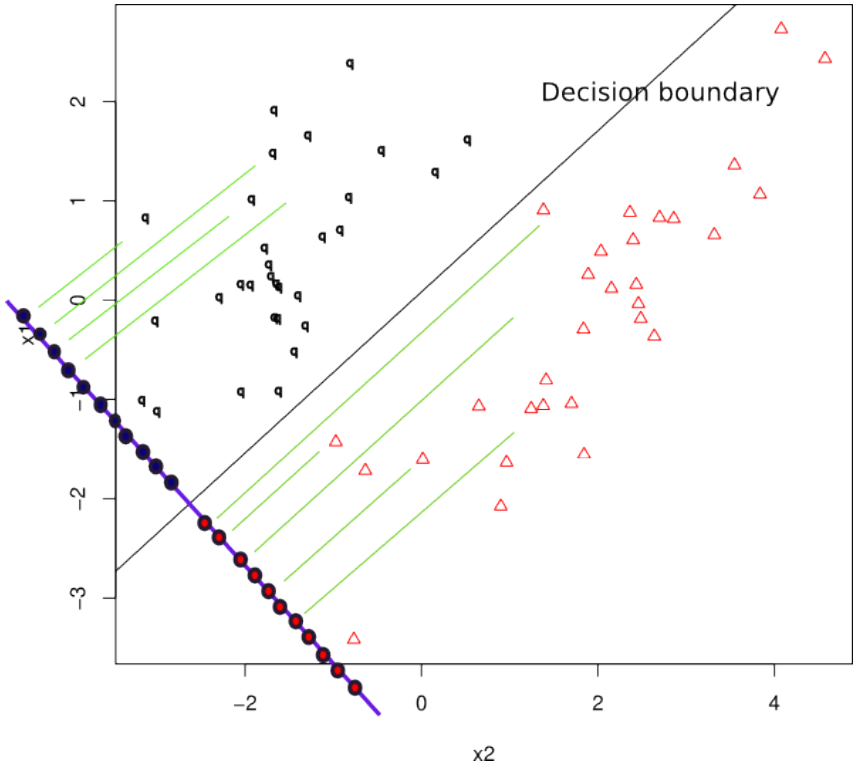

Figure 2.11: LDA finds a projection of the data where the separation of the data is clear.

2.5.3 Embedded methods

Embedded feature selection methods are methods that are build-in for some machine learning algorithms.

Lasso regression

Lasso regression uses L1 regularization, that adds a penalty $\alpha \sum_{i=1}^{n} |w_i|$ to the loss function. the result is that the coefficients of weak features are forced to zero, as each non-zero feature adds to the penalty. This form is regularization is thus quite aggressive, it removes weak features completely and selects the good features. The problem with this is, similar to linear regression, stability. Coefficients can vary significantly, even for small changes in training data, when there are correlated features.

Ridge regression

Ridge regression uses L2 regularization, which add a L2 norm penalty to the loss function, given by $\alpha \sum_{i=1}^{n} w_i^2$. Where the L1 norm forces the coefficients to zero, the L2 regularization forces the coefficients to be spread out more equally. The result is that correlated features tend to get similar coefficients, as this minimizes the loss function, which in turn results in a more stable model. The disadvantage of ridge regression is that bad features still have low weights. This means that they still have an influence on the output.

Random forests

A random forest (RF) is an efficient learning algorithm based on model bagging and aggregation ideas[27]. The Random forests work by creating different decision trees. On their own, decision trees are very prone to overfitting. Random forests solve this problem by creating an aggregation of trees.

The word random in random forest indicates that some randomness is included. Each tree in a random forest looks at a random subset of the samples and a random subset of the features. This principle is shown in Figure 2.12. This random subset of samples is called the bootstrap sample and is selected out of N samples, by picking N samples with replacement. This results, on average, in 2/3 of the samples being selected (with some doubles). The other 1/3 of the samples are then used as out of bag (oob) set. Averaging the performance of each tree on the out of bag set, offers an indication of the generalisation of the random forest.

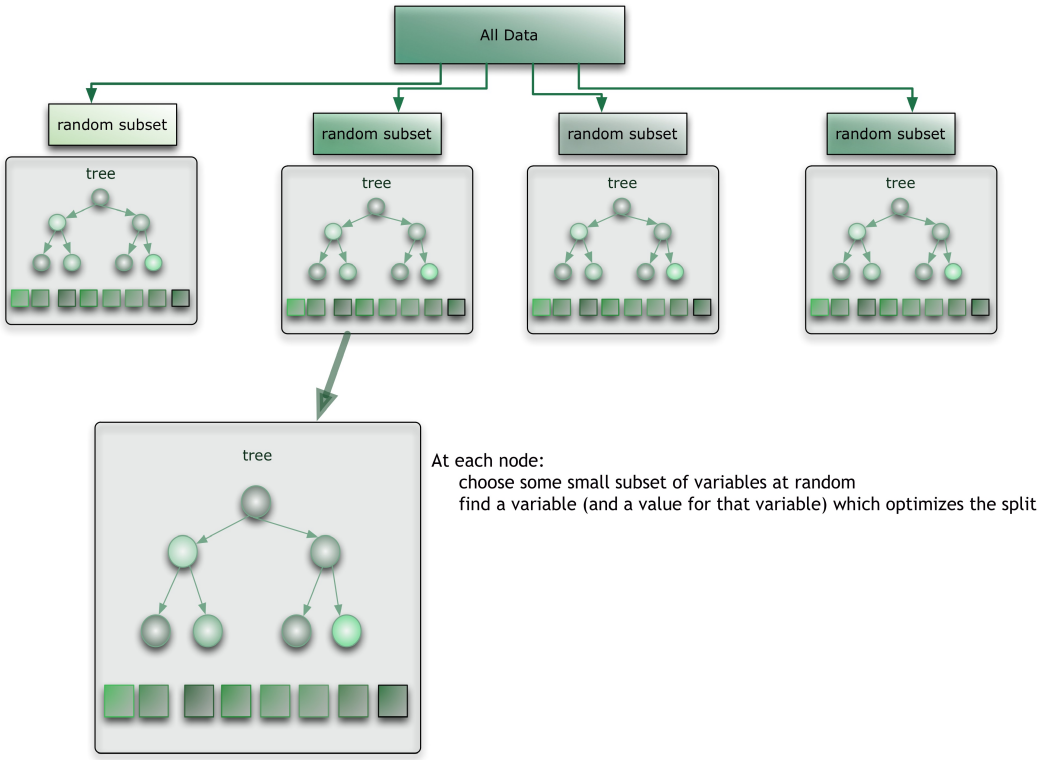

Figure 2.12: The structure of a random forest, found at [46]

To understand which features are good, one needs to understand the internal workings of a decision tree. Suppose the following example[6], where one tries to find an algorithm to predicted whether or not a person will play tennis on a given day. Suppose the training data is given by Table 2.3 and a prediction for the 15^{th} sample needs to be made.

[6]This example is based extensively on this youtube video: https://www.youtube.com/watch?v=eKD5gxPPeY0

Table 2.3: suppose the following training examples for a decision tree.

Day	Outlook	Humidity	Wind	Play tennis
1	sunny	high	weak	no
2	sunny	high	strong	no
3	overcast	high	weak	yes
4	rain	high	weak	yes
5	rain	normal	weak	yes
6	rain	normal	strong	no
7	overcast	normal	strong	yes
8	sunny	high	weak	no
9	sunny	normal	weak	yes
10	rain	normal	weak	yes
11	sunny	normal	strong	yes
12	overcast	high	strong	yes
13	overcast	normal	weak	yes
14	rain	high	strong	no
15	rain	high	weak	?

A decision tree will take a feature and split the data based on the possible outcomes of this feature. In case the features are continuous values, ranges are selected. In some cases the leafs will be pure, meaning that all samples in this leaf belong to a single class. The pure leaves in Figure 2.13 are displayed in green. In case the leave is not pure, another split is needed. Note that not all random forests split until all leaves are pure; random forest can be limited in depth, in that case the output is chose by a majority voting of the samples.

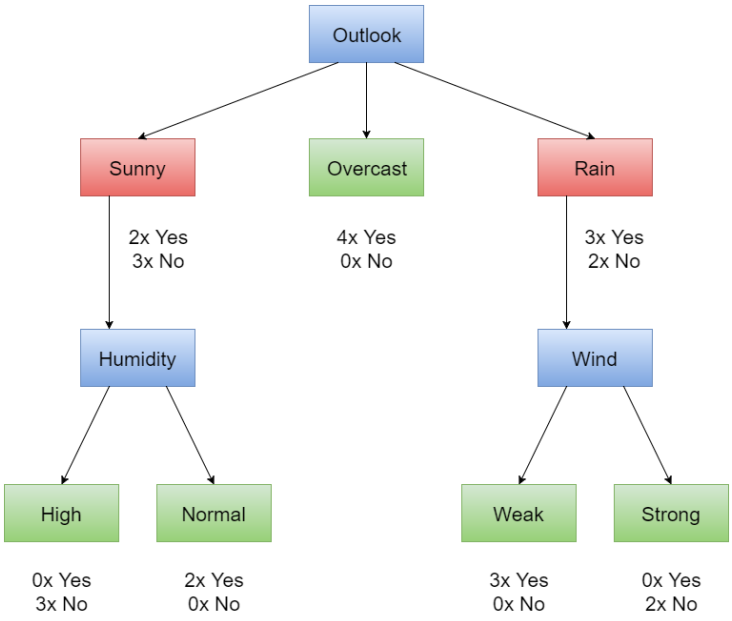

Figure 2.13: A decision tree for the data in Table 2.3

Once the tree is constructed it becomes clear that the predicted output of sample 15 is 'Yes'. This is obtained simply by following the tree branches. Even though the features are selected at random, they have influence on the accuracy. Good features will reduce the impurity significantly, thus the impurity reductions are a good indication for how important a feature is.

The importance is averaged over different nodes and different trees. As a results, random forest are also capable of detecting combinations of features that work well. One feature may not be important on its own, but might be a very good feature when combined with other features. Suppose the following example in Table 2.4:

label	**feature A**	**feature B**
Happy	+	+
Happy	-	-
Sad	-	+
Sad	+	-

Table 2.4: Some feature are not significant on its own, but a might be part of a combination of features.

It is clear that feature A and B are very important when it comes to predicting whether or not a person is happy or sad. When both features have the same sign, the person is happy, otherwise he is not. Combinations of features are often not found by feature selection methods as they look for correlations between a single feature and the output.

This problem does not occur for random forest though, as combinations of features are also 'tested' in the sense that a tree might split on them in different stages. Once the combination of features occurs randomly in a decision tree, the impurity will drop significantly, which will result in higher importance rankings.

Principal Component Analysis

Principal Component Analysis (PCA) is a technique to do dimension reduction. Intuitively, PCA can be seen as fitting an n-dimensional ellipsoid to the data. The Principal components are then the axes of the ellipsoid. Less variation in one direction, corresponds to a smaller axis. Removing that axis, will only remove a small fraction of the information, as there is only little variation in that direction. This is shown in Figure 2.14, where the ellipsoid covers a three dimensional features space. The ellipsoid has three axes: a,b and c. Intuitively, one can see that there is more variation (information) in the c ans b direction, while the a axis is relatively small.

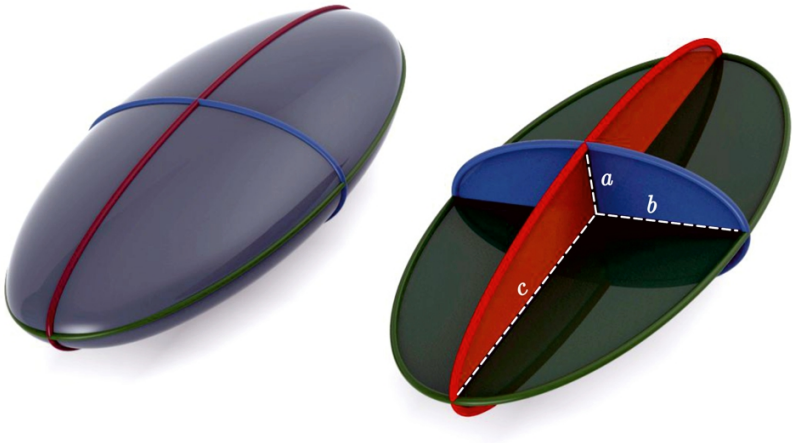

Figure 2.14: Suppose a three-dimensional feature space, where all points lie in the ellipsoid in the left.

Removing the a axis by projecting the data on the plane given by vectors b and c, will result in a two dimensional projection of the data in the form of an ellipse. This would be the black plane in Figure 2.14. This process can be repeated for higher dimensional features spaces. In other words, PCA will thus, without going into too much detail, start with an n-dimensional ellipsoid and iteratively remove the smallest axis in each iteration until the desired number of dimensions is obtained. Note that the ellipsoid should be adjusted in each step.

The major disadvantage of PCA is that the algorithm is unsupervised, meaning that it does not look at the corresponding labels of the given samples. Suppose the difference between two classes was clearly given by looking at the a axis in Figure 2.14. Applying PCA would, in that case, result in a total loss of all information.

3

Results

todo

3.1 Person specific

At first research was done in a person specific setting. This means that that algorithm was trained on several samples from a single person, before it was tested on the same person. This section will go over the results.

The first step was to transform the continuous valence and arousal values to classes. This was done by performing a simple binary classification. Given that the dimension range from 1 to 9, all labels with a valence or arousal below 5 were reported as low valence or arousal respectively. The remaining valence or arousal were placed in the high valence or arousal respectively.

3.2 Used approach

This thesis compares the aforementioned features and the aforementioned feature selection methods. For this a two stepped algorithm, inspired by the advanced random forest method, explained in Section 2.3.5, was used. In short, the first step is to rank all the features and only take the top X of the features. This threshold is applied to limit computation times and throw out features that have very low importance. The next step is to iteratively build a model by selecting features out of the remaining feature set. This approach is depicted in Figure 3.1.

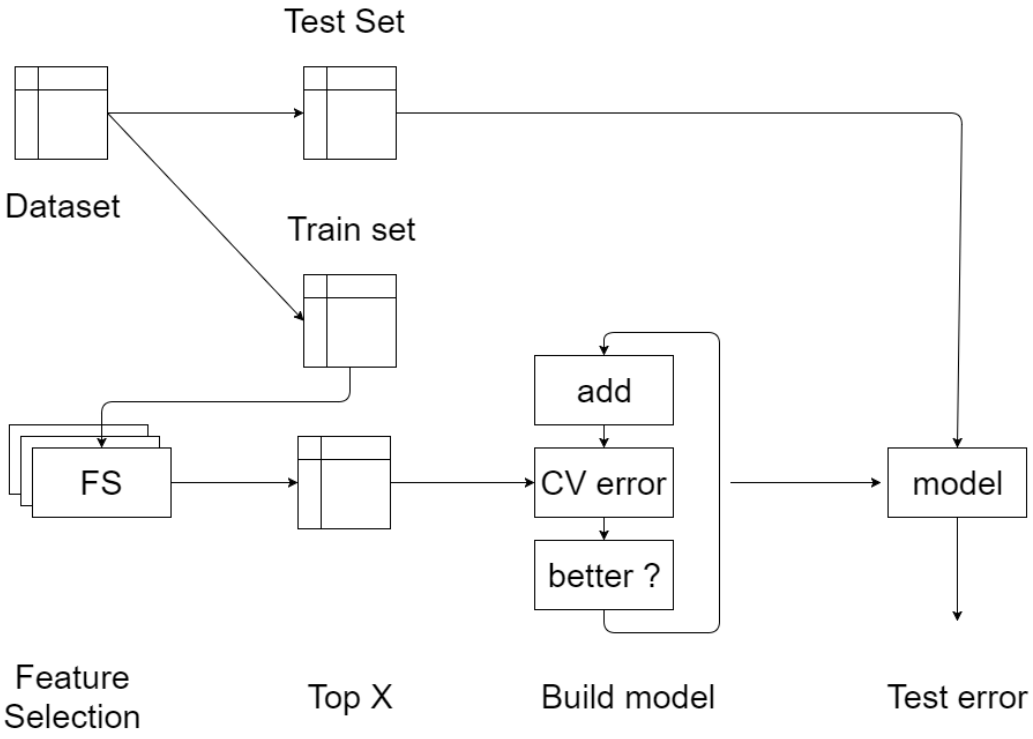

Figure 3.1: The used approach of this thesis.

As you can see in Figure 3.1, the approach starts by separating a test set to evaluate the final performance of the algorithm. This test set contained 10 of the 40 samples. Next, the aforementioned feature selection methods are applied to the train set. A top X of the features is then kept.

In the next step different models are build. This is done iteratively by starting with an empty feature set. In the add step, a feature is added to this set and the cross validation error is determined. Cross validation is a technique that separates the data in N folds, as shown in Figure 3.2. Next the algorithm is trained on N-1 blocks and tested on the remaining blocks. This is done N times and the average of the performance is then reported as cross validation error.

Figure 3.2: Cross validation

The advantage of using a cross validation scheme is that it gives a pretty good estimation of the generalisation of the algorithm, while still using all train data. This step is important because it ensure that the chosen features have good generalisation properties. Good feature should perform well on unseen samples. Note that the test set, displayed in red is not used during cross validation. The test set is kept completely separate to ensure that a fair estimate of the generalisation is achieved.

Next the average of the cross validation errors and the standard deviation is calculated. The average cross validation minus the standard deviation is b then compared to the previous best performance. If the performance is better, the feature is kept in the feature set. If the performance is not better, the feature is neglected. The standard deviation is included to increase the stability of the algorithm. By making sure that the new model performs better in a statistical way, one can avoid that small differences in averages lead to a different model.

In the final step the performance of the test set is determined by the accuracy metric. Accuracy is chosen as metric, because this metric gives a clear and intuitive measurement of performance.

The first parameter of this flow is the threshold parameter, indicated in the figure as X. This threshold cancels features with low importance, by simply taking the best X features from the feature ranking. Assigning a high value to the threshold will increase calculation times as more features are available for the building phase. The performance of the model, will not be better, since a lot of the additional features will have low importance values. Setting a low threshold is also not good, as this might cancel out important features.

In this work, the parameter was fixed to 30 for all feature selection methods for the following reasons. First, considering that there are 30 samples in the feature set, having 30 features is already more than enough. Note that a well-known rule of thumb is to have at least 10 times more samples than features[47, 48][1]. Second, looking at the features that were selected during training, one can see that usually around 5-7 features remain. The last selected feature usually has a rank around 20, meaning that the last 10 available features in the building phase are rarely used.

A second parameter of this model is a model to estimate the performance. For this, two different models were compared. The first model is an SVM with a radial basis functions kernel. This model was chose because it has proven itself in multiple emotion recognition studies. Additionally, SVM are capable to handle small dataset, which gives this method an advantage in this experiment. The next model is a random forest with 2000 estimators. This model was mainly chosen, because feature selection with random forest deliver good results, according to literature[27]. Using the same algorithm in both feature selection and building phase, means that the algorithm can select its own features. In other words features that work well for a certain algorithm are more likely to be chosen.

[1]Note that this is just a rule of thumb, and therefore not proven theoretically. In practice however, it turns out to work quite well.

4

Conclusion

todo

5

Future Research

This chapter gives an overview of research that can be done on this subject

5.1 Applications for emotion recognition

Emotion recognition has many different applications, e.g. as an improvement for brain computer interfaces or marketing analysis. A Brain Computer Interface (BCI), creates a direct link between the brain and the computer[49], that enables a subject to control the computer using only his mind. This means that physical actions like moving a mouse or typing on a keyboard are no longer needed to control a computer. A BCI is usually composed of two components. The first component is the extraction component, which extracts brain signals from the brain. The second component is a decoder that interprets signals translates them to device commands.

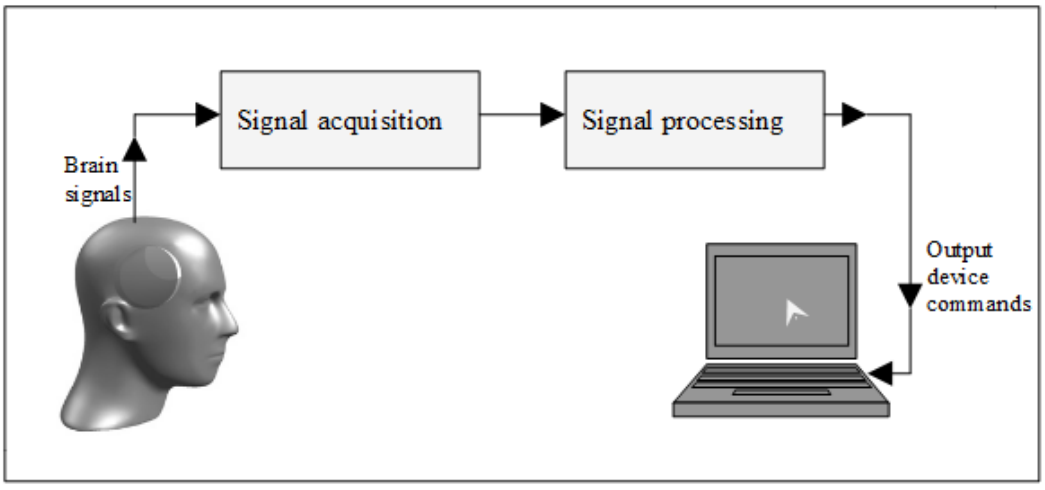

Figure 5.1: The basic components of a BCI system[50]

A very well-known BCI is the P300 speller. The P300 speller is an active topic of research. It uses EEG signals to enable patients with a locked in syndrome to communicate[51]. The basic version uses a six by six grid of characters, each row and column is flashed in a random order while the subject silently counts the number of flashes of a certain character, as shown in figure

5.2. This procedure, where a train of stimuli with some infrequent occurring target stimuli is applied, is called the oddball paradigm[52]. It is known that this technique triggers an increase in the potential difference in the EEG around the parietal lobe. When a potential difference in the brain occurs as a reaction to an event, it is referred to as en event-related potential (ERP) . The P300 ERP occurs roughly 300 milliseconds after the stimulus is flashed, hence its name[53]. The presence or absence of the P300 waveform is used by the P300 speller to determine what character the subject was focusing on, which basically allows the subject to spell text.

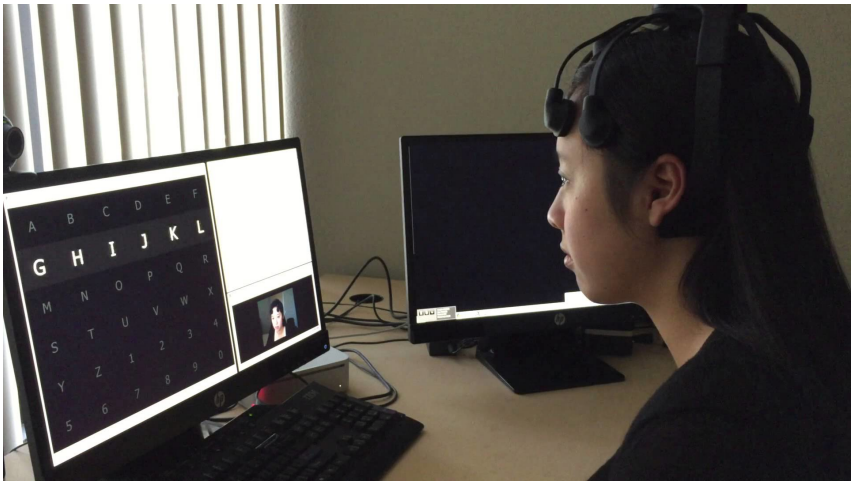

Figure 5.2: Different parts of the P300 speller, found at [54].

Research with visual stimuli on healthy subjects, has shown that emotion has an effect on the auditory P300 wave[55]. Both the P300 peak amplitude and area were highest when viewing neutral pictures and descended further, in decreasing order, for sadness, anger and pleasure. The latency of the P300 ERP speller was shortest for subjects in an emotionally neutral state. The latency increased for pleasure, anger and sadness. It is expected that a visually triggered P300 wave, will also be influenced by emotion. Having a good emotion recognition system, can help a P300 detector in finding the correct latency of the P300 wave. This can then, in turn improve the detection of P300 waves. Additionally knowing a subject's emotional state can help detecting when a subject gets frustrated, e.g. because of mistakes he makes.

An improvement in performance is not the only advantage an emotionally aware P300 speller has. Contrary to what subjects might think, the P300 speller is unable to read the mind and know what a person is thinking about[51]. The P300 speller provides no more than a means of communication that the subject can use. Should he chose to ignore the instructions and focus his attention elsewhere, then the recordings become useless. Nevertheless, ethical questions often remain unanswered. Knowing how the subject feels, can provide more insight for ethical issues, e.g. "How does the subject think about the P300 speller recording and analysing his brain activity?". Information about the subject's emotional state can help answering some of these ethical questions. Integrating the results from this thesis with the P300 speller, is an opportunity for future research.

Another application for emotion recognition is in the field of marketing and customer satisfaction research. Discovering how a person feels about a product is often tricky. Questionnaires is one way to go, but they might contain a lot of noise. Being able to 'read' the emotion straight from a subject's mind, is expected to give more accurate results as it avoids any form of social masking.

Bibliography

[1] K. Takahashi, "Remarks on emotion recognition from multi-modal bio-potential signals," vol. 3, pp. 1138–1143 Vol. 3, Dec 2004.

[2] K.-E. Ko, Hyun-Chang, and K.-B. Sim, "Emotion recognition using eeg signals with relative power values and bayesian network," *International Journal of Control, Automation, and Systems*, 2009.

[3] Y. Lio, O. Sourina, and M. K. Nguyen, "Real-time eeg based human emotion recognition and visualization," 2010.

[4] D. O. Bos, "Eeg-based emotion recognition," 2007.

[5] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1970–1973 vol.3, Oct 1996.

[6] J. Wagner, J. Kim, and e. André, "From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification," *IEEE*, 2005.

[7] P. C. Trimmer, E. S. Paul, M. T. Mendl, J. M. McNamara, and A. I. Houston, "On the evolution and optimality of mood states," *behavioral sciences*, vol. 3, pp. 501–521, August 2013.

[8] Y. Han, S. Suk Oh, J. K. Kang, and H. Park, "Simultaneous measurement of fmri and eeg - principles and applications," *intech*, May 2014.

[9] M. Garces Agustina Correa and E. Laciar Leber, "Noise removal from eeg signals in polisomnographic records applying adaptive filters in cascade," June 2011.

[10] M.-K. Kim, M. Kim, E. Oh, and S.-P. Kim, "A review on the computational methods for emotional state estimation from the human eeg," *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 573734, p. 13, 2013.

[11] T. C. Technologies, *10/20 System Positioning manual.* Fortis Tower, 2012.

[12] unknown, "Electrode placement," 2015.

[13] Brainworks, "What are brainwaves?," 2015.

[14] J. Mellinger, G. Schalk, C. Braun, H. Preiss, W. Rosenstiel, and A. Kübler, "A meg-based brain-computer interface (bci)," *Neuroimage*, July 2008.

[15] S.-g. Kim Cmrr and K. Uĝurbil, "Comparison of blood oxygenattion and cerebral blood flow effect in fmri: Estimation of relative oxygen consumption change," *Magnetic Resonance in Medicine*, vol. 38, no. 1, pp. 59–65, 1997.

[16] A. Momose, "Demonstration of phase-contrast x-ray computed tomography using an x-ray interferometer," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 352, no. 3, pp. 622 – 628, 1995.

[17] T. Castermans, D. Matthieu, G. Cheron, and T. Dutoit, "Towards effective non-invasive brain-computer interfaces dedicated to gait rehabilitation systems," *brain sciences*, vol. 4, 2014.

[18] A. Villringer, J. Planck, C. Hock, L. Schleinkofer, and U. Dirnagl, "Near infrared spectroscopy (nirs): A new tool to study hemodynamic changes during activation of brain function in human adults," *Neuroscience Letters*, vol. 154, pp. 101 – 104, 1993.

[19] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, pp. 18–31, Jan 2012.

[20] P. LanG, M. Greenwald, M. Bradley, and A. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions.," *Psychophysiology*, vol. 30, pp. 261–273, May 1993.

[21] S. U. A. NG, "Machine learning," 2016.

[22] L. Eikvil, "Optical character recognition," *citeseer. ist. psu. edu/142042. html*, 1993.

[23] Konstantin, "A brief history of intelligence," 2011.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[25] X. Zhu, Z. Ghahramani, J. Lafferty, *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," vol. 3, pp. 912–919, 2003.

[26] stackexchange, "When is a model underfitted," 2016.

[27] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, pp. 2225–2236, 2010.

[28] D. O. Bos, "Eeg-based emotion recognition the influence of visual and auditory stimuli."

[29] Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 1798–1806, July 2010.

[30] L. Brown, B. Grundlehner, and J. Penders, "Towards wireless emotional valence detection from eeg," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2188–2191, Aug 2011.

[31] M. Murugappan, "Human emotion classification using wavelet transform and knn," in *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on*, vol. 1, pp. 148–153, June 2011.

[32] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *CoRR*, vol. abs/1601.02197, 2016.

[33] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *IEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[34] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," January 1994.

[35] R. N. Duan, J. Y. Zhu, and B. L. Lu, "Differential entropy feature for eeg-based emotion classification," pp. 81–84, Nov 2013.

[36] Y. Lio and O. Sourina, "Eeg databases for emotion recognition," *International Conference on Cyberworlds*, 2013.

[37] Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, "Fusion of electroencephalographic dynamics and musical," 2014.

[38] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[39] Y. Peng, S. Wang, S. Long, and L. B.-L., "Discriminative graph regularized extreme learning machine and its application to face recognition," *Neurocomputing*, vol. 149, pp. 360–353, 2015.

[40] S. L. Crawford, "Statistical primer for cariovascular research," 2006.

[41] J. P. Pluim, A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions of Medical Imaging*, vol. 22, pp. 986–1003, August 2003.

[42] G. J. Szekely and M. L. Rizzo, "Brownian distance covariance," *The annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[43] A. Coone, "A study on different preprocessing and machine learning techniques for the detection of error-potentials in brain-computer interfaces," afstudeerwerk, Ghent university, June 2011.

[44] F. Lee, R. Scherer, R. Leeb, C. Neuper, H. Bischof, and G. Pfurtscheller, "A comparative analysis of multi-class eeg classification for brain computer interface," in *Proceedings of the 10th Computer Vision Winter Workshop*, pp. 195–204, 2005.

[45] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer, "Current trends in graz brain-computer interface (bci) research," *IEEE Transactions on rehabilitation Engineering*, vol. 8, pp. 216–219, JUNE 2000.

[46] Citizennet, 2016.

[47] Stackexchange, "Application of machine learning techniques in small sample clinical studies," 2016.

[48] Stackexchange, "Number of features vs. number of observations," 2016.

[49] H. Verschore, "A brain-computer interface combined with a language model: the requirements and benefits of a p300 speller," afstudeerwerk, Ghent University, June 2012.

[50] A. Sangari, "What is a brain computer infterface," 2016.

[51] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and clinical Neurophysiology*, vol. 70, no. 70, pp. 510–523, 1988.

[52] T. Verhoeven, "Brain-computer interfaces with machine learning: an improved paradigm for the p300 speller," afstudeerwerk, Ghent University, June 2013.

[53] N. V. Manyakov, N. Chumerin, A. Combaz, and M. M. V. Hulle, "Comparison of classification methods for p300 brain computer interface on diabled subjects," *Computational intelligence and neuroscience*, 2011.

[54] Cognionics, "Cognionics dry eeg p300 speller demo," 2015.

[55] Y. Morita, K. Morita, M. Yamamoto, Y. Waseda, and H. Maeda, "Effects of facial affect recognition on the auditory {P300} in healthy subjects," *Neuroscience Research*, vol. 41, no. 1, pp. 89 – 95, 2001.