Physiological signal-based emotion recognition from wearable devices

Master of Science Thesis
University of Turku
Department of Computing
Health Technology
February 2023
Tiina Nokelainen

Supervisors:
PhD Antti Airola
M.Sc Ismail Elnaggar

UNIVERSITY OF TURKU
Department of Computing

TIINA NOKELAINEN: Physiological signal-based emotion recognition from wearable devices

Master of Science Thesis, 51 p.
Health Technology
February 2023

The interest in computers recognizing human emotions has been increasing recently. Many studies have been done about recognizing emotions from physical signals such as facial expressions or from written text with good results. However, recognizing emotions from physiological signals such as heart rate, from wearable devices without physical signals have been challenging. Some studies have given good, or at least promising results. The challenge for emotion recognition is to understand how human body actually reacts to different emotional triggers and to find a common factors among people.

The aim of this study is to find out whether it is possible to accurately recognize human emotions and stress from physiological signals using supervised machine learning. Further, we consider the question what type of biosignals are most informative for making such predictions. The performance of Support Vector Machines and Random Forest classifiers are experimentally evaluated on the task of separating stress and no-stress signals from three different biosignals: ECG, PPG and EDA. The challenges with these biosginals from acquiring them to pre-processing the signals are addressed and their connection to emotional experience is discussed. In addition, the challenges and problems on experimental setups used in previous studies are addressed and especially the usability problems of the dataset.

The models implemented in this thesis were not able to accurately classify emotions using supervised machine learning from the dataset used. The models did not perform remarkably better than just randomly choosing labels. PPG signal however performed slightly better than ECG or EDA for stress detection.

Keywords: affective computing, biosignal, machine learning, ECG, EDA, PPG

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Research questions 2

1.3 Thesis structure 3

2 Background **4**

2.1 Challenges in biosignal based emotion recognition 5

2.2 Automatic Nervous System 6

2.3 Biosignals in emotion recognition 7

2.3.1 Electroencephalogram 7

2.3.2 Electrocardiogram 10

2.3.3 Electrodermal activity 13

2.3.4 Photoplethysmogram 16

2.4 Emotion models 18

2.4.1 Discrete emotional model 19

2.4.2 Affective dimensional model 19

2.5 Machine learning approaches in emotion recognition 20

2.5.1 Supervised learning approaches 20

2.5.2 Unsupervised learning approaches 21

2.5.3 Feature extraction from biosignals 21

3 Related work **23**

3.1 Datasets 23

3.2 Experiment setups 25

3.3 Supervised 26

3.4 Deep learning 26

3.5 Unsupervised 27

3.6 Feature Extraction 27

3.7 Validation methods 29

3.8 Subject dependency 29

3.9 Summary of related work 30

4 Materials and methods **33**

4.1 Data - CLAS 33

4.1.1 Data collection setup 33

4.1.2 Data quality 34

4.2 Methodology 37

4.2.1 Pre-processing of biosignals 37

4.2.2 Feature extraction 39

4.2.3 Feature selection 39

4.2.4 Classifiers 41

5 Experiment **43**

5.1 Training supervised classifiers 43

5.2 Results 44

5.2.1 Performance metrics 44

5.2.2 Performance results of the experiment models 45

6 Conclusion **49**

References **51**

1 Introduction

The interaction between humans and computers is a vastly researched field, and the detection of emotions has been increasingly a target of interest in the past decade. Emotion recognition from the written text, for example, from social media posts, has been a known research target for a while now. However, emotions are still perhaps the least researched field in human-computer interaction.

Computers are perceived as logical and rational machines while human emotions are more individually and illogically, indescribably experienced states of mind. Training emotions to a computer is not as trivial as, for example, the grammar of a language [1]. The less researched field of emotion recognition is detecting emotions with physiological signals using biosensors. As smartwatches and smart rings have become more common and biosensors have evolved to be used on more mobile and wearable devices, the utilization of biosignals is more present. Smartwatches can detect sleep stages, human activity, and even stress levels by using biosignals from the human body.

1.1 Motivation

Mental health has become an essential topic of discussion in modern society. Nowadays, you can switch lights on and off through your smart device and even change

the tone of the lights to bright red or green. But what if your smart device could detect your mood and change the hue of the lights automatically or the genre of music to listen to according to the way you are feeling? When feeling anxious or stressed, certain hues of color on lights and genres of music may calm the human mind and body. Detection of emotions could help detect changes in the mood even before human themself can.

Emotions are a central part of communication among people; thus, it is - or could be - an essential part of human-computer interaction. Emotions have been detected from speech and facial expressions; however, there are some challenges with using cameras or microphones for emotion detection. To detect facial expression, a camera is needed, and it is not very practical for everyday use, and finding an optimal lightning setup causes more challenges. Likewise, speech recognition raises some difficulties since it is not always desired or possible to speak out to your smart device, for example, during a movie.

The difficulties in using speech and facial recognition raise interest in developing alternative systems and technologies for detecting emotions. Biosignals measured directly from the human body have shown to be a potential and good alternative for detecting emotions. In addition, the increased development of wearable devices provides even more possibilities for biosignals in machine learning models. Acquisition of biosignals is much easier nowadays and can be even more accessible in the future. Biosensors can be found in jewelry, clocks, or even in clothes.

1.2 Research questions

This thesis aims to fulfill the following research questions:

RQ1: Is it possible to recognize different emotions from phsysiological signals using supervised machine learning?

RQ2: What is the best suited supervised machine learning algorithm for making the most accurate predictions for emotion recognition?

RQ3: What are the best physiological signals for emotion recognition?

1.3 Thesis structure

The basic idea behind emotions, biosignals, and emotion recognition is covered in chapter 2. In chapter 3, earlier studies in the emotion recognition field are described and compared. Most common methods are brought to attention and examined. The dataset used in this thesis is covered in chapter 4, along with methods implemented in the experimental part of the thesis. Chapter 5 goes into more depth with the designed model with some visualizations and the results. The results and experiment results are discussed in more detail in chapter 6 with conclusions.

2 Background

Human emotion plays a vital role in human interaction between human to human. Therefore, it is not unheard of that emotions also play a significant role in human-computer interaction. The use of physiological signals has been rising when the bio-sensors have been available in wearable, off-the-shelf devices. Modern wearable devices can detect, for example, sleep stages or stress levels from the person wearing the device. Sleep detection presents exciting data for the subject about their quality of sleep, and they can change their daily routines if needed to get their quality of sleep higher. Similarly, the device detecting stress level can notify the user if their stress levels are getting abnormally high to react to it as soon as possible. Commonly used biosignals in emotion recognition include:

- EEG from the brain
- ECG from the heart
- EDA sweat signal
- The blood volume PPG signal

Different biosignals are described in more detail in chapter 2.3.

2.1 Challenges in biosignal based emotion recognition

What makes emotion recognition from physiological signals difficult is that emotions are not measured in any particular dimension. There is not an objective perspective of someone's emotion their feeling. The only known way to define an emotion is by describing their feelings aloud. Indeed some emotions are easy to recognize by examining the expression of others' faces or from the tone of their voice. However, there is no universal way to express feelings in any numeric or categorical measure, which is necessary for machine learning. Every individual experiences emotions differently, and the human body works differently between individuals. One might feel afraid on the top of a building, and another can feel excited or happy to be in an exciting and dangerous place.

Off-the-shelf wearable devices that record biosignals have not been on the market for an extended amount of time. Previously researchers had to rely on professional, clinically used electronic devices. Recording a biosignal is quite different when the subject sits quietly unmoved in supervised lab conditions as opposed to actually moving i.e. "in the wild". Biosignals are more prone to artifacts when the subject is moving. There are studies about emotion recognition from subjects that are moving. Kanjo et al. [2] used location data also in addition to biosignals from the subjects. Their model performed better with the environmental data than using only the biosignals which implifies there is a correlation between environment and emotions.

Because the human body can react unintentionally to different kinds of triggers, external or internal, there is no certainty that a change in the signal is a product of emotional experience. This means that while analyzing and processing biosignals, it

needs to be considered that not all reactions are desirable and presumed outcomes. For example, suppose a reaction is expected from some emotional trigger e.g., while watching a video of someone jumping from a plane. In that case, it can affect sweating and risen heart rate, but it can also result from pain like an acute stomach ache or a headache [3].

From a machine learning perspective, biosignals can be very informative signals and reliable, but they are prone to artifacts to a considerable extent. Biosignals (non-invasive) are very sensitive to any motion or other distraction. Though they can be filtered and cleaned from any noise and artifacts, the signal cannot be saved every time. While biosignals gather information about emotional response, distinguishing negative and positive experiences i.e. arousal from each other, is not straightforward.

2.2 Automatic Nervous System

The autonomic nervous system (ANS), in other words, the involuntary response system, plays a significant role in emotional experience. It regulates the smooth muscles, the secretion glands of internal organs, and cardiac muscles. The autonomic nervous system is divided into the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS). The rhythm of the heart is controlled by the medulla oblongata, which locates in the brain and is part of ANS.

Emotional stimuli affect the autonomic nervous system and its activity. For example, the sympathetic nervous system is connected to sweat glands that react to external stimuli, e.g., temperature changes or an emotional response. It is known to be described as *"the flight or fight response"*. The parasympathetic nervous system is more known to stimulate the *"rest and digest"* states. For example, it controls the

constriction of pupils and airways and the heartbeats when the pulse is low [4].

2.3 Biosignals in emotion recognition

Biosignal-based emotion recognition is an intricate task; thus, selecting suitable physiological signals for own research is the key element. Although emotions are reasonably easy to recognize from facial expressions using EMG or even from the brains using EEG, they are not entirely practical signals since they both require several electrodes to be attached to the subject. Attaching electrodes to the subject is not always the preferable or even possible solution therefore wearable devices come in handy. Photoplethysmogram, skin conductance response, and skin temperature are often used as signal resources for emotion recognition because they are often implemented in wearable devices. Some wearable devices can even collect some ECG signal, which is valuable in emotional experience [5]. The selected signals should reflect the activity of the autonomic nervous system [4]. The signals used in this thesis experiment are electrocardiogram (ECG), electrodermal activity (EDA), and photoplethysmogram (PPG). The most used biosignal in emotion recognition is electroencephalography (EEG), which measures the brain's electrical signals.

2.3.1 Electroencephalogram

Electroencephalogram (EEG) measures the electrical activity of the brain. The brain's electrical activity is a product of the current of ions within the electrically charged neurons in the brain. The activity's rhythmic and periodic patterns of brainwaves are formed, captured as EEG. EEG is commonly used to detect epilepsy or brain damage. Since emotions are seen to be a product of psychological experience e.g. feeling happy or stressed, EEG is widely used in emotion detection experiments.

Acquisition

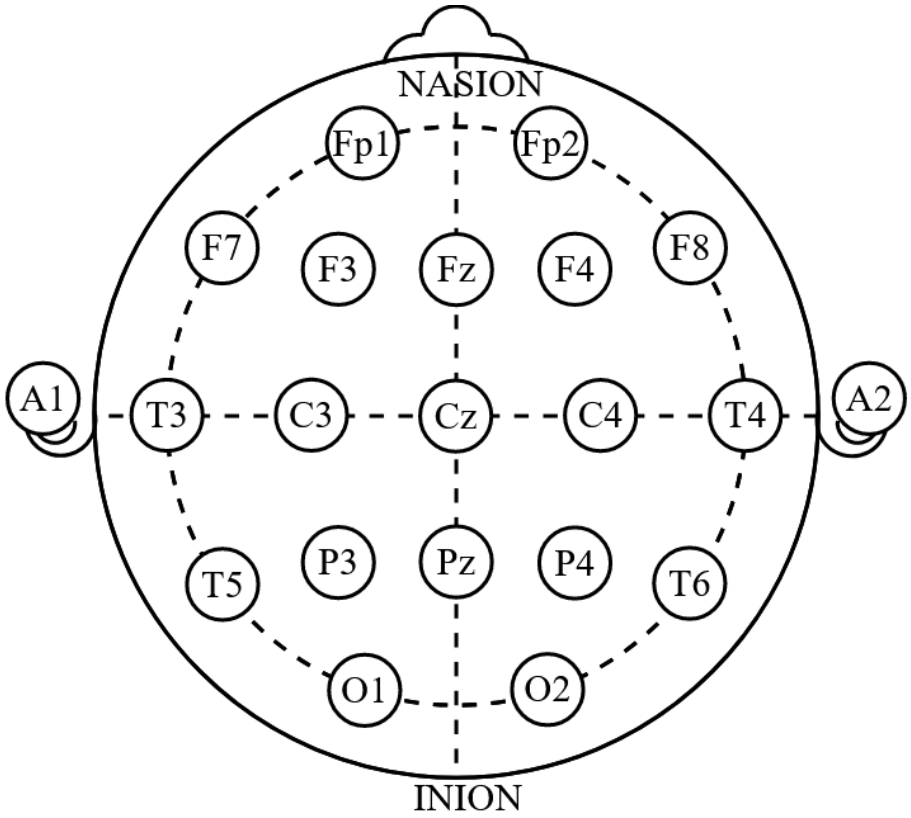

Figure 2.1: EEG electrode placement using 10-20 system. Each circle represents an electrode. [6]

EEG is measured by placing several electrodes on the skin of the scalp. Electrodes have a conductive gel or paste on them. Mostly 19 electrodes are used to collect EEG signals. The International 10-20 system is used to place the electrodes in a standard pattern on the scalp, as seen in figure 2.1. The signal is divided into different frequency bands, which are the main interest of EEG. Frequency bands are listed in the table 2.1. Different frequencies are present during a different state of the human mind. Delta brainwaves are active when a human is asleep, and theta brainwaves represent being alert. In a relaxed state, alpha brain waves are activated. When brains are focused, beta brainwaves are present in EEG. Brainwaves with gamma frequencies are known to be multi-processing states of the brain. Mu brainwaves stop

Table 2.1: EEG bands

Band	Frequency (Hz)	Location
Delta	< 4	Frontally
Theta	$4 - 7$	-
Alpha	$8 - 15$	Posterior regions of head
Beta	$16 - 31$	Both sides
Gamma	> 32	Somatosensory cortex
Mu	$8 - 12$	Sensimotor cortex

idling when the human body moves a significant part of their body. [7] Sampling frequency of EEG should be at least 100 Hz, but usually, a higher sampling rate is used, typically 300 Hz.

Common artifacts

Common artifacts in EEG signals are caused by eye movement, swallowing, and poor electrode contact.

Challenges

Recording EEG signals is not very practical since many electrodes are attached all around the head. Some consumer-level devices have been present in the market recently, but their signal quality does not necessarily reach the required or desired level. EEG being already vulnerable to artifacts and having relatively low signal quality, measuring good quality signals in day-to-day life is challenging. Extracting robust features from EEG can also cause some difficulties. [8]

2.3.2 Electrocardiogram

Electrocardiography (ECG) is a non-invasive method to measure the muscle activity of the heart. It is commonly used to detect various arrhythmias i.e. abnormal heart rhythm, and it is an essential tool used in hospitals. It is crucial to detect cardiac arrhythmia as early as possible since some arrhythmias can be fatal if they are not treated without delay. Emotions have an impact on heart activity as well. When a human gets excited, it is expected that their heart rate might increase, and when the body feels relaxed, the heart rate decreases.

ECG provides information about many things besides arrhythmias. Heart rate is possible to extract from ECG after finding R-peaks (specified in 2.2) and calculating the heart rate from it. In addition, respiration rate is possible to extract from ECG. The respiration signal is usually filtered out from ECG signal by eliminating low frequencies e.g. with a high-pass filter. However, it can provide important information about respiration if needed.

Components of ECG signal

One cycle of a heartbeat in ECG signal consists of a few different components as seen in figure 2.2: P-wave, QRS-complex, and T-wave from which the most crucial component is the QRS-complex. All these components represent a particular stage in one heartbeat cycle. The P-wave represents atrial depolarization, the QRS-complex ventricular depolarization, and the T-wave ventricular repolarization.

Acquisition

In lab conditions (hospital, clinics) ECG is measured with ten electrodes attached to specific body parts. From 10 electrodes, 12 ECG leads are formed as listed in the table 2.2. Each lead represents the electrical activity of the heart from a certain

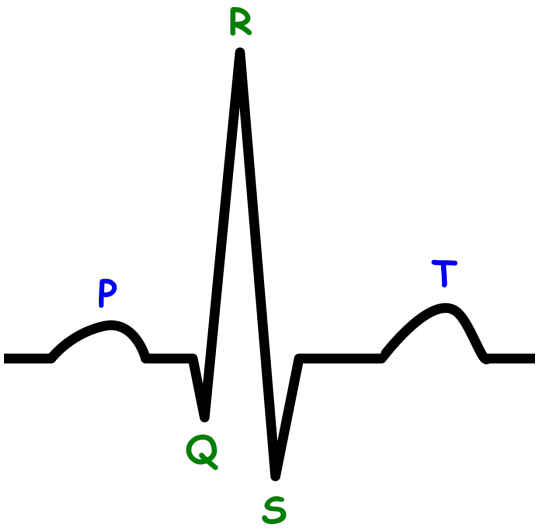

Figure 2.2: Illustration of one heart beat cycle and ECG's different components.

angle. Most information is gathered with 12 leads; however, ECG signal is possible to measure with only one lead which gives the least information because of the signal represents only one angle of the heart. In clinical settings regular sampling rate can be up to 1000 Hz; however, 500 Hz is commonly used in a wearable device or even lower frequency. It is not recommended to use a much lower sampling rate than 100 Hz [9]. However, heart rate variables have been successfully extracted from ECG signals with a sampling rate equal to as low as 50 Hz. [10]

Common artifacts

Common artifacts or noise in ECG signals are typically muscle artifacts, contact noise, power line interference (50/60 Hz), baseline wanderer, and noise from the data collecting device [13]. The preceding three noises are easily filtered out from the signal using low-pass and high-pass filters.

Challenges

Noise is the greatest challenge when processing ECG signals. Detecting the R-peaks can be difficult with much noise due to e.g. poor electrode attachment or movement

Table 2.2: 12 leads from 10 electrode/sensors [12]

Lead	Negative electrode	Positive electrode	Angle of heart
Lead I	RA	LA	Lateral
Lead II	RA	LL	Inferior
Lead III	LA	LL	Inferior
aVR	LA + LL	RA	*None*
aVL	RA + LL	LA	Lateral
aVF	RA + LA	LL	Inferior
V1			Septal
V2			Septal
V3			Anterior
V4			Anterior
V5			Lateral
V6			Lateral

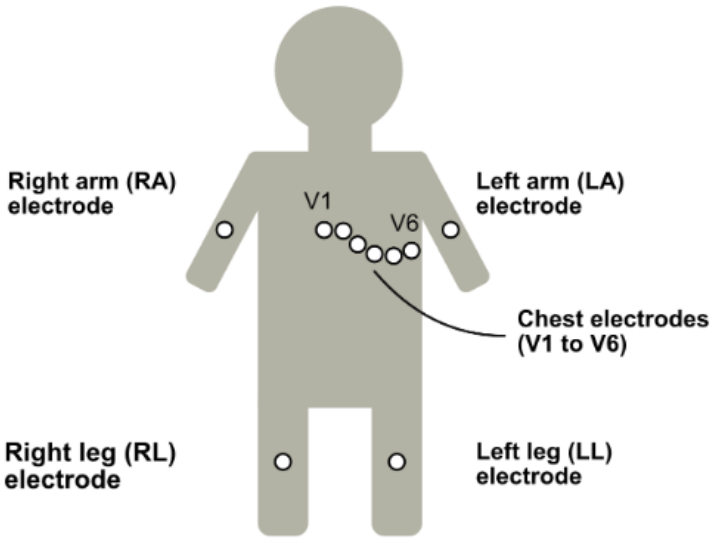

Figure 2.3: Placement of ECG electrodes [11]

of the subject (e.g. coughing). Noise can be filtered out with high pass and low pass filters and with a set of rules e.g. how close two R-peaks can physically be with each other. However, filtering should be carefully executed to avoid losing any desired information.

2.3.3 Electrodermal activity

Electrodermal activity measures the changes in the electrical properties of the human skin - the so-called *"sweat signal"*. When human skin starts to produce sweat even mildly, the skin conductance level rises since sweat, which is mostly water, conducts electricity [4]. EDA is found to be a good and sensitive signal for emotion recognition for its close relationship with ANS, and it is known to be the key signal in lie detectors for that reason. Because EDA is one of the oldest used and researched biosignals and is reasonably easy to measure, it is implemented in many wearable devices.

Acquisition

Electrodermal activity is easy to measure from the human body. It requires two Ag/AgCl electrodes to be placed on the skin's surface. Usually, electrodes are placed on the fingertips. In addition, EDA can be measured from the palm, wrist, or even from the foot. When the subject is sweating even a little, the conductance of the skin changes. Sweat glances produce more sweat which makes the resistance of skin drop, and on the contrary, the skin's conductance rises [14]. Skin conductance is linear to the activity of sweat glands - the more sweat glands push sweat to the surface of the skin, the higher the skin conductance rises. The skin conductance is measured in siemens units (S), more specifically in micro siemens (μS). Siemens is used as a unit of electric conductance. The usual sampling rate for a good quality signal is 200 to 400 Hz.

Different units of EDA

The EDA signal is usually split into different components: tonic skin conductance level (SCL) and phasic skin conductance response (SCR). SCL represents the slow changes in EDA signal - the tonic levels. SCR component represents the fast-changing signal in skin conductance. SCR can tell the rapid reaction of the stimulus, which is usually 1-5 seconds after the stimulus. After a stimulus, the SCR level changes rapidly and forms peaks. Therefore SCR is an excellent signal when focusing on the instant reactions to different stimuli. The tonic level of the signal changes more slowly, representing the overall skin conductance level without the SCR peaks giving information about the subject's emotional or physical state.

Common artifacts

Common artifacts in EDA signal are due to poor sensor contact or motion artifact such as subject tapping on the sensors. A poor sensor contact or a dry electrode can

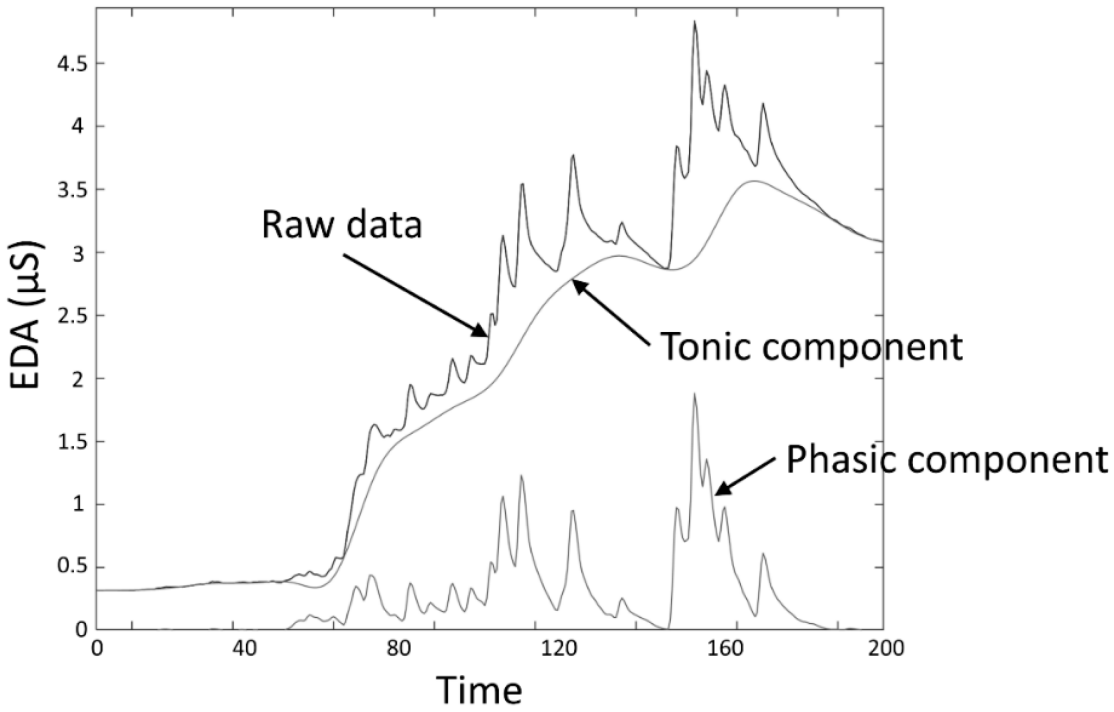

Figure 2.4: Different components in EDA signal [15]

cause artifacts to the signal, such as an unusually low signal, or they can exasperate the electrical noise (50 Hz/60 Hz), which is not usually a problem in EDA signal since the EDA signal is low-pass filtered. Usually, artifacts in EDA signals are unusual rises or drops in the signal level that are not physiologically possible. SCR peaks can reach their highest point in 1-3 seconds, and the SCR signal can lower 50% in 2-10 seconds. If there is a rapid change in the EDA signal, let us say 2 μS drop in less than two seconds, it is impossible to be from a normal physiological cause. [16] [17]

Challenges

The way a body produces sweat differs significantly between individuals. EDA is the most accurate when the conductance is over 0.5 μS. However, due to low production of sweat or even too high air conditioning or cold environment can lead to EDA signal under 0.5 μS [16]. The signal under 0.5 μS is most likely useless since different EDA

components are nearly impossible to distinguish. Artifacts from loose electrodes cannot be recovered during signal processing. The best practice is to ensure that the electrodes are correctly attached to the skin's surface and that the electrodes do not move during the recording. Also, it should be considered that EDA responses occur with a bit of delay after a stimulus, usually after 1-2 seconds.

2.3.4 Photoplethysmogram

Photoplethysmogram (PPG) is a non-invasive optical measurement technique to detect cardio-vascular pulse waves, usually from a fingertip using a light source and a detector. The light source sends infrared, which is low-intensity light, or/and green light, through the finger. Then the detector measures the amount of backscattered infrared, which corresponds with the blood volume variation [18]. Blood volume is the volume of blood cells (red cells, plasma) in the blood. The blood volume changes as wave-like pulses when the heart contracts blood to the vessels all the way to the fingertips. The blood volume rises when the heart contracts since more blood cells travel through vessels; hence, less light backscatters to the detector. An example of a PPG signal is seen in figure 2.5. The high point of the signal refers to the contraction of the heart - the systolic peak. The second peak, which is certainly lower, refers to the relaxation of the heart when the heart is filled with fresh blood - a diastolic peak. From PPG, there is a possibility to extract a few different signals. PPG can detect pulse rate, respiration rate, and even oxygen saturation.

Acquisition

Commonly PPG is measured by attaching a pulse oximeter around the subject's fingertip as seen in figure 2.6 LED being the light source and PD the detector. There are other ways to collect PPG data, such as attaching the oximeter to the earlobe, which is how this thesis' PPG data is collected. There are two typical ways

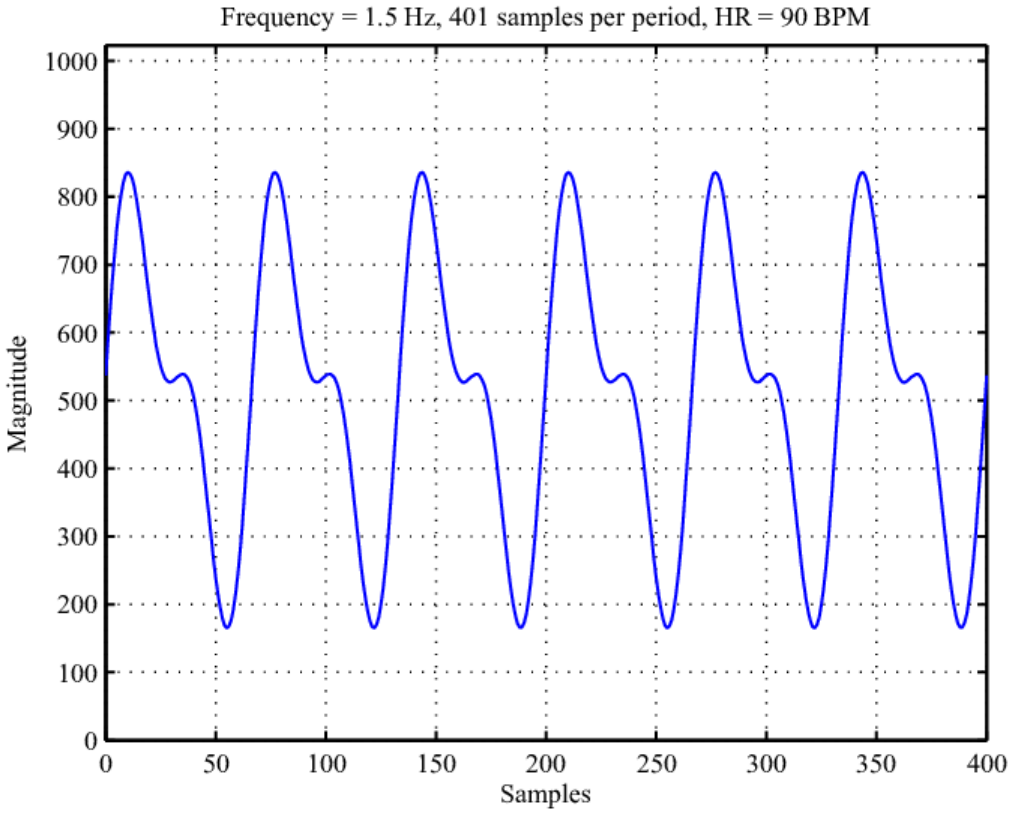

Figure 2.5: Illustration of a pure PPG signal with no noise or artifacts. [19]

of setting up the light source and detector. The source and the detector can be next to each other when the detector measures how much light backscatters to the detector. The other way is for the source and detector to be placed on opposite sides of the finger when the receiver detects the amount of light coming through the finger. In addition, PPG can also be measured from the wrist as modern smart clocks do. In both, they have the same idea of having the light source and the light receiver. PPG is usually obtained using a 125-1000 Hz sampling rate. [19]

Common artifacts

Common artifacts are muscle and motion artifacts and respiration rate, though respiration rate is sometimes a wanted outcome from the PPG signal. When the light source and light detector are side-by-side, the light source can cause some interfer-

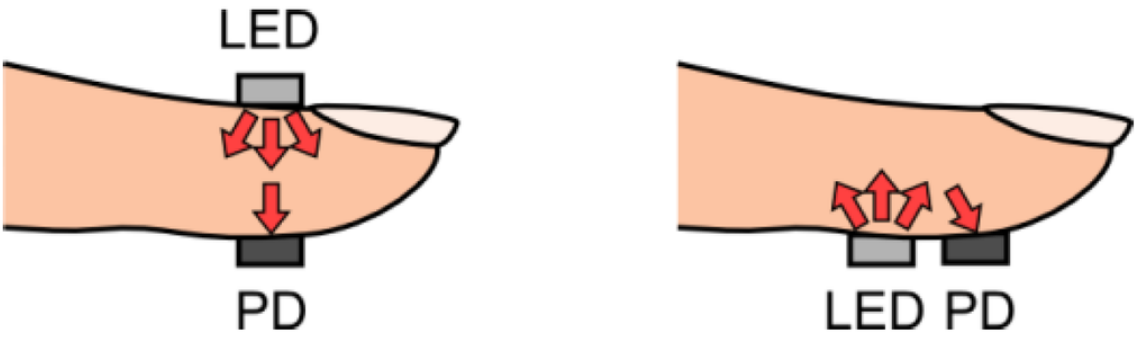

Figure 2.6: Two main setup style of PPG sensors with LED light source and a receiver. [20]

ence to the signal. However, it can be maintained by proper probe attachment and filtering. [21]

Challenges

PPG being based on optical recording method skin tone can have some affect on the signal; different skin tones absorb light differently.

2.4 Emotion models

There are many approaches to measuring emotions. The approaches can be categorized into two categories by the emotional model used. The two categories broadly used are discrete emotional models (DEM) and affective dimensional models (ADM) [22]. Both models require the subject to report what they are feeling and even how strongly they are feeling them. There is a well-used reporting tool as Self Assessment Manikin (SAM) [23]. Some studies even use both emotion models since they are in a way linked to each other, as shown in figure 2.7.

2.4.1 Discrete emotional model

In discrete emotional models, the subject recognizes different emotional states and reports them, such as feeling joy or disgust. Common wanted emotions are anger, disgust, happiness, fear, sadness, and surprise. These emotions are said to be the six basic emotions. However, humans may interpret their emotions differently and their ability to recognize their own emotions according to the chosen emotions is questionable.

2.4.2 Affective dimensional model

In the affective dimensional model, we look at emotions through two different parameters: valence and arousal. Arousal measures the intensity of emotional stimulation on a scale of low to high. Valence measures the pleasantness of the emotional experience on a scale from very pleasant to very unpleasant (high to low). Different combinations of arousal and valence often imply a specific type of emotion, as shown in figure 2.7.

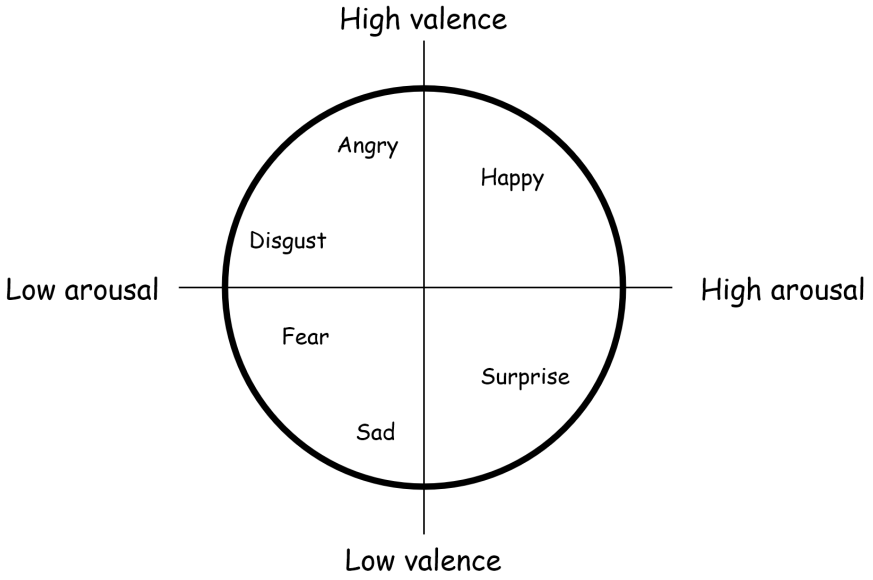

Figure 2.7: Emotions in affective dimensional model coordinator [24]

2.5 Machine learning approaches in emotion recognition

Machine learning is a field of computer science where a computer is programmed to operate without its response to input being accurately programmed. This means that the computer learns to operate independently with the help of features learned from the data. The computer builds a model from the sample feeds from which it learns to make decisions or predictions. In data analysis, machine learning is used to find complex models in the data that can be used to make predictions. Researchers can make reliable decisions and results from the data used based on these predictions. Machine learning methods are often divided into supervised learning and unsupervised learning. Supervised learning is a more commonly used approach in emotion recognition models, which might be result of supervised learning being broadly used in many machine learning problems, and most machine learning algorithms are supervised [25]. However, unsupervised learning models have given promising, or at least acceptable results in emotion recognition also.

2.5.1 Supervised learning approaches

In supervised learning, the training data is known in advance. The task is to determine the output y for the input x using the known training data. The training data consists of pairs where the outputs Y are defined for the points X. The supervised model can make a decision - prediction - for the unknown data input x based on the training data. How the model makes this decision depends on the different supervised algorithm approaches. Supervised models aim to reduce the error rate of the model. The error rate is calculated by comparing the predictions made to the training data and the actual outputs. The most popular supervised learning approach in emotion recognition is Support Vector Machines (SVM) [26].

Deep learning methods have raised interest lately, and for instance, simple deep learning method Convolutional Neural Network (CNN) has performed excellently in few emotion recognition studies [2].

2.5.2 Unsupervised learning approaches

In unsupervised learning, only the inputs x_i are known from the training data, but the outputs y_i are unknown. In such cases, the task of the computer is usually to cluster the data, i.e. to group the same type of data points into their own clusters. Therefore, the purpose is to find interesting characters in the data, like shapes. This is much less well-defined than supervised learning, as the characters you are looking for are unknown in advance. Unsupervised learning methods also do not have an exact error measure, as there is no training data to compare with the prediction made. Unsupervised learning can be used for marketing by targeting customers with similar attributes. In this situation, an unsupervised learning method identifies or groups the same type of customers into their own groups, which allows you to target your advertising to specific customer groups better. Attributes can include tracking social media accounts, customer age, and gender [27] [25]. Unsupervised methods are rarely used in emotion recognition systems particularly based on physiological signals. In emotion recognition, an unsupervised learning method would divide data points into several clusters. Each cluster would represent a certain emotion.

2.5.3 Feature extraction from biosignals

A feature extraction takes place before the machine learning model can be trained. When having a large dataset of the raw signal, processing it can be pretty overwhelming. Feature extraction helps to reduce data redundancy and speed up the machine learning process since the number of data decreases while still keeping the vital information. Finding the most informative and appropriate features is essen-

tial and can have a huge impact on the model's performance. The most common features extracted from biosignals are statistical and frequency domain features. Statistical features include e.g. the number of peaks in the EDA signal or the mean of heart rate from ECG signals. Frequency domain features can include the power of a particular frequency band e.g. power in 0-0.04 Hz or the prominent frequency in the signal.[1][16]

After feature extraction, we are left with a dataset that can have dozens of features per entity or even hundreds. Training a model with multidimensional data may lead to overfitting it or losing informative data by drowning the crucial features. A good practice is choosing a subselection of features to train the model. This is called feature selection [28]. There are many different techniques and approaches to feature selection, yet they all have the same goal to find the best features for the model. Some techniques even combine different features to reduce dimensionality while still describing the original data and keeping valid information such as Principal Component Analysis (PSA).

3 Related work

Physical signals such as facial expression and speech in emotion recognition have been much more studied for years than physiological signals. The oldest studies with physiological signals are, however, from the 1950s. Most of the oldest studies with more than one signal usually consist of a combination of physical and physiological signals [29]. Emotion recognition with only physiological signals has become more common in the 2010s; this might be due to having wearable devices available for consumers.

This thesis is more focused on physiological signals that are reasonably easy to measure from the human body. Some studies included here have used physical signal EMG, which measures the muscle activity, usually the muscle activity of facial expressions.

3.1 Datasets

Reacting to external stimuli with physiological emotions is remarkably dependent on the subject since everyone reacts to different events differently. Likewise, human bodies react physiologically differently to different emotions. Age, gender, and mental health have been seen to have a considerably significant effect on different physiological signals.

In K. H. Kim et al. [4] their dataset contains biosignals from 50 individual children aged from seven to eight years. Their initial target was younger subjects; however, younger children had difficulties inducing emotions and reporting them. Many other studies used their own acquired data too as [30] [31] [32] [33]. Some studies used ready-made datasets such as AMIGOS or DEAP. Several databases are openly reachable via the Internet, like the database used in this thesis. These publicly available datasets consist of data recorded from 23-32 subjects. Most of the studies have used their own data. Two studies used the same dataset, consisting of data from only one male subject recorded over 25 days, with four emotion recordings per day [34][35]. Wagner et al. also compared their model with another dataset from MIT Media Lab, achieving similar accuracies for valence (81-88 %) and slightly lower accuracies for arousal (87.50 % versus 96.59 %). The MIT Media Lab dataset also contains data from only one subject. Das et al. [33] used their own dataset with only four medically healthy subjects suggesting that their method achieves close to 100 % with using only EDA signal. These high results might mean overfitting of the model thus the model wouldn't perform well with new unseen data.

Since biosignals depend highly on the subject, many features affect the signal output. Medical conditions, age, even gender can impact the signal. It is good to review the subjects used in the studies. Many datasets consist demographically similar subjects. Relatively young people (18-35 years old) were desirable subjects in many - if not all - studies since the quality of the signal is known to be better from a younger person. [36] [30] [33] [31] studies even wanted to make sure that their subject were overall healthy, not on any medication. [31] even made sure that their subjects did not have any history of psychological or neurological conditions. [32] had 101 volunteers that were all first-year students who are presumably 18 years old. The gender distribution was quite even in all studies which expressed the distribution except in [2] they had data recorded from all female subjects.

3.2 Experiment setups

There is not only one specific way to trigger emotions and collect data. Many studies have found videos and pictures to induce emotions successfully, and the materials are easy to label even beforehand, though it is not desirable to assume the emotions experienced. Self-assessment reports are a common way to help label the data properly. Usually, the emotion model used for self-assessment reports and labeling is the ADM with valence and arousal, though DEM is also broadly used.

In Dar et al. [24], their model performed distinctly better on AMIGOS dataset than on the DREAMER dataset. They suggest that this is due to more accurate self-assessment reporting. In AMIGOS, subjects reported emotions on a scale from 1 to 9, and in DREAMER, emotions were reported from 1 to only 5. Thus choosing the reporting type for the experiment should be carefully deliberated.

Many experiments used ready-made emotion eliciting picture, video, or audio databases such as International Affective Picture System (IAPS) or International Affective Digitized Sounds (IADSs). In these databases, every picture and sound is labeled with arousal and valence rates, and they are supposed to elicit these labeled emotions or feelings from the subjects.

Many studies noticed that emotions are easier to separate on the arousal axes than on valence. Arousal rate was over 10 % better recognized than valence rate in [35] in a scale low/high arousal and positive/negative valence. They found similar results in other literature as well.

3.3 Supervised

Many studies use a common machine learning model, Support Vector Machine (SVM), a supervised learning model. However, SVM performance varied between studies. Das et al. [33] achieved promising results on their SVM model with GSR and ECG signals, accuracy being well over 90 %. They used DEM as their emotional model with three different emotions: happy, sad, neutral. Interestingly, their study achieved 100 % accuracy with only GSR signal in all their models (SVM, NB, KNN) with statistical features. They suggest that GSR is a highly reliable source for emotion recognition. In [4] and [31] studies GSR signal wasn't one of the chosen biosignals for their models. Their best accuracy was only close to 70 % in both studies.

3.4 Deep learning

Studies show promising results in using deep learning methods to recognize different emotions. Convolutional neural network (CNN) gave good results in [24] and [2] with accuracies 98.8 % and 87.3 %. They both used the LSTM layer in their CNN model. In Dar et al. [24], ECG and GSR signal were sent through the LSTM layer (Long Short Term Memory) in 1D-CNN. EEG, being a different kind of signal than ECG and GSR, was converted to PNG images and run through 2D-CNN. Multi-modal fusion was used to decide the classes by major vote. They got good results of 99% accuracy by using ECG and EEG signals together; GSR did not help the model's performance. Kanjo et al. [2] used different biosignals: ECG, GSR, BT, and they also recorded movement of the subjects with an accelerometer. They adopted the LSTM as an additional layer in their CNN model. LSTM helped determine which information from earlier steps should be remembered for the next state and which information should be forgotten. Interesting with their study is

that they used environmental and location sensor data in their model. Using only data from biosignal sensors, they achieved an average of 87.3 % accuracy with five different output classes. They managed to improve their performance to 94.7 % by including environmental and locational data in the model. Their study also tested emotion recognition with Multi-layer perception (MLP), which CNN outperformed by 6 % of accuracy MLP having average accuracy of 79 %. MLP is a fairly simple deep learning method that still performed quite well though not matching the CNN models' results.

3.5 Unsupervised

There are close to none unsupervised systems in the literature for biosignal-based emotion recognition. Only one study found that uses the unsupervised learning method in their model. In [37], only EEG signal was used in their model of hypergraph Laplacian-based partitioning. They adopted K-means to cluster the data into the number of emotion classes. They used ADM for the emotion model with arousal, valence, dominance, and liking classes. Their results varied between 54.61 and 65.12 %. Though their results do not compare to supervised learning methods, it is interesting to have a study in the unsupervised learning field as well. In addition, the results suggest that unsupervised learning methods should be more researched in emotion recognition.

3.6 Feature Extraction

Extracting appropriate features from the biosignals for the model is an essential task in emotion recognition machine learning models, especially in supervised learning methods. Many studies use traditional statistical features in their models - time and frequency domain features. Some studies tested specific algorithms for feature

extraction.

In [36] they suggest that their HAF-HOC feature extraction for EEG signal improves classification efficiency for emotion recognition getting an accuracy rate up to 85.18 %. HOC-based analysis constructs the feature vector (FV^{HOC}) as follows.

$$FV^{HOC} = |D_1, D_2, \ldots, D_L|, 1 < L \leq J$$

where J denotes the maximum order of the estimated HOC, and L is the HOC order up which they were used to form the FV^{HOC}. HOC measures the relationship between zero-crossing rate and autocorrelation [36].

Zong et al. [34] suggest that their feature extraction technique outperforms traditional methods. Their technique is based on the Hilbert-Huang Transform (HHT). Their technique with 'fission' based features improved their model from 71% to 76%. HHT method is based on decomposing signal into IMFs (*Intrinsic Mode Functions*), and the fission approach of HHT aims to extract features from each IMF, and the feature vector used in the model is the combination of these features. Also [22] study used HHT as a feature extraction method.

Wen et al. [32] extracted features using local scaling dimension (LSD). LSD calculates how the signal fluctuates in different time scales, giving information about the strength of the fluctuations. The LSD is defined at each time scale ε as follows:

$$D_m(\varepsilon) \equiv \frac{1}{m-1} \frac{\delta log \chi_m(\varepsilon)}{\delta log \varepsilon}$$

where $\chi_m(\varepsilon)$ are the moments.

A popular method for frequency-domain features is Welch's method to estimate the power of the signal at different frequencies. At least [33] [37] mentioned using the method for power spectrum density. Another algorithm for spectrum analysis was

used in [4]: ARMAsel.

Studies using CNN, a neural network, learned features directly from the raw signal without manual feature extraction. Those studies were [24] and [2] which both had models with combination of CNN and LSTM.

3.7 Validation methods

Almost all studies used either leave-one-out cross-validation (LOOCV) or leave-one-subject-out cross-validation (LOSOCV), which is no surprise since biosignals are very subject-dependent. [24] and [2] papers did not disclose validation method other than that they trained their model with 70 % of the data and tested with the other 30 %. [24] tested their model twice on both datasets (AMIGOS & DREAMER) with randomly splitting their data into train and test sets. The accuracy of their model was calculated by the mean of both performances.

3.8 Subject dependency

Emotions are very personally experienced states in human body and each individual experiences them differently. When having data in training set from the same subject or subjects than in test set, this makes the model subject dependent. Models can perform better when they are subject dependent since training and test sets contains data points from same subject which can make the classification easier. As seen in the table 3.1 most of the studies are subject dependent. Couple studies ([36] and [22]) tested subject independency but their models' performance were significantly poorer than subject dependency. However, in [24] they achieved accuracy of 98.8 % having used subject independency which is a great result.

3.9 Summary of related work

Support Vector Machine is the most common learning method used in emotion recognition problems. In addition, SVM models achieve desirable results for classifying different emotions. However, SVM is not always the best model to recognize emotions. In [30] KNN performed better than SVM with an accuracy of 82 %. It is worth mentioning that many of the databases used have data only from 50 or fewer subjects. [32] used their database with 101 subjects which is a fairly extensive database. However, their results with Random Forest only achieved 74 % accuracy, which is not the worst performance, but other models have more promising results. Recently, deep learning has increased popularity in machine learning approaches altogether; however, only a few studies were found within the emotion recognition scope. [24] and [2] both used CNN deep learning method with promising results 98.8 % and 87.3 % indicating that neural networks should be more researched as an emotion recognition model. It should be addressed that many of the studies used subject-dependent models resulting better performances than subject-in-dependency. Although Dar et al. [24] study did manage to get high performance rate of 98.8 % with subject independency.

Understandably, unsupervised models do not attract researchers as the selected model since emotions rely heavily on labeled data. Nonetheless, [37] shows that unsupervised models should not be excluded entirely. However, it should be considered that they used only EEG signals in their unsupervised model, and other studies presented in this thesis usually had more than one biosignal. EEG cannot be collected with wearable off-the-shelf devices, making it a problematic biosignal for this challenge.

Keeping in mind that this thesis concentrates primarily on signals collected with wearable devices, the most common biosignal in emotion recognition is ECG, while

overall, the most popular biosignal seems to be EEG, often used by itself. ECG presents vital data about heart rate, which makes it a valuable and reliable signal. When looking at the emotion models, it seems to be distributed evenly between ADM and DEM. If studies are grouped by their emotion models, the average accuracy in models with the ADM emotion model is 76 %. Models with DEM emotion model performed only two percentage points better if [24] is excluded from the calculations since they used both DEM and ADM emotion models. All related work reviewed in this thesis is presented in table 3.1.

Table 3.1: Related work of emotion recognition from physiological signals

Study	Year	Dataset	No. subjects	Subject independence	Biosignals used	Emotion model	Labels	ML-model	Best result	validation method
Kim et al. [4]	2004	own data	125+50	independent	ECG, PPG, EDA, SKT	DEM	sad, stressed, angry, surprised	SVM	78.43 %	not addressed
Wagner et al. [35]	2005	MIT+own	1	??	EMG, ECG, SC, Resp	DEM+ADM	joy, anger, sadness, pleasure	LDF	92.05 %	LOOCV
Zong et al. [34]	2009	Uni. Augsberg	1	signal dependent	ECG, EMG, SC, Resp	ADM	joy, anger, sadness, pleasure	SVM	76 %	10-fold CV
Petrantonakis et al. [36]	2010	own	16	independent	EEG	DEM	happiness, anger, fear, disgust, sadness, surprise	SVM	85.15 %	LOOCV
Kolodyazhniy et al. [30]	2011	own data	28	dependent & independent	ECG, EDA, RR, Temp, EMG	DEM	fear, sadness, neutral	LDA, KNN	82 %	34-fold, 204-fold and leave-one-subject-out
Agrafioti et al. [22]	2012	own	31	dependent *(tested independent)*	ECG	ADM+DEM	erotica, excitement, disgust, fear, gore & positive valence, negative valence	LD	76.19 %	LOOCV
Wen et al. [32]	2014	own	101	independent	GSR, ECG (HR), OXY	DEM	amusement, anger, grief, fear, baseline	RF	74 %	LOOCV
Das et al. [33]	2016	own?	4	dependent?	GSR, ECG	DEM	happy, sad, neutral	SVM	98.53 %	not addressed
Kanjo et al. [2]	2019	EnvBodySens	40	dependent *(tested independent)*	HR, GSR, BT, ACC	ADM	5-step valence (low to high)	CNN-LSTM	87.3 %	says only that 70-30 train/test
Pinto et al. [31]	2019	own	23	independent	ECG, EDA, BVP, Resp	ADM	arousal, valence, dominance, liking	SVM	ca. 68.44 %	4-fold CV
Liang et al. [37]	2019	DEAP	32	dependent	EEG	ADM	negative, positive	Hypergraph partitioning	65.12 %	LOOCV
Dar et al. [24]	2020	AMIGOS & DREAMER	23 & 23	independent	EEG, ECG, GSR	ADM	HVHA, HVLA, LVHA, LVLA	LSTM-CNN	98.8 %	randomly splitted the data twice average accuracy of both performance

4 Materials and methods

4.1 Data - CLAS

For the experimental part of the thesis, an open dataset downloaded from Mendeley Data [38] is used to test the model. The dataset is A Database for Cognitive Load, Affect and Stress Recognition (CLAS) from the University of Varna, Bulgaria [39]. The data was collected from 62 volunteers who were students between 20 and 27 of age, except one in their thirdies and one over fifty. Among these volunteers 17 were women and 45 were men. Dataset consists of three different biosignals recorded while subjects performed different cognitive tasks and watched emotion eliciting videos and pictures. Biosignals recorded from the subjects were ECG, EDA, and PPG. In addition, accelerometer data is collected, representing the subject's physical movement. The accelerometer is not in the scope of this thesis. In this thesis, only the data from the emotion eliciting part of the test is processed.

4.1.1 Data collection setup

The CLAS database contains data from three different cognitive stimuli: math problems, the Stoop test, and logic problems, as well as emotional stimuli. Each session started with a one-minute baseline recording before the cognitive part of the experiment, followed by the emotion eliciting audio-visual material. The baseline repre-

sents the normal state of the subject's body and mind. The cognitive stimuli data is excluded from this thesis; therefore, they are not described in detail. Different audio-visual materials were shown to the subjects to evoke emotions. The subjects watched 16 different emotionally tagged videos from the DEAP database [40] and 16 different emotionally tagged pictures from the IAPS database. The videos are organized in four blocks with four 30-second videos each. A neutral video stimulus was shown between the blocks. After the videos, the pictures were shown to the subjects also in four blocks with neutral video separating the pictures.

Three biosignals were recorded with Shimmer3 GSR+ Unit and Shimmer3 ECG Unit. The PPG signal was captured with an optical pulse sensor linked to the Shimmer3 GSR+ Unit. A sampling rate of 256 Hz was used to acquire all the biosignals with a 16-bit resolution per sample. The Shimmer3 GSR+ records the resistance of the skin, which was measured in kilo-ohms ($k\Omega$). The resistance can be converted to conductance with equation $G = 1/R$, where G is conductance and R is resistance. Therefore conductance is the reciprocal of resistance.

4.1.2 Data quality

Few subjects whose data were excluded from this thesis had problems in at least one of the signals (ECG, EDA, or PPG). The signals were missing data from the start or end of the experiment; some had unusable EDA or ECG signals having, e.g., no changes in the signal. The EDA signal is susceptible to artifacts if the electrodes are not attached well or if the signal is under 0.5 μS as covered in chapter 2.3. Figure 4.1 is an example of an EDA signal which was excluded from this experiment since there is not a proper informative signal. A good quality EDA signal is below that in the same figure where the EDA peaks are clearly noticeable, and the tonic level also changes. In the poor signal, these EDA peaks are not present, indicating that

the signal was not recorded correctly.

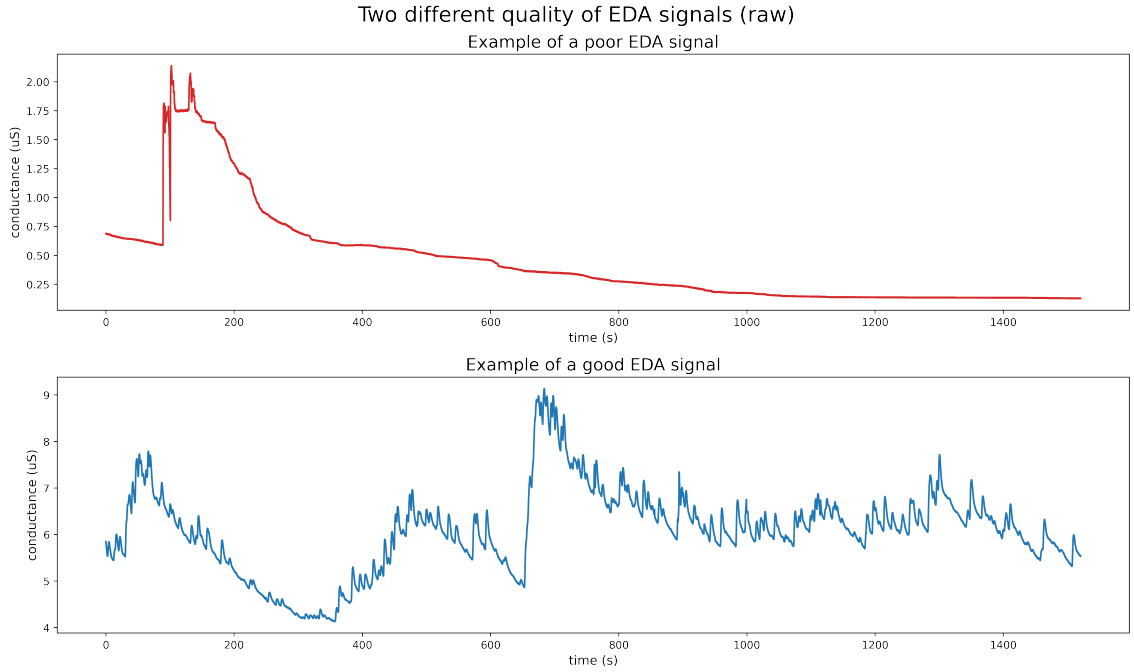

Figure 4.1: Comparison of a good EDA signal versus poor.

Data from 58 subjects were included in the model. A few signals had quite noticeable power line interference of 50/60 Hz in ECG signals and some EDA and PPG signals, which is fortunately easily filtered out - more on filtering and preprocessing methods in subsection 4.2.1. Overall, PPG signals were almost 100% usable signals on every subject, which might be due to the acquisition style. PPG was measured by attaching the meter to the subject's earlobe, which is more likely to stay unmoved compared to ECG electrodes on the subject's body or EDA electrodes on the fingers of the subject.

Most of the signals were fairly good quality with minor artifacts from movement and power line interference. In figure 4.2 is sample signals of ECG, PPG, and EDA from one subject. Note that each signal is a sample of different experiment times, and they are not synchronized in this figure. In ECG, the QRS-complex is easily distinguished while having noticeable power line interference. The PPG signal is very

Table 4.1: Emotion setup

(a) Setup for subjects ids 1-11

block	block type	dur (s)	emotion
10	neutral	30	neutral
11-14	videoclip	4 x 60	excitement
25	neutral	30	neutral
16-19	videoclip	4 x 60	bored
20	neutral	30	neutral
21-24	videoclip	4 x 60	calm
25	neutral	30	neutral
26-29	videoclip	4 x 60	stress
30	neutral	30	neutral
31	pictures	4 x 20	excitement
32	neutral	30	neutral
33	pictures	4 x 20	calm
34	neutral	30	neutral
35	pictures	4 x 20	bored
36	neutral	30	neutral
37	pictures	4 x 20	stress

(b) Setup for subjects ids 12-62

block	block type	dur (s)	emotion
10	neutral	30	neutral
11	pictures	4 x 20	excitement
12	neutral	30	neutral
13	pictures	4 x 20	calm
14	neutral	30	neutral
15	pictures	4 x 20	bored
16	neutral	30	neutral
17	pictures	4 x 20	stress
18	neutral	30	neutral
19-22	videoclip	4 x 60	excitement
23	neutral	30	neutral
24-27	videoclip	4 x 60	bored
28	neutral	30	neutral
29-32	videoclip	4 x 60	calm
33	neutral	30	neutral
34-37	videoclip	4 x 60	stress

good with noticeable "artifact" from respiration. The phasic and tonic levels are easily discovered from the EDA signal with little noise. These signals are somewhat ideal signal samples of the biosignals.

In figure 4.3 it is possible to see how emotions or how the subject feels can affect on EDA signal and heart rate (beats per minute). The red lines represent the start of a trigger which lasts until the next trigger starts. While the subject watched videos labeled as neutral or bored, the EDA signal level stays quite low and the heart rate

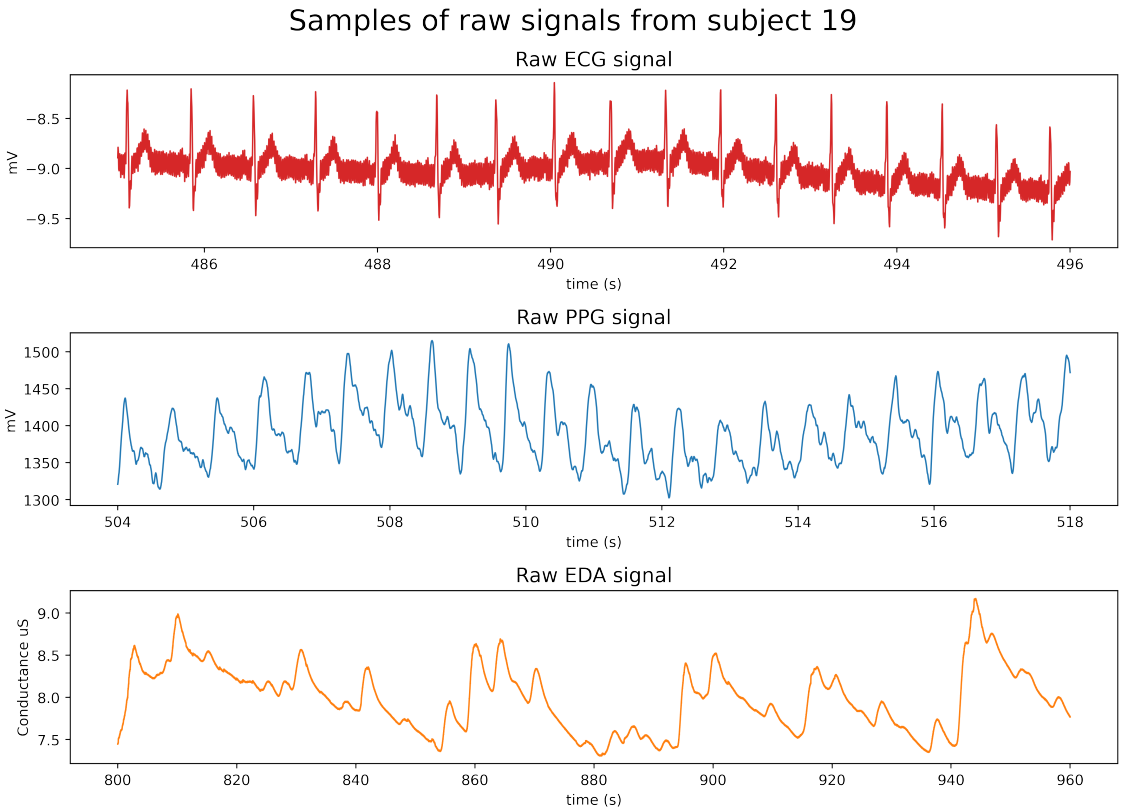

Figure 4.2: Samples of all three signals

does not necessarily elevate. Actually it can be seen that during neutral periods the heart rate lowers and during stress trigger it elevates a little. Noticeable elevation can be seen immediately after first excitement trigger. The heart rate reaches its peak during the third excitement trigger and likewise the highest point of EDA signal is reached during the same period.

4.2 Methodology

4.2.1 Pre-processing of biosignals

In the preprocessing of the filters, background noise were filtered out from ECG signal with Butterworth band-pass filter with cutoff frequencies of 0.5 Hz and 20 Hz. Fedotov 2016 [41] suggested that band-pass filter 8-20 Hz is the best filter

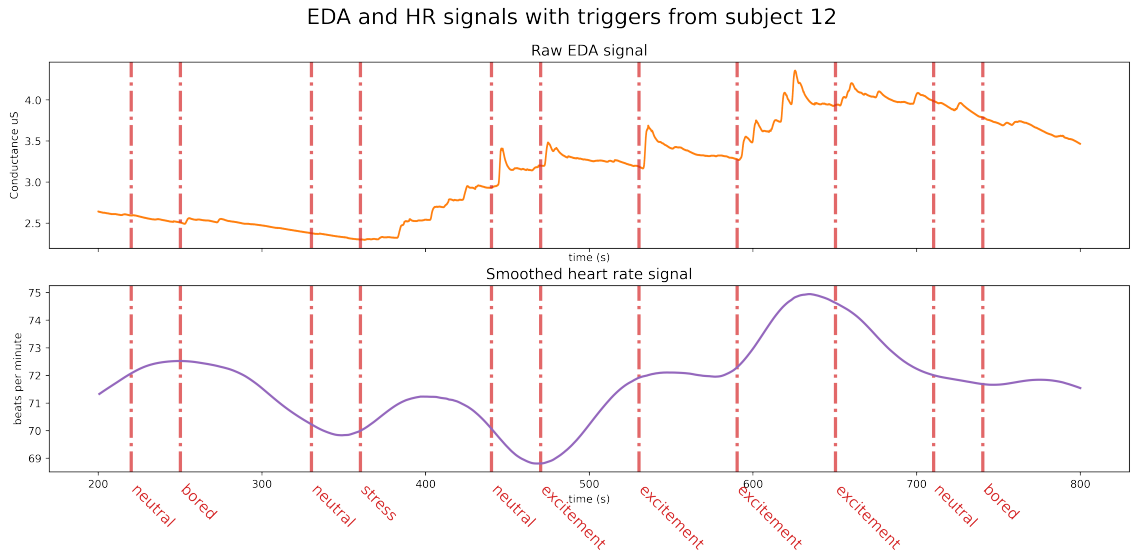

Figure 4.3: Example of how triggers affect on EDA signal and heart rate.

to measure R-R intervals with least amount of error. The low cut frequency of 8 Hz eliminates P- and T-waves of the ECG signals which are not relevant for our model since only R-peaks are calculated and features are extracted from R-peak information. However, cutoff frequencies of 0.5 Hz and 20 Hz were chosen since it they are common cutoff frequencies used in ECG preprocessing. The suggested frequencies were also tested but they didn't improve the model significantly. Within the preprocessing phase, few signals were excluded having a lot of noise in the

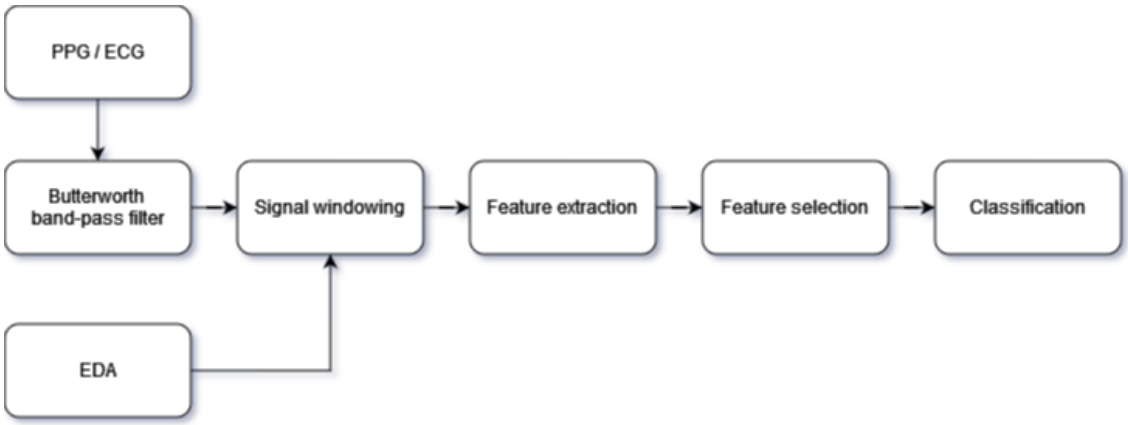

Figure 4.4: Experiment model pipeline.

signals.

The neutral signals were excluded from the experiment for not representing any emotion in interest.

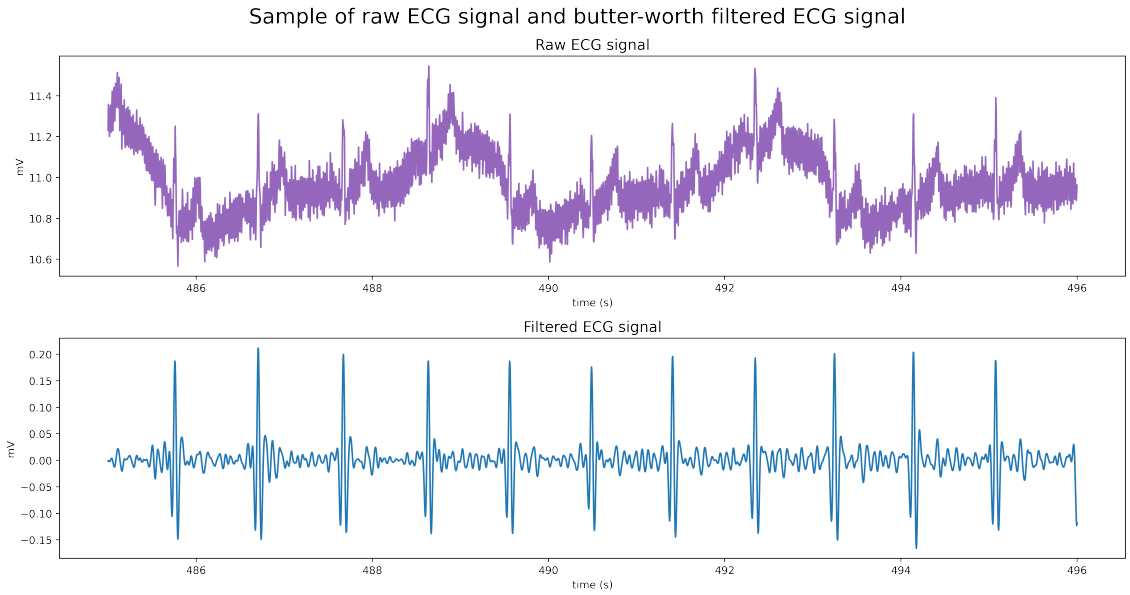

Figure 4.5: Butterworth band-pass filter to ECG signal with frequencies 0.5-20 Hz.

4.2.2 Feature extraction

Each signal were divided in different blocks based on their emotional trigger and each emotion block were divided into 20 second windows. 17 features were extracted from each signal from these 20 second windows. Features included statistical and frequency based features presented in table 4.2. Features were chosen based on the original study of the chosen dataset [39]. Note that no frequency based features were extracted from EDA signal.

4.2.3 Feature selection

Not all features were used for the model since too many features can result in loss of important features which can lead to poor performance. A Recursive Feature

Table 4.2: Features extracted from the signals

Signal	Statistical	Frequency
ECG+PPG	mean heart rate mean RR max NN interval pNN50 SDNN RMSSD std of the difference of successive NN intervals SD1 (short term variability) SD2 (long term variability)	powerband 0 - 0.4 Hz powerband 0.04 - 0.15 Hz powerband 0.15 - 0.4 Hz VLF percent LF percent HF percent LF/HF ratio
EDA	number of peaks max amplitude of the peaks min amplitude of the peaks mean conductance of the peaks RMS std of the peaks mean absolute value of the peaks skewness of the peak distribution kurtosis of the peak distribution mean resistance first quartile third quartile interquartile range percentile 2.5 percentile 10 percentile 90 percentile 97.5	*no frequency based features*

Elimination (RFE) were used to select a subset of features which contains the most efficient features for the classification model.

4.2.4 Classifiers

Two different supervised classifiers is trained; Random Forest and Support Vector Machines.

Random Forest

Random Forest is a simple classifier based on majority decision. It operates by constructing multiple decision trees at training and the output is selected based on the most selected output of the decision trees. A basic idea of how random forest classifies an instance is described on figure 4.6. Typically at least 100 decision trees are constructed. Each tree can pick only from a subset of features and the features are randomly picked for all the decision trees.

For this model, 150 decision trees are constructed with random state 20 with criteria of entropy.

Support Vector Machine

Support vector machine algorithm finds an optimal hyperplane that can distinctly classify different data points. Hyperplane is in N-dimensional space where N is the number of features. In figure 4.7 is a basic sample of a 2-dimensional classification problem with two different classes (blue and green). Red line represents the optimal hyperplane to divide the two classes.

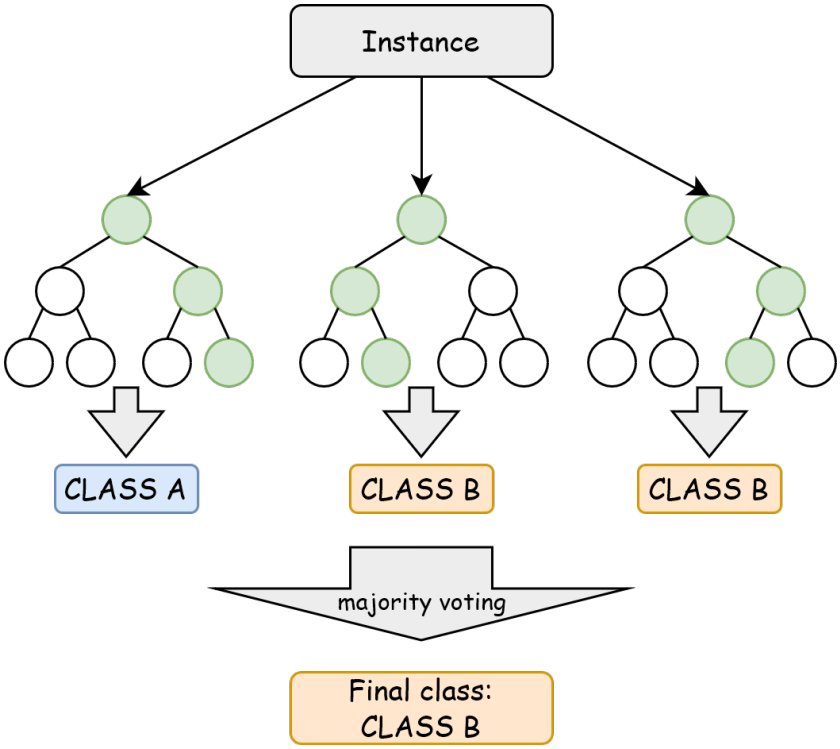

Figure 4.6: Basic structure of a Random Forest classifier

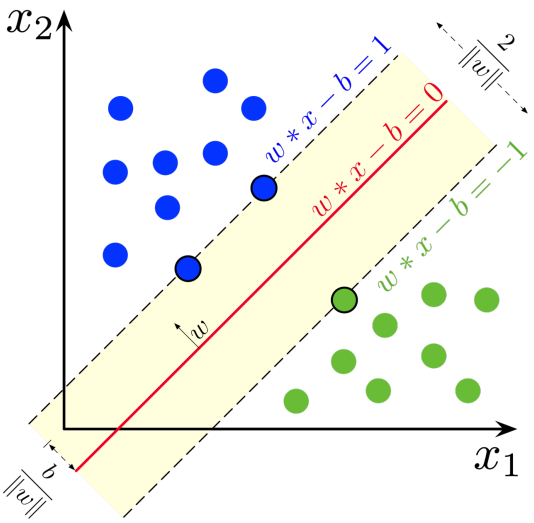

Figure 4.7: Example of 2-dimensional SVM classification hyperplane [42]

5 Experiment

The goal of this experiment was to find the best supervised machine learning models for emotion recognition using physiological signals. Test subjects were shown different videos and emotion to trigger different emotions which were excitement, calm, bored and stress. The signals were also labeled with stress and no-stress labels which helped to narrow down the experiment to first experiment if the models are able to differentiate these two classes from each other and plan was to then broaden the stress detection to four different emotion recognition.

The first objective was to find best model for emotion recognition or stress detection. Second objective was to figure out which biosignal is the most suitable for emotion or stress recognition. With biosignals the acquisition and pre-processing of the signal play big roles in the machine learning challenge and these were taken into account in analysing of the results.

5.1 Training supervised classifiers

Two different supervised classification models were trained; Random Forest and Support Vector Machines. Random Forest is a simple classification model that is had not been used in many related works. SVM was used in many previous studies and was therefore selected as one classifier for this experiment. Random Forest

was selected since it is quite simple algorithm and not the first choice for emotion recognition models. Leave-One-Subject-Out Cross-Validation were used to validate both models. Therefore the models were subject independent since there was not data from same subject in test and training sets. Subject dependent models were also tested for an attempt to improve the model performance rate.

Models firstly classified the data to two different labels; stressed and no stressed. This was on easier problem to tackle than to classify to four different emotions (excitement, calm, bored, stress). Excitement, calm and bored labels were labeled as *no stress* and stress as *stress*. This made the data imbalanced having different amount of samples for each class which was taken into account when training the data. SMOTE (*Synthetic Minority Over-sampling Technique)* [43] were used to tackle the imbalanced data. Since there is less data for stress-label, SMOTE duplicates stress-data to balance the data resulting in same amount of data between *stress* and *no stress* samples in the training sets.

5.2 Results

5.2.1 Performance metrics

The performance of the models were measured with precision, recall, f1-score and overall accuracy. These metrics are quite basic measurements of machine learning model performances. In all the metrics used, the higher the value is, the better the performance of the model is, highest value being 1.0.

Precision of the model measures how many of the positive identifications were actually correct. Precision is calculated as follows

$$Precision = \frac{TP}{TP+FP}$$

where TP = True Positives and FP = False Positives.

Recall attempts to measure how many of the actual positives were identified correctly. Recall is calculated as follows

$$Recall = \frac{TP}{TP+FN}$$

where FN = False Negatives.

F1-score is the mean of the precision and recall.

$$\text{f1-score} = 2 \times \frac{precision \times recall}{precision + recall}$$

where precision and recall is calculated as previously.

Accuracy is simple way to measure the performance of the model by dividing correctly classified examples by all examples as following

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

where TN = True Negatives. The overall accuracy can be quite good but it still doesn't give information on how the model actually performs on identifying true positives from negatives (stress from no stress).

5.2.2 Performance results of the experiment models

The results of the stress classifiers were not as good as expected which indicated that recognizing four different emotions instead of stress and no stress from the dataset would be very difficult task thus leaving emotion recognition out of the scope and it was not implemented.

The results of Random Forest classifier is presented in table 5.2 and SVM in table

Table 5.1: Results of all classified as one label

Target	precision	recall	f1-score	overall accuracy (stress+no stress)
All stress	24 %	100 %	39 %	24.47 %
All no stress	76 %	100 %	86 %	75.53 %

Table 5.2: Results of Random Forest classifier

signal(s)	precision	recall	f1-score	overall accuracy (stress+no stress)
ECG	23 %	29 %	26 %	59.46 %
EDA	24 %	25 %	25 %	62.72 %
PPG	26 %	36 %	31 %	59.88 %
ECG+EDA	24 %	28 %	26 %	60.79 %
PPG+EDA	26 %	28 %	27 %	62.66 %

5.3. It can be said that both models performed poorly on recognizing stress from no stress signals. SVM had good recall score of 85 % using only EDA signal but the overall accuracy suffered being only 31.72 % meaning that it classified many no stress examples as stressed even though it labeled many true stressed as stressed.

It is good to compare the results of classification models to results of classifying all the instances as one label. In table 5.1 is presented results when all the instances are classified as stressed or no stressed. When classifying all instances to stressed, precision is only 24 % which is almost the same performance result of the experiment models.

When looking at the area under the curve, it can be seen that the models performed very poorly. If the ROC curve approaches diagonal line from lower left side to upper right side, it indicates that the model probably just randomly classifies the labels to instances rather than actually learns anything from the data. The ROC cure of RF

Table 5.3: Results of Support Vector Machines classifier

signal(s)	precision	recall	f1-score	overall accuracy (stress+no stress)
ECG	28 %	35 %	31 %	62.48 %
EDA	24 %	85 %	38 %	31.72 %
PPG	23 %	51 %	32 %	47.13 %
ECG+EDA	28 %	27 %	28 %	65.26 %
PPG+EDA	24 %	47 %	32 %	50.39 %

model is presented in figure 5.2 and SVM in figure 5.2. Neither of the curves differ notably from a straight diagonal line. This indicates well that the models performed poorly in the stress detection challenge.

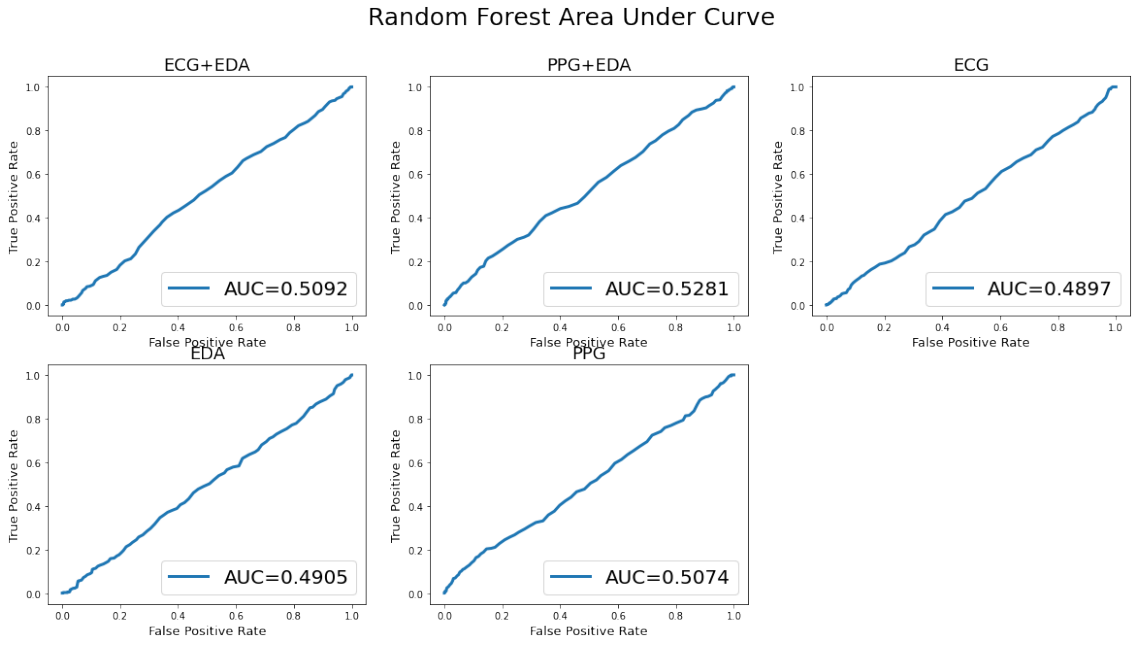

Figure 5.1: Area Under Curve from Random Forest classifier with different biosignals.

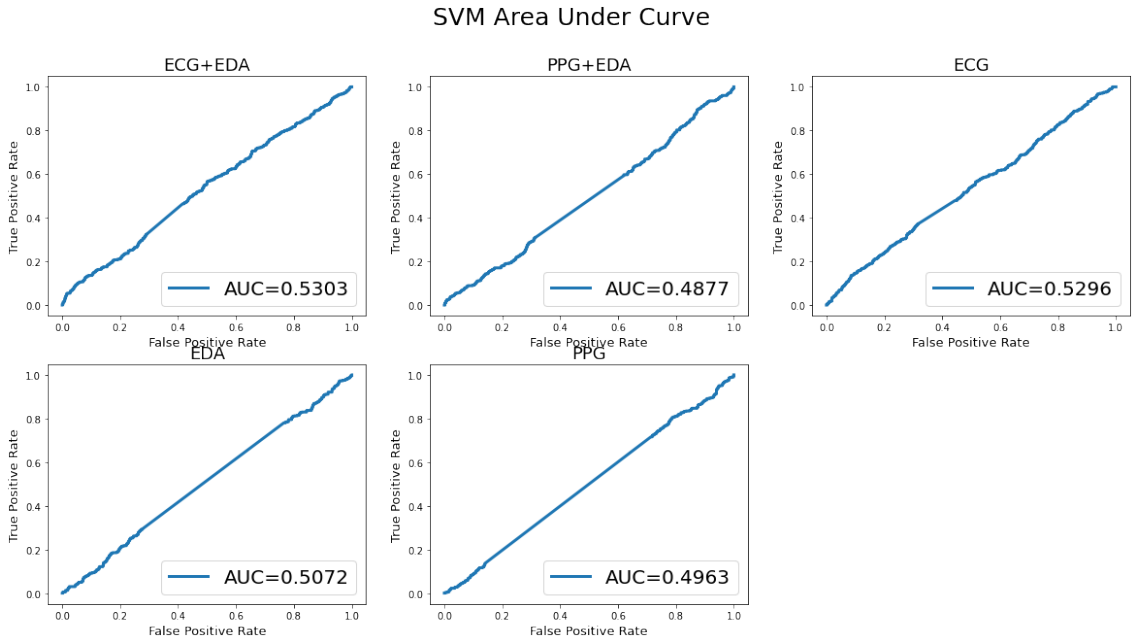

Figure 5.2: Area Under Curve from Support Vector Machines classifier with different biosignals.

6 Conclusion

The hypothesis were proven to be incorrect; emotions are difficult to recognize from biosignals using methods presented in this thesis (RQ1). Answering to the RQ2 is difficult since both of the models in this experiment performed poorly. However, taking into account previous works, SVM seem to be a promising algorithm for supervised machine learning model for emotion recognition. Although the results should be critically examine since emotions are difficult to recognize even by a human being and emotions are personally experienced thus common factors - if any - are difficult to detect.

To find the best biosignal for emotion recognition - anwering to the RQ3 - is not that easily selected. EDA is an important signal for emotion recognition since it has a strong correlation with emotional arousal. However, EDA signal can be surprisingly difficult to acquire reliably. It is prone to different artifacts and the placement of sensors plays a big role in the signal acquisition and the environment can also have a huge impact on the signal (cool and dry environment versus warm and humid environment). ECG and PPG gives important features about the heart and PPG did even perform slightly better than ECG with the same features. PPG is easier to acquire with wearable devices and still managing to generating important information about emotions.

With the dataset used proved that the quality of the data is an important aspect in this problem. Looking at the dataset provided in this experiment might not have been the most compatible data with this problem. The signals were measured using commercial devices and the quality of the signals was not accurate enough with some subjects and some of the samples had to be cut off. Especially the EDA signal had problems since the skin conductance should be at least 0.5 mS to be accurate and many of the signals were under 0.5 mS.

The plan was to implement emotion recognition using the CLAS dataset but when the dataset was tested to recognize stress from other emotions, the models didn't even manage to detect stress from no-stress signals. Performing a multi-class detection is far more complex than two class classification, therefore emotion recognition with four emotions was sadly not implemented.

Previous studies do present optimistic results which indicates that emotion recognition with supervised machine learning models is possible and should be researched more. The challenge is that the experiment setup should be well designed and implemented which can have some problems in execution. Designing a study protocol with emotion recognition is extremely difficult since emotions are remarkably subjective therefore trying to trigger same emotions across the study subjects is challenging achieve.

References

[1] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system", in *Tutorial and research workshop on affective dialogue systems*, Springer, 2004, pp. 36–48.

[2] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection", *Information Fusion*, vol. 49, pp. 46–56, 2019.

[3] E. Syrjälä, M. Jiang, T. Pahikkala, S. Salanterä, and P. Liljeberg, "Skin conductance response to gradual-increasing experimental pain", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 3482–3485.

[4] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals", *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[5] M. Strik, T. Caillol, F. D. Ramirez, S. Abu-Alrub, H. Marchand, N. Welte, P. Ritter, M. Haissaguerre, S. Ploux, and P. Bordachar, "Validating qt-interval measurement using the apple watch ecg to enable remote monitoring during the covid-19 pandemic", *Circulation*, vol. 142, no. 4, pp. 416–418, 2020.

[6] *Wikipedia 10-20 system (eeg)*, https://en.wikipedia.org/wiki/10-20_system_(EEG), Accessed: 2021-11-26.

[7] W. W. Ismail, M. Hanif, S. Mohamed, N. Hamzah, and Z. I. Rizman, "Human emotion detection via brain waves study by using electroencephalogram (eeg)", *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 1005–1011, 2016.

[8] X. Hu, J. Chen, F. Wang, and D. Zhang, "Ten challenges for eeg-based affective computing", *Brain Science Advances*, vol. 5, no. 1, pp. 1–20, 2019.

[9] O. Kwon, J. Jeong, H. B. Kim, I. H. Kwon, S. Y. Park, J. E. Kim, and Y. Choi, "Electrocardiogram sampling frequency range acceptable for heart rate variability analysis", *Healthcare informatics research*, vol. 24, no. 3, pp. 198–206, 2018.

[10] S. Mahdiani, V. Jeyhani, M. Peltokangas, and A. Vehkaoja, "Is 50 hz high enough ecg sampling frequency for accurate hrv analysis?", in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2015, pp. 5948–5951.

[11] J. Laitala, "Using lstm network to detect r-peaks from noisy ecg signals", *Health Technology*, 2020.

[12] *Emergency Medical Technician 12-lead ecg placement & corrections*, http://www.emtresource.com/resources/ecg/12-lead-ecg-placement/, Accessed: 2021-11-24.

[13] R. J. Martis, U. R. Acharya, and H. Adeli, "Current methods in electrocardiogram characterization", *Computers in biology and medicine*, vol. 48, pp. 133–149, 2014.

[14] J. Montagu and E. M. Coles, "Mechanism and measurement of the galvanic skin response.", *Psychological Bulletin*, vol. 65, no. 5, p. 261, 1966.

[15] B. Choi, H. Jebelli, and S. Lee, "Feasibility analysis of electrodermal activity (eda) acquired from wearable sensors to assess construction workers' perceived risk", *Safety science*, vol. 115, pp. 110–120, 2019.

[16] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments", *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[17] *Biopac Systems, Inc. eda data analysis & corrections*, `https://www.biopac.com/eda-faq-data/`, Accessed: 2021-11-23.

[18] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals", *Current cardiology reviews*, vol. 8, no. 1, pp. 14–25, 2012.

[19] J. Wannenburg and R. Malekian, "Body sensor network for mobile health monitoring, a diagnosis and anticipating system", *IEEE Sensors Journal*, vol. 15, no. 12, pp. 6839–6852, 2015.

[20] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable photoplethysmographic sensors—past and present", *Electronics*, vol. 3, no. 2, pp. 282–302, 2014.

[21] J. Allen, "Photoplethysmography and its application in clinical physiological measurement", *Physiol. Meas*, vol. 28, R1–R39, 2007.

[22] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection", *IEEE Transactions on affective computing*, vol. 3, no. 1, pp. 102–115, 2011.

[23] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential", *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[24] M. N. Dar, M. U. Akram, S. G. Khawaja, and A. N. Pujari, "Cnn and lstm-based emotion charting using physiological signals", *Sensors*, vol. 20, no. 16, p. 4551, 2020.

[25] P. Ongsulee, "Artificial intelligence, machine learning and deep learning", in *ICT and Knowledge Engineering (ICT&KE), 2017 15th International Conference on*, IEEE, 2017, pp. 1–6.

[26] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods", *Sensors*, vol. 20, no. 3, p. 592, 2020.

[27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, ISBN: 978-0-262-01802-9.

[28] M. L. Samb, F. Camara, S. Ndiaye, Y. Slimani, and M. A. Esseghir, "A novel rfe-svm-based feature selection approach for classification", *International Journal of Advanced Science and Technology*, vol. 43, no. 1, pp. 27–36, 2012.

[29] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, T. A. Ito, *et al.*, "The psychophysiology of emotion", *Handbook of emotions*, vol. 2, no. 01, p. 2000, 2000.

[30] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm, "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions", *Psychophysiology*, vol. 48, no. 7, pp. 908–922, 2011.

[31] J. Pinto, A. Fred, and H. P. da Silva, "Biosignal-based multimodal emotion recognition in a valence-arousal affective framework applied to immersive video visualization", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 3577–3583.

[32] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals", *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 126–140, 2014.

[33] P. Das, A. Khasnobish, and D. Tibarewala, "Emotion recognition employing ecg and gsr signals as markers of ans", in *2016 Conference on Advances in Signal Processing (CASP)*, IEEE, 2016, pp. 37–42.

[34] C. Zong and M. Chetouani, "Hilbert-huang transform based physiological signals analysis for emotion recognition", in *2009 IEEE international symposium on signal processing and information technology (ISSPIT)*, IEEE, 2009, pp. 334–339.

[35] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification", in *2005 IEEE international conference on multimedia and expo*, IEEE, 2005, pp. 940–943.

[36] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis", *IEEE Transactions on affective computing*, vol. 1, no. 2, pp. 81–97, 2010.

[37] Z. Liang, S. Oba, and S. Ishii, "An unsupervised eeg decoding system for human emotion recognition", *Neural Networks*, vol. 116, pp. 257–268, 2019.

[38] *Mendeley Data clas: A database for cognitive load, affect and stress recognition*, `https://data.mendeley.com/datasets/8hm59ryzb8/1`, Accessed: 2021-03-24.

[39] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition", in *2019 International Conference on Biomedical Innovations and Applications (BIA)*, IEEE, 2019, pp. 1–4.

[40] 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", *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[41] A. Fedotov, "Selection of parameters of bandpass filtering of the ecg signal for heart rhythm monitoring systems", *Biomedical Engineering*, vol. 50, Sep. 2016. DOI: 10.1007/s10527-016-9600-8.

[42] *Wikipedia support vector machine*, https://en.wikipedia.org/wiki/Support-vector_machine, Accessed: 2022-01-07.

[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.