Rose Lu

EEG-Based Emotion Experiment

Automatic Recognition with Machine Learning

Master's thesis in Cybernetics and Robotics
Supervisor: Marta Molinas
Co-supervisor: Mohit Kumar
May 2023

NTNU

Norwegian University of
Science and Technology

Rose Lu

EEG-Based Emotion Experiment

Automatic Recognition with Machine Learning

Master's thesis in Cybernetics and Robotics
Supervisor: Marta Molinas
Co-supervisor: Mohit Kumar
May 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

NTNU
Norwegian University of
Science and Technology

Preface

The master's thesis is the end product of finalizing my Master's in Cybernetics and Robotics conducted during the spring semester of 2023 with the Department of Engineering Cybernetics at the Norwegian University of Science and Technology (NTNU). Part of this thesis is based on the previously written specialization project entitled "Automatic Emotion Decoding from EEG Signals Using the SEED Dataset"[1] done in the autumn semester of 2022 with guidance from the same supervisor and co-supervisor. Parts of Chapter 2 and Chapter 4 are updated and extended versions of the specialization project. In addition, some of the models used in this thesis have already been presented in two conference papers for the 10th International BCI Meeting[2] and IEEE EMBC 2023[3].

Firstly, I would like to express my gratitude to my supervisor *Marta Molinas* and my co-supervisor *Mohit Kumar* for introducing me to this field of research. I will also thank them for their invaluable guidance, feedback and support throughout the research and writing process. It has been helpful and valuable. Also, thanks to *Andres Soler* for helping out with the experiment setup and preparation for EEG data collection. I would like to acknowledge my friends and family. My completion of this thesis would not have been possible without great support from them. A special thanks for helping me keep my spirits high during challenging times. And also many thanks to all the volunteers for their time and participation in the experiments. Finally, to *Embla Celine Stengel Neverlien* for always being encouraging and supportive. It has been a pleasure collaborating with her on the collection of EEG data and paper writing. Thanks for an excellent partnership.

It is my hope that this research can contribute to further research and exploration within EEG-based emotion recognition. I had no prior experience with EEG signals or machine learning when diving into this research area. Hopefully, this thesis disclose and reflects the new knowledge and experience gained this year.

Rose Lu
Trondheim, 31.05.2023

---

[1]http://dx.doi.org/10.13140/RG.2.2.33244.67205
[2]https://bcisociety.org/bci-meeting/
[3]https://embc.embs.org/2023/

Abstract

Emotions have a large impact on our daily lives. Understanding and being able to control our emotions can lead to improved self-awareness, empathy, and quality of life. Conventional approaches to emotion recognition involve asking individuals to describe their emotional experience or use observer ratings, making them subjective and prone to biases. To overcome some of these issues, researchers have turned to more objective approaches, including electroencephalography (EEG) and machine learning algorithms.

This thesis explores how different emotional states are reflected in EEG signals and the possibilities for automatic decoding and classification of human emotions in terms of valence and arousal. The experiment involves EEG data collected from 20 subjects with two 8-channel Unicorn systems. Each subject watched 52 non-audio movie clips of approximately 40 seconds each and rated the elicited emotions according to the 9-point self-assessment manikin (SAM) scale. The data was then filtered and preprocessed before numerical features from the time domain, i.e. Hjorth parameters, and the frequency domain, i.e. power spectral density (PSD) and differential entropy (DE), were extracted from the signals. This work also explores features with nonlinear characteristics, including fractal dimension. The mentioned features and combinations of them were extracted from epochs of four different sizes, i.e. 5, 10, 20, and 40 seconds. Some were also extracted from five frequency bands of $\delta, \theta, \alpha, \beta$ and γ. The extracted features were fed into the classifiers support vector machine (SVM), K-nearest neighbour (KNN) and multilayer perceptron (MLP), and evaluated on the newly-created dataset.

Further analysis is needed to make a finite conclusion on which classification method and feature combination had the best performance. Based on the experimental results, no final conclusive observations were made. Subject variability, unbalanced classes in the dataset, and in between valence and arousal discrimination are factors that might have prevented this. However, most of the results were above 50%, i.e. better than random guessing. In addition, most of the best-performing models had a performance somewhere between 59%-73%, depending on whether the used metric is average accuracy or average F1-score. High/low arousal classification outperformed high/low valence classification on average. The performance for all models is subject-independent, i.e. the models are trained and predicted across subjects.

The work is also compared with the previously written specialization project on emotion decoding on SEED dataset, other publicly available datasets and state-of-the-art methods. The accuracies and F1-scores from the specialization project and the state-of-the-art methods outperformed models evaluated on the newly-created dataset. The models used in this work have been selected based on the specialization project, which might lead to some overfitting. Further, most of the state-of-the-art methods are complex, involving deep learning methods and complicated architecture. However, the performance, compared to other EEG-based emotion datasets (i.e. DEAP, DREAMER, AMIGOS, IDEA), is consistent with the average accuracy obtained by the others. Some of the proposed models outperform those datasets that reported F1-score with more than 13 percentage points.

Sammendrag

Følelser er en viktig del av livet og har en stor innvirkning i vårt daglige liv. Det å forstå og kunne kontrollere følelser kan for noen føre til bedre selvbevissthet, empati og livskvalitet. Tradisjonelle metoder for identifisering av følelser innebærer blant annet å be enkeltpersoner beskrive sine egne følelser eller bruke observeringsmetoder. Dette er metoder som er subjektive og utsatt for bias. For å løse noen av disse problemene har forskere begynt å utforske alternative tilnærminger, som blant annet involverer elektroencefalografi (EEG) og maskinlæring.

Denne masteroppgaven ser på hvordan emosjonelle tilstander reflekteres i EEG-signaler, samt mulighetene for automatisk gjenkjenning og klassifisering av følelser. I eksperimentet ble EEG-data fra 20 deltagere samlet inn ved hjelp av to Unicorn systemer med åtte elektroder hver. Hver deltager fikk se 52 filmklipp uten lyd på 40 sekunder hver, før de vurderte de fremkalte følelsene i henhold til 9-punkt SAM-skalaen. SAM er en metode for å måle graden av glede, opphisselse og dominans. EEG-dataen ble deretter filtrert og preprosessert før signalets *features* fra tidsdomenet og frekvensdomenet ble hentet ut. Oppgaven utforsker også *features* med ikke-lineære egenskaper. *Features* som brukes i denne oppgaven er *differential entropy* (DE), effektspektral tetthet (PSD), Hjorth parametre og fraktale dimensjoner hentet ut fra epoker av fire ulike størrelser på 5, 10, 20 og 40 sekunder. Noen *features* er også hentet ut fra de fem velkjente frekvensområdene $\delta, \theta, \alpha, \beta$ and γ. *Features* ble deretter matet inn i klassifiseringsmetodene støttevektormaskin (SVM), k-nærmeste naboer (KNN) og flerlags perceptron (MLP) for å evaluere dem på det nye datasettet.

Mer forskning er nødvendig for å komme til en endelig konklusjon om hvilke modeller som yter best basert på resultatene. Variasjon mellom deltagere, ubalansert datasett og vanskeligheter med å skille mellom glede og opphisselse er faktorer som kan ha forhindret dette. Imidlertid var de fleste resultatene over 50%, altså bedre enn tilfeldig gjetting. Litt avhengig av om den rapporterte ytelsesmetrikken er gjennomsnittlig nøyaktighet eller gjennomsnittlig F1-poengsum, haddde de beste modellene en ytelse mellom 59%-73%. Modellene er bedre til å predikere høy/lav opphisselse enn høy/lav glede. Ytelsen for alle modellene er uavhengig av deltagerne, altså er modellene er trent og predikert på tvers av forsøkspersoner.

Arbeidet blir også sammenlignet med prosjektoppgaven, andre tilgjengelige EEG-datasett (dvs. DEAP, DREAMER, AMIGOS, IDEA) og metoder fra eksisterende litteratur. Prosjektoppgaven og metoder fra den eksisterende litteraturen presterte bedre enn modellene som ble evaluert på det nye datasettet. Modellene som er brukt i dette arbeidet, er valgt basert på arbeidet gjort i prosjektoppgaven, og kan dermed lide av overtilpasning. Dessuten involverer de fleste metodene fra eksisterende litteratur dyp læring og komplisert arkitektur. Ytelsen, sammenlignet med andre EEG-baserte datasett, er i tråd med gjennomsnittlig nøyaktighet oppnådd av de andre. Noen av F1-poengsummene til de beste modellene presterer mer enn 13 prosentpoeng bedre enn F1-poengsummen til enkelte datasettene.

Table of Contents

List of Tables

List of Figures

Abbreviations and Acronyms

CV	Cross-validation
DE	Differential entropy
DNN	Deep neural network
EEG	Electroencephalogram
FD	Fractal dimension
GNB	Gaussian Naïve Bayes
HA / LA	High arousal / low arousal
HC	Hjorth complexity
HFD	Higuchi fractal dimension
HM	Hjorth mobility
HV / LV	High valence / low valence
ICA	Independent components analysis
KFD	Katz fractal dimension
KNN	K-nearest neighbour
ML	Machine learning
MLP	Multilayer perceptron
PSD	Power spectral density
SAM	Self-assessment manikin
SVM	Support vector machine

AMIGOS	Multimodal dataset of **A**ffect, personality traits and **M**ood on **I**ndividuals and **Gr**O**up**S
DEAP	Database for Emotion Analysis using Physiological signals
DECAF	Multimodal dataset for **de**coding user physiological responses to **af**fective multimedia content
EMDB	Emotional movie database
IDEA	Intellect database for emotion analysis
SEED	SJTU Emotion EEG Dataset

1

Introduction

Emotions have a vital impact on our daily lives as they affect how we handle different situations in different ways, our communication abilities, physical health, mental state, etc. In other words, it is essential in non-verbal communication due to its relation to human behaviour and cognition. It is common to split emotions into two main categories: non-physiological and physiological signals. Physiological signals are mainly based on a person's background and are considered subjective. Examples of non-physiological signals are heart rate, neuro signals, skin impedance, and brain signals. Many people find it challenging to grasp and express their emotions accurately with facial expressions, body language or speech (Suhaimi et al., 2020). This can be a challenge for the fields where correct emotion recognition is essential, including psychology, artificial intelligence and medical treatment. The physiological signals tend to be disguised more easily. To get more reliable results, non-physiological signals can be used. They allow for direct assessment of the "inner" states and tend to be more objective to a certain extent (Zheng et al., 2019).

1.1 Background and Motivation

Emotion research is an interdisciplinary field that encloses research within psychology, computer science, control theory, neuroscience, etc. The scientific knowledge of human emotions has been greatly exceedingly investigated in psychology. What is less known is electroencephalography-based emotion recognition (Suhaimi et al., 2020). With the new research branch known as affective science, a more objective way of getting to know emotions is developed. It has shown huge potential within both the medical and commercial fields. According to the Gartner Hype Cycle report of 2022[1], emotional artificial intelligence (AI) technology, including affective computing, is one of the emerging technologies

---

[1]https://www.gartner.com/document/4016927

that is expected to make a transformative impact within the next decade. The literature review conducted by Al-Nafjan et al. (2017) also shows an explosive interest in EEG-based emotion detection and recognition.

Further, EEG-based emotion recognition is helpful in the diagnosis of depression, PTSD, and other mental disorders (Liu et al., 2021). A person's emotional responses when exposed to different environments and stimuli, can be non-invasively extracted from small electrodes on the scalp. In addition, it can bring a more satisfactory user interface and usability regarding brain-computer interfaces (BCI) and human-machine interaction (HMI) (Pathirana et al., 2018). This can promote effective communication among individuals and human-to-machine information exchange. Another interesting aspect is the understanding of how human emotions work. There is no simple mapping between emotions and specific brain structures. Different emotions can activate similar locations in the human brain. However, a single emotion can also activate several brain structures.

1.2   Problem Description

This work aims at creating an emotion-induced dataset to contribute to the further development of models to decode human emotions through EEG signal analysis. An investigation of both the traditional machine learning techniques and the simpler deep learning techniques of emotion recognition evaluated on SEED will also be evaluated on the newly created EEG-based dataset.

Most current state-of-the-art methods are based on complex and computationally-heavy deep learning architectures. Although they can output high-performance accuracy, their computational configuration will not be practical and beneficial for the limited number of classes. Simpler machine learning techniques would be preferable as they are more intuitive to implement and understand. According to Asadur Rahman et al. (2020), using too complicated architectures to obtain high accuracy is like "*using cannon to kill a mosquito*". The generalization of the less complex machine-learning techniques and the combinations of feature extraction in the specialization project will be further investigated in this thesis. The results obtained in this work will be compared to the results from the specialization project on SEED (2015) and the already publicly available EEG-based emotion datasets: AMIGOS (2021), DEAP (2012), DREAMER (2018), and IDEA (2022).

Figure 1.1 shows the framework for emotion classification. Yang et al. (2020) suggests a workflow is split into four phases, which will be followed in this work. Phase one focuses on emotion elicitation and data collection, the second phase includes preprocessing and filtering the obtained raw EEG signals, the third phase is feature extraction, and the final phase is classification. The main goal is for the models to decode and recognize emotions in EEG signals.
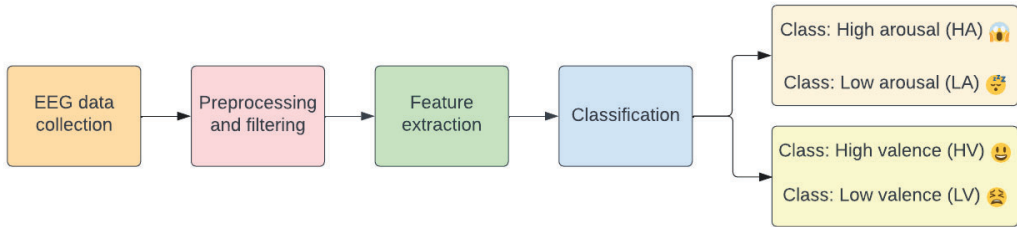
Figure 1.1: The framework for emotion classification

1.2.1   Main Objectives

The following main objectives have been formulated and will be investigated throughout this thesis:

1. Experiment design and data collection for an EEG-based emotion dataset with non-audio movie clips as stimuli

2. Study of existing feature extractors and different combinations of them:

   - Hjorth parameters: Mobility (HM) and complexity (HC)
   - Fractal dimensions: Higuchi (HFD) and Katz (KFD)
   - Power spectral density (PSD)
   - Differential entropy (DE)

3. Classify human emotions with traditional machine learning and simple deep learning techniques:

   - Support vector machine (SVM)
   - K-nearest neighbour (KNN)
   - Multilayer perceptron (MLP)

1.3   Related Works

In literature, the interest in EEG for automatic emotion decoding and recognition has increased. However, it is a challenging field. One of the challenging parts of emotion recognition is the bottleneck issue with the lack of enough available training data (Yang et al., 2020). Other challenges include different responses towards the same stimuli as the participants might elicit incorrect emotions. Other factors like the participant's background, workload, personal preferences, or sociability might affect the elicited emotions as well (Zheng and Lu, 2015).

3

1.3.1   Relevant Datasets

In recent years several publicly available datasets with stimuli media and emotional states through different modalities have been explored, e.g. magnetoencephalogram (MEG), electrocardiogram (ECG), heart rate (HR), and galvanic skin response (GSR). These let the researchers gain more insight into emotions and EEG, in addition, to allowing them to compare methods. A more strcutured summary can be found in Appendix B.

In 2012, Koelstra et al. presented the DEAP dataset. It consists of EEG data from 32 participants watching 40 one-minute excerpts of music videos eliciting different emotions. For each video, the participants rated the videos regarding the levels of valence, arousal, and dominance, inspired by Russell's circumplex emotion model. In addition, their familiarity with the videos and their liking of them were also reported. The EEG signals were recorded with 32 electrodes, placed according to the international 10-20 system, with a sampling rate of 512 Hz. Further, the data were downsampled to 128Hz and a bandpass frequency filter from $4 - 45$ Hz was applied before being segmented into 60-second trials. From their results, valence showed a stronger correlation with EEG signals, while for the EEG-based features, arousal achieved higher accuracy. Further details are presented by Koelstra et al. (2012).

A similar dataset is DECAF (Abadi et al., 2015). However, it uses the MEG sensor to acquire data from 30 participants. MEG is a non-invasive technology that captures the magnetic activity of the brain, thus requiring little physical contact between the user and the sensing coil. The stimuli used here are the same 40 one-minute music videos used in DEAP, in addition to 36 extra movie clips. The MEG signals were downsampled to 300 Hz. The preprocessing also involved channel correction and filtering with lowpass and highpass filters. Further details are presented by Abadi et al. (2015).

SEED dataset (Zheng and Lu, 2015) uses 15 four-minute movie clips as stimuli to elicit discrete emotions, i.e. positive, negative, and neutral emotions. Fifteen healthy students performed three different sessions each with an interval of about one week, which makes 45 trials in total. The EEG signals were collected with the 62-channel ESI NeuroScan system with a sampling rate of 1000 Hz according to the international 10-20 system. These signals were further downsampled to 200 Hz and preprocessed with a bandpass frequency filter from $0 - 50$ Hz. Their investigation of critical frequency bands indicates that higher frequency bands of EEG data are more related to emotion recognition. Further details are presented by Zheng and Lu (2015).

Moving forward, AMIGOS dataset (Miranda-Correa et al., 2021) consists of data from three modalities, namely EEG, ECG, and GSR. The EEG data from 40 participants watching videos were processed and collected with a sampling frequency of 128 Hz according to the international 10-20 system. A bandpass frequency filter from $4 - 45$ Hz is also applied. They discovered that social context has an impact on valence and arousal expressed by participants. Further details are presented by Miranda-Correa et al. (2021)

Katsigiannis and Ramzan (2018) created DREAMER containing recorded EEG and ECG signals of 23 participants during affect elicitation with 18 different movie clips as stimuli.

Emotions in terms of valence, arousal and dominance are evaluated after each stimulus. The data were recorded at a sampling rate of 128 Hz with 14-channel EEG recordings according to the international 10-20 system. For the filtering process, three separate band-pass Hamming sinc linear phase FIR filters are applied and shifted by the filter's group delay. Further, they were preprocessed with Artefact Subspace Reconstruction (ASR) and Common Average Reference (CAR). Further details are presented by Katsigiannis and Ramzan (2018).

One of the newest contributions to the emotion datasets is IDEA (Joshi and Ghongade, 2022). The dataset is generated for positive and negative emotions with audio-video clips and mental mathematical problems as stimuli. The emotions are further divided into four subclasses. Positive emotions correspond to cheerfulness, pleased, relaxation and zest. While anger, distress, restlessness, and sadness are the subclasses known for negative emotions. EEG data is collected from 14 participants using 24 electrodes for EEG recording at a sampling rate of 256 Hz. These data were further notch filtered to remove power line interference and preprocessed with a sixth-order Butterworth pass-band filter in the range of $4 - 44$ Hz. Based on their work, selecting fewer channels might achieve higher accuracy as reduces excessive information and computational costs. Further details are presented by Joshi and Ghongade (2022).

Not all datasets use auditory content as stimuli, and EMDB (Carvalho et al., 2012) is one example. They used 52 non-auditory movie clips to evoke emotions from different quadrants of affective space. The resulting categories are horror, erotic, social positive interactions, social negative interactions, scenery, and object manipulation. The recorded data are the heart rate (HR) and the skin conductance level (SCL) of 32 participants. Scores for valence, arousal and dominance are assessed using the self-assessment manikin (SAM) scale. Movie clips have been proven to be more effective in eliciting emotions for a longer time than images. Further details are presented by Carvalho et al. (2012).

1.3.2   Relevant Feature Extraction and Classification Methods

Various methodologies within feature extraction and classification methods for EEG signals have been explored widely lately. Some related works associated with feature extraction and emotion classification based on the EEG are presented here.

Different approaches for feature extractions have been proposed by researchers. These features are mainly in the time domain, and frequency domain, or have nonlinear characteristics. They are often extracted from EEG signals decomposed into frequency bands. The number of sub-bands varies between four (θ theta, α alpha, β beta, γ gamma) and five (δ delta, θ theta, α alpha, β beta, γ gamma). Commonly used features in the literature from the frequency domain are PSD and DE. It is proved that the PSD of EEG signals is correlated with human emotions (Koelstra et al., 2012; Miranda-Correa et al., 2021). In contrast to PSD, DE is known for its ability to differentiate between low and high-frequency energy (Duan et al., 2013; Zheng et al., 2019). In addition, other features, such as the Hjorth parameters (Joshi and Ghongade, 2022; Topic and Russo, 2021; Yuvaraj et al., 2023) from the

time domain and nonlinear features like fractal dimensions (Asadur Rahman et al., 2020; Topic and Russo, 2021; Yuvaraj et al., 2023), are also used in emotion recognitions.

For recognizing human emotions, various classification methods have been used. Deep learning-based methods have recently been developed into one of the most popular methods in the research field. Zheng and Lu (2015) introduced a method based on a deep belief network (DBN) to investigate the critical frequency bands of EEG signals. The convolutional neural network (CNN) has shown its potential in the generalization of classifying unseen subjects (Keelawat et al., 2021). In the work of Liu et al. (2020), a network combining the CNN, the sparse autoencoder (SAE), and the deep neural network (DNN) is proposed. The network is tested on the DEAP and SEED datasets and is shown as an efficient method with a faster convergence than the conventional CNN. Xiao et al. (2022) proposed a 4D-aNN architecture which captures discriminative patterns in EEG signals while fusing information on different domains. Another 4D-structured architecture, 4D-CRNN, is proposed by Shen et al. (2020). This method is capable of capturing frequency, spatial and temporal information from EEG signals.

However, in most of the proposed networks, the results are not compared to traditional machine learning techniques such as SVM and KNN. When working with deep learning methods a large amount of data is needed, which is not realistic when collecting EEG data (Moctezuma et al., 2022). Collecting EEG data is time-consuming, and doing so for many hours is demanding and unrealistic for participants. Kumar and Molinas (2022) explored simpler architectures with a focus on models based on CNN and MLP. Their research showed that simpler architecture can achieve performance comparable to the more complex deep learning architectures. In addition, Zheng and Lu (2015) and Asadur Rahman et al. (2020) have shown that common machine learning techniques such as SVM and KNN can get quite good performance with the correct feature combination.

1.3.3   Specialization Project

The specialization project[2] serves as a pre-project for the master's thesis. In the specialization project, different feature extraction techniques and classification methods were evaluated on the publicly available SEED dataset[3]. The dataset consists of recordings of 15 participants watching movie clips labelled as negative, positive and neutral emotions. The labels correspond to the horizontal axis (valence) of the valence-arousal emotion model.

The work explores three main types of classifiers, SVM, KNN and MLPs, with features, like HM, HC, PSD, KFD, HFD, and combinations of them, as input. The mean accuracies are the average performance accuracies across all 45 trials on 15 participants. In addition, the data samples were split into different subsets. For SVM and KNN, the data samples were split into a training set and a testing set of 80% and 20%. Data samples used in MLPs were split into a validation set besides the ones mentioned earlier. The ratio is 60%, 20% and 20% for the training, validation, and testing sets, respectively. In addition, a

---

[2]http://dx.doi.org/10.13140/RG.2.2.33244.67205
[3]https://bcmi.sjtu.edu.cn/home/seed/seed.html

5-fold CV and a 10-fold CV are used to verify the models. From the obtained results, it can be observed that HM and HC, extracted channel-wise from the epochs, achieve a good performance compared to other existing models on the same dataset. In addition, the performances of SVM and MLP-v2 are superior to KNN, with accuracies of 83.865% and 86.892%, respectively. In this work, the features and classifiers will be further evaluated and analyzed on a new dataset. The best-performing MLP, i.e. MLP-v2, will hereafter be referred to as MLP.

1.4   Outline Structure

This thesis is arranged as follows: An introduction of EEG-based emotion recognition and an overview of related works are presented in Chapter 1. Chapter 2 deals with relevant theory related to EEG. The experiment protocol and data collection methods are described detailed in Chapter 3, followed by descriptions of the different methodologies used to obtain the experimental results in Chapter 4. The results are presented in Chapter 5. And a discussion of the results is done in Chapter 6. The thesis is concluded in Chapter 7 followed by appendices.

2

Theoretical Background

This chapter[1] provides the reader with some relevant background knowledge considered usefully related to emotion recognition based on EEG. This includes information about emotions, EEG signals and their features and concepts, and common classification methods.

2.1 The Human Brain

The human brain is a part of the central nervous system and is responsible for controlling and coordinating various body functions, as well as cognition activities, e.g. memories, thoughts, perceptions, and emotions (Chen and Mehmood, 2020). The brain is split into three structures: the brain nucleus, the brain margin, and the cerebral cortex. Furthermore, the cerebral cortex, which is the outermost layer of the brain, can further be divided into four regions: the frontal lobe, parietal lobe, occipital lobe, and temporal lobe (Figure 2.1). The frontal lobe controls voluntary movements and is involved in various cognitive functions, including emotional expression and thinking. The parietal lobe is primarily responsible for processing sensory information, e.g. temperature, pain and touch pressure. The occipital lobe is located in the back of the skull and is responsible for receiving and processing visual information from the eyes. The temporal lobe is dedicated to auditory and memory-related functions (Liu et al., 2021). The lobes work together, along with other brain structures, to control people's daily behaviour and facilitate emotional experiences. Emotion-related activity in the brain can also be detected through specific brainwave patterns. These patterns can be linked to different emotional states (Knyazev, 2013). In addition, emotions are known to be associated with the activation of brain regions.

---

[1]This chapter is an updated version of the theoretical background chapter presented in the author's specialization project
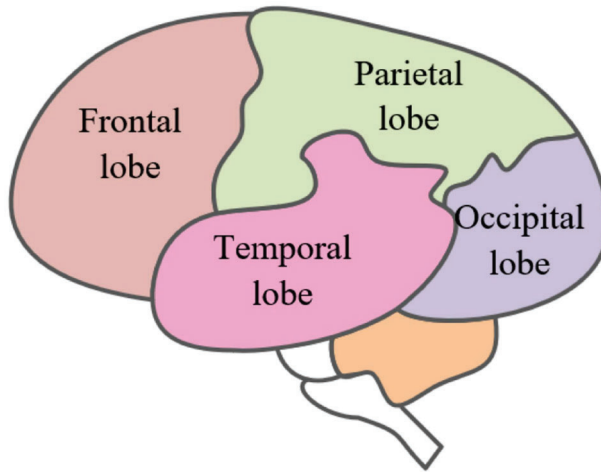
Figure 2.1: The regions of the cerebral cortex, including the frontal lobe, parietal lobe, temporal lobe, and occipital lobe (Liu et al., 2021)

2.1.1   Brain Frequency Bands

Brain activity is a response to stimuli in the central nervous system, which can be represented in different ways, where the oldest and most popular is its characteristic in frequency bands. The bands describe the repetitive waveforms of similar shape and duration (Sanei and Chambers, 2007). In general, five main sub-bands can be detected from low to high frequencies as follows: δ delta, θ theta, α alpha, β beta and γ gamma. Table 2.1 presents the associated mental state with the corresponding frequency bands and ranges. The degree of awareness increases with the increase of frequency range.

Table 2.1: An overview of the brain rhythms and their respective frequency range (Sanei and Chambers, 2007)

Brain rhythm	Frequency range [Hz]	Meaning
Delta δ	$0.5 - 4$	Deep sleep
Theta θ	$4 - 7.5$	Drowsiness
Alpha α	$8 - 13$	Relaxed awareness
Beta β	$14 - 26$	Active thinking
Gamma γ	> 30	Deep focus

2.2 Electroencephalogram

Electroencephalogram (EEG) was discovered by Hans Berger (1873 - 1941) almost a century ago and expresses indirectly the electrical activity of the brain (Schomer and Lopes da Silva, 2017). The activity is recordable by electrodes if the active neurons inside the brain can generate enough potential. Raw EEG signals are considered nonlinear and nonstationary by nature due to the different properties of the head layers.

To acquire EEG signals, a non-invasive, lightweight device is being used. Some of the advantages of using an EEG device are the relatively low cost and it can easily be used without complex lab setups. Compared to other more space-consuming and complex methods, such as magnetic resonance imaging (MRI) and computed tomography (CT), EEG devices are portable, relatively cheap and less space- and time-consuming (Stancin et al., 2021). However, a limitation of EEG measurement is the low spatial resolution which makes it hard to extract meaningful information just by observation. Furthermore, it is easily exposed to noise and artefacts. They are often caused by the power source, muscle movements, eye blinks, heart rate, and the surrounding environment (Sanei and Chambers, 2007).

2.2.1 Electrode Placement

To record the EEG signals, electrodes placed on the head are used. Each electrode represents a channel. These can be placed in numerous ways, depending on the area of interest. However, the most common approach is placing them according to the international 10-20 system proposed by the International Federation of Clinical Neurophysiology (Klem et al., 1999). The numbers refer to the percentage of distances between neighbouring electrodes of the total front-back or right-left distance of the brain. An extension of the 10-20 system, known as the international 10-10 system, also exists (Chatrian et al., 1985; Oostenveld and Praamstra, 2001). This system has a higher density of electrodes which makes it easier to measure the brain signal in a specific region. The electrode placement used in this experiment is illustrated in Figure 2.2.
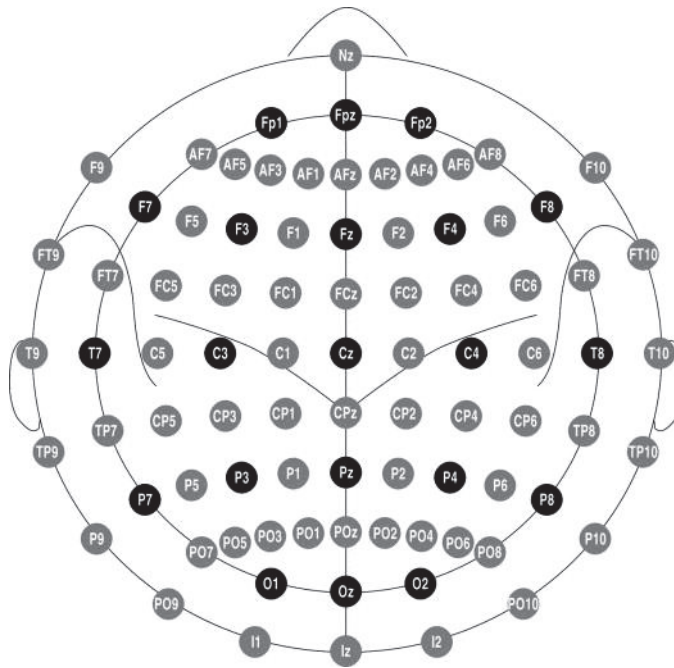
Figure 2.2: Electrode placement and labels in the international 10-20 and 10-10 systems. Black dots indicate the original 10-20 system, grey dots indicate additional positions introduced in the 10-10 system (Oostenveld and Praamstra, 2001)

2.3 Emotion Model

In the field of psychology, human emotions have been vastly researched. Emotions are typically pretty subjective and are often associated with the personality, mood, and desire of a person (Liu et al., 2021). With an emotion model, the different emotions are somehow categorized, and a more objective insight is given into emotions. The emotion models include discrete models and dimensional models.

The discrete model suggests that emotions are composed of basic emotions and that all emotions can be formed through a combination of one or more of these. The most known of these models is the Six Basic Emotions presented by Ekman (1992) consisting of the emotions: anger, disgust, fear, joy, sadness, and surprise, There is no agreement about the precise number of basic emotions. Plutchik (2001), for example, proposed eight bipolar emotions as the basics, and they are as follows: joy, sorrow, anger, fear, acceptance, disgust, surprise, and expectancy. He also proposed a dimensional description of the basic emotions with the emotion wheel. The dimensional model views emotions as a set of dimensions varying independently. Russell (1980) proposed that emotions can be represented as a two-dimensional coordinate system with axes representing the degree of valence and arousal (Figure 2.3). The horizontal axis (valence) represents positive and negative

stimuli, i.e. misery and pleasure. While the vertical axis (arousal) describes the level of intensity a stimulus gives with a range from uninterested and sleepy to alarmed and excited. A three-dimensional emotion model with a dominance dimension, in addition to valence and arousal, also exists (Koelstra et al., 2012). The dominance dimension ranges from submissive to empowered. In the dimensional emotion models, emotions can be represented by different coordinate positions.
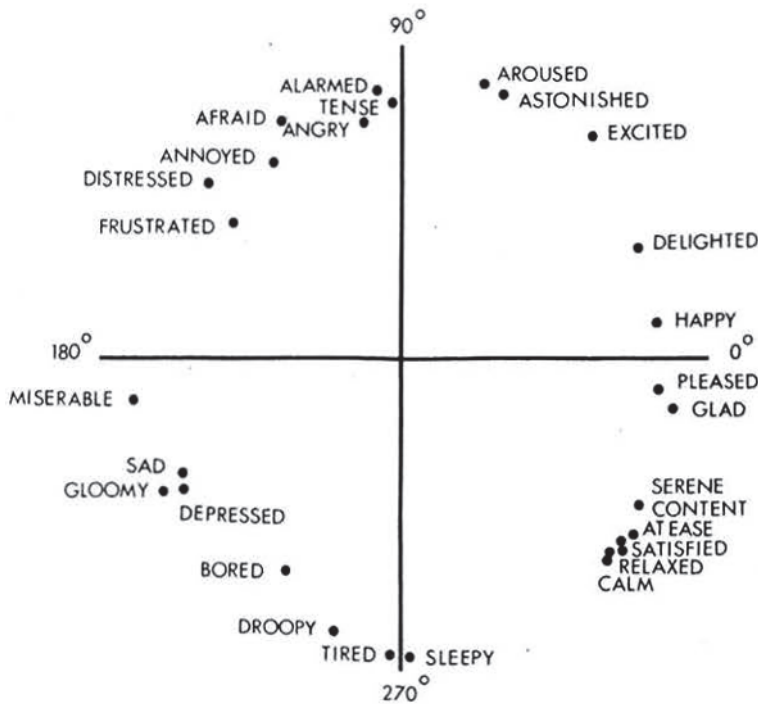
Figure 2.3: Two-dimensional emotion model of valence and arousal (Russell, 1980)

2.4 Feature Extraction

The main goal of feature extraction is to separate and transform relevant parts of a signal from irrelevant components and express them in a compact or meaningful form (Liu et al., 2021). This process can result in improvements in accuracy and reduce overfitting. It might also increase the understanding of a model due to the decrease in the complexity of the signals. Features can be extracted from different domains. Typical features extracted from EEG signals are from the time domain and the frequency domain, i.e. mean, standard deviation, Hjorth parameters, and PSD (Stancin et al., 2021). In addition, there are features with nonlinear characteristics such as fractal dimension.

2.5 Machine Learning

Machine learning (ML) is a method that can extract patterns from raw data and increase its knowledge (Goodfellow et al., 2016). It uses training data and the theory of statistics to build mathematical models and algorithms to enable us to handle complex tasks. Typical tasks for ML models are classification and regression. In classification, the model is asked to categorize where some input belongs to. In regression, however, a numerical value is asked to be predicted based on the input data. These two tasks are quite similar except for the output.

It is common to divide ML into three subcategories: supervised learning, unsupervised learning, and reinforcement learning (Goodfellow et al., 2016). Supervised learning allows the models to train and learn from labelled datasets, while for unsupervised learning the objective is to find patterns or trends in unlabelled data. Reinforcement learning learns through the trial and error method where the so-called best action gets some reward.

2.5.1 Deep Learning

Deep learning is a subset of ML. It consists of neural networks with more than two hidden layers which abstract different features (Sreeshakthy et al., 2016). The networks imitate the learning process and the intelligent behaviour of the human brain. Moreover, the number of layers and the ability to do complex things are correlated. This allows the computer to build complex models out of simpler ones. One downside of deep learning is that it requires large amounts of data and is quite computationally expensive to run, depending on the model size.

3

Experiment Design and Data Collection

In this chapter, the experimental setup and methods used are described. Some information about the experimental protocol, including the selection of the recording system, stimuli and channels, is mentioned here. A thorough description of the SAM scale and software used are also provided.

3.1   Materials and Methods

Before the experiment, each participant signed a consent form (Appendix C) and got information about the upcoming procedure. They were also warned about the potentially shocking scenes from some movie clips and were reminded that they were free to discontinue participation at any time. The experiments were performed in a quiet environment with controlled illumination during the daytime. At least one experimenter was present during the experiments in case of questions.

There are four main steps in each experiment, with a total of 52 trials:

1. Introduction to the experiment

2. General instructions on how to rate the questionnaire

3. SAM instructions (Figure 3.3)

4. Fifty-two trials (Figure 3.7)

The movie clips were presented on a Dell UltraSharp 38 Curved Monitor (3840 x 1600). Participants were instructed to sit comfortably and limit their movements while watching movie clips to avoid the production of unrelated artefacts. EEG signals were recorded continuously with a sampling rate of 250 Hz. Figure 3.1 showcase the experiment scene. To protect the privacy and confidentiality of participants, one of the experimenters is present in the photo.

Figure 3.1: A photo of the experiment scene during recordings

3.1.1   EEG Recording System Selection

Deciding whether to use dry or wet electrodes for data acquisition can be challenging. A recent study concluded that wet electrodes had a slightly improved signal quality (Kam et al., 2019). Hence, the plan was to use the Explore+ device from Mentalab[1]. It has 32 flat electrodes placed according to the international 10-20 system. Conductivity gel is applied on the electrodes to increase conductivity between the electrodes and the skin. However, due to delayed delivery and connectivity issues, two devices from Unicorn[2] were used to collect the EEG data instead. Each Unicorn device captures EEG signals with eight dry electrodes of conductive rubber. The electrodes were put on the g.tec's g.GAMMAcap[3] which is based on the 10-10 system (Figure 3.2). An overview of the used EEG channel names and which device they belong to is given in Table 3.1.

---

[1]https://mentalab.com/
[2]https://www.unicorn-bi.com/
[3]https://www.gtec.at/

Figure 3.2: The Unicorn EEG headset (left) and the g.tec's g.GAMMAcap

Table 3.1: Overview of EEG channel names according to the international 10-10 system

Device number	Channel number	Channel name
68	1	FP1
	2	FPz
	3	F5
	5	FC1
	5	CP1
	6	P1
	7	PO3
	8	PO5
69	1	FP2
	2	F8
	3	F4
	5	FC6
	5	FC4
	6	CP6
	7	P8
	8	Cz

3.1.2   Self-Assessment Manikin

Self-assessment manikin (SAM) is used for rating the elicited emotions after each clip. It is an inexpensive, pictorial assessment technique to facilitate measurements of valence, arousal and dominance (Bradley and Lang, 1994). The SAM scale description is depicted in Figure 3.3. The valence dimension is represented by five figures ranging from a frowning, unhappy figure to a smiling, happy figure. For the arousal dimension, the scale ranges from a relaxed, unaffected figure to a passionate, wide-eyed figure. For the dominance dimension, the figure ranges from submissiveness and feelings of control to dominance and power. All dimensions are accompanied by a score between 1 to 9, which allows the

participant to rate the elicited emotions more finely.
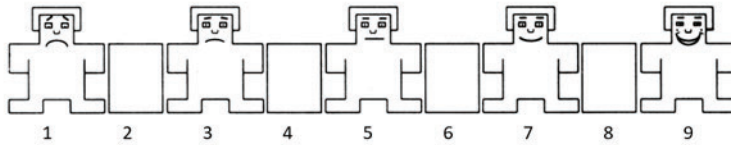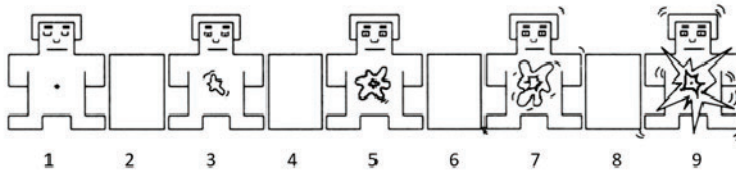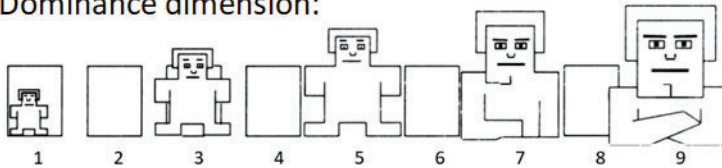
Figure 3.3: Self-assessment manikin (SAM) used to rate dimensions of valence, arousal, and dominance (Bradley and Lang, 1994)

Emotions are examined in terms of two dimensions in this thesis: valence and arousal. The two-dimensional valence-arousal emotion model can be split into four classes, i.e. low valence (LV), high valence (HV), low arousal (LA), and high arousal (HA). On the 9-point SAM scale, the threshold is placed in the middle, as reported by Koelstra et al. (2012). Following Betella and Verschure (2016), if a participant's rating on valence or arousal is greater than or equal to the value of 4.8 the rating is considered as HV or HA. However, if the rating is less than 4.8, the rating is considered LV or LA. This may lead to unbalanced classes, which are taken into account by reporting the average F1-score, along with the average accuracy. The distribution of HV/LV ratings and HA/LA ratings from the subjects are shown in Figure 3.4.
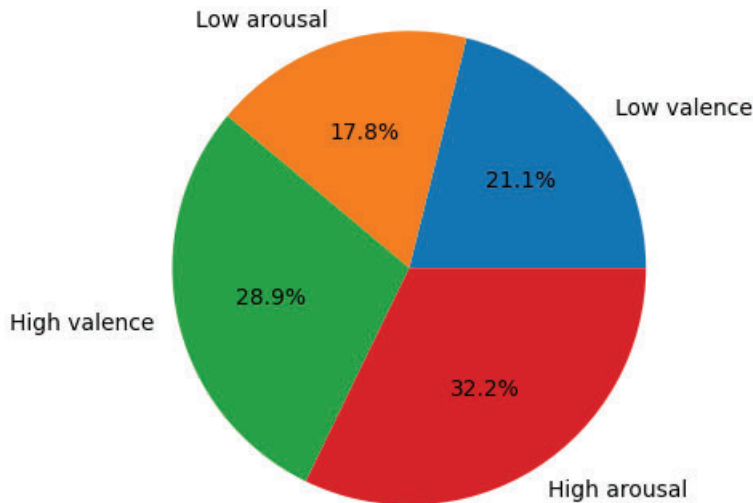
Figure 3.4: Distribution of high/low valence/arousal rated by 19 participants

3.2   Emotion Elicitation

The non-auditory stimuli selection is designed at Tsukuba University Human Sleep Lab[4] by PhD candidate Felix Ipanaque and its supervision team. Fifty-two movie clips of around 40 seconds each are selected to elicit emotions. Figure 3.5 shows examples of movie clips used in the experiment. To target the emotions of all four endpoints of the valence-arousal space, clips from six categories were present: horror, erotic, positive social content, negative social content, scenery and object manipulation (Carvalho et al., 2012). In the horror category, movie clips of life-threatening and gruesome situations are shown. The erotic category includes clips with couples engaged in sexual intercourse. The social positive and social negative categories show clips of happy social interactions and sad or angry interactions without horrifying situations, respectively. The scenery category shows clips of landscapes, e.g. waterfalls, mountains, the aurora borealis, etc. And the object manipulation category includes two clips showing a hand moving white objects around on a table.

---

[4]https://www.u.tsukuba.ac.jp/ abe.takashi.gp/laboratory.html

(a) Horror

(b) Erotic

(c) Social positive

(d) Social negative

(e) Scenery

(f) Object manipulation

Figure 3.5: Screenshot of movie clip samples from all six categories: (a) horror, (b) erotic, (c) positive social content, (d) negative social content, (e) scenery, and (f) object manipulation

3.3 Channel Selection

The EEG electrodes were placed according to the international 10-10 system with reference electrodes on both ear lobes (Figure 3.6). The positions of the electrodes on the scalp are FP1, FPz, FP2, F5, F4, F8, FC1, FC4, FC6, CP1, CPz, CP6, P1, P8, PO5, and PO3 inspired by Pane et al. (2018)'s channel selection for 15 electrodes. Pane et al. (2018) used stepwise discriminant analysis (SDA) with Wilks lambda as the selection criteria to find the optimal channels. Furthermore, their research confirms that higher frequency bands

such as α, β, and γ are more related to emotion recognition, which fits well with the results from Zheng and Lu (2015). EEG signals are obtained from 16 electrodes (eight electrodes in each system).
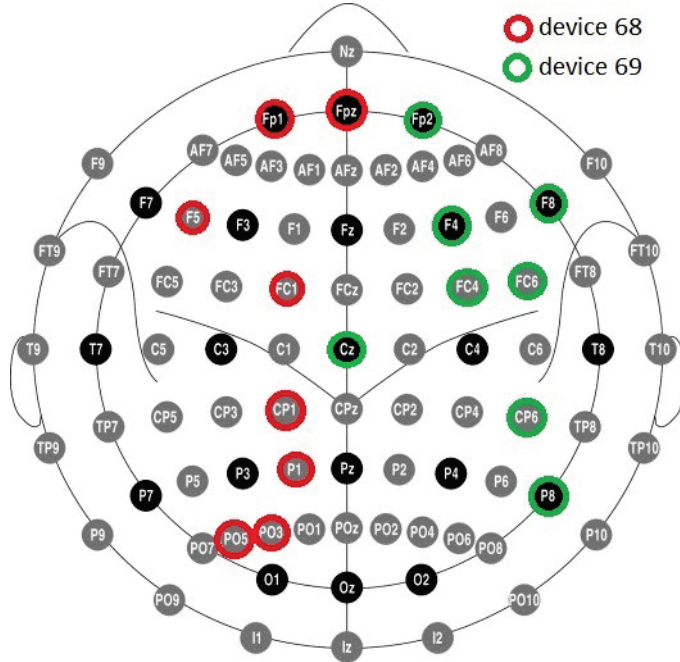
Figure 3.6: The selected electrodes placed according to the international 10-10 system

3.4 Participants

Twenty university students with general good health participated in the experiments. According to Suhaimi et al. (2020) such studies should have at least ten participants to have reliable and meaningful results statistically. Both biological genders (ten males and ten females) were represented with an age of 23.80 ± 0.87 years. All participants are students or recent graduates from the Norwegian University of Science and Technology (NTNU). Participants with known neurological diseases and heavy use of medicine or drugs are excluded from the experiment, to make the results generalizable to the larger population. All participants also had to give their informed written consent and be above 18 years before being included in the experiments. In the consent form, the participants were informed about the procedure and got an introduction to SAM. They received a gift card as a small compensation for their time. Only data from 19 participants (ten males and nine females) are used for emotion recognition due to some missing recordings from one of the participants.

3.5   Experiment Protocol

One session lasts about 75 minutes, where approximately 35 of these minutes are for the collection of emotion-elicited EEG signals using the Unicorn system. A detailed protocol is shown in Figure 3.7. There are a total of 52 trials in one experiment. The subject participated only in one experiment each. Before each movie clip, there is an 11-second hint in the form of a fixation cross. The period with fixation cross can work as a baseline for future studies. After each movie clip, participants are asked to perform a self-assessment to rate their elicited emotions according to the SAM scale description. To weaken any carry-over effects between movie clips, a rest of 15 seconds are followed before a new trial. Synchronization markers are sent from the stimuli presenter to the EEG recorder to mark the beginning and end of each fixation cross, in addition to the beginning and end of each movie clip.
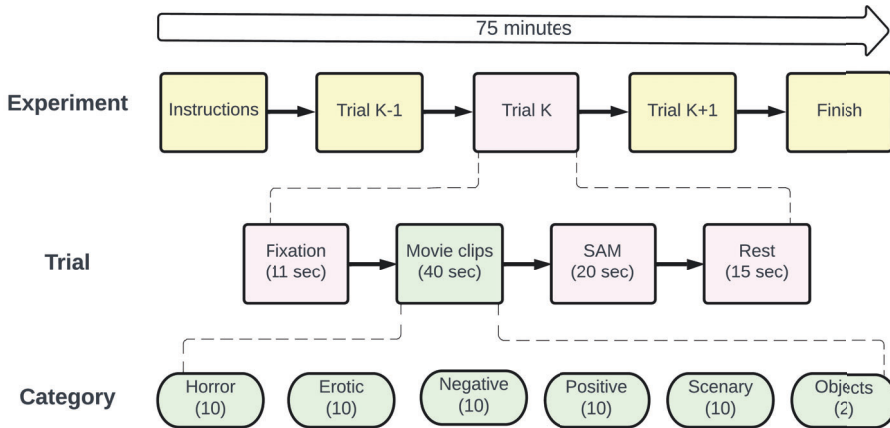
Figure 3.7: Protocol of the EEG experiment

4

Methods

This chapter[1] provides the reader with relevant methods considered usefully related to EEG signals and emotions. The methods for preprocessing the raw EEG signals will be introduced first, followed by details of feature extraction and classification methods.

4.1 Preprocessing Techniques

The human brain is a complex system. The signal's characteristics change with time and the influence of external stimuli. To improve the classification performances of emotions and get closer to the true signals, the acquired EEG signals should be denoised and filtered, since the EEG signals are known for being non-stationary, nonlinear and noisy (Klonowski, 2009). Preprocessing aims to maximize the signal-to-noise ratio (SNR) to get cleaner data.

4.1.1 Epoched Data

The raw EEG data are recorded continuously with synchronization markers to identify timestamps for the start and end of each movie clip. Data that do not contain an emotional state, e.g. breaks, fixation cross, SAM, etc., are cropped out. The data is then split into the same-length epochs for each channel with the dimensions of (epoch x channel x segment size). Different segment sizes of each epoch without overlapping are explored: 5 seconds, 10 seconds, 20 seconds, and 40 seconds with a sampling rate of 250 Hz.

---

[1]This chapter is an updated version of the materials and methods chapter presented in the author's specialization project

4.1.2   Filtering Raw EEG Signals

Typical noises found in EEG signals are background noises generated from the power lines at 50 Hz (in Norway). This type of noise can be detected by the measurement system. The power line noises in the EEG signals are suppressed with a notch filter at 50 Hz and 100 Hz. Following Zheng and Lu (2015), a bandpass filter between 0.3 Hz to 50 Hz is further applied to reduce more noises and artefacts. Figure 4.1 shows the PSD of the obtained EEG signal from participant 11 in different stages: the original signal, only the emotion-elicited signal, and the filtered signal. The filtered signal is used for feature extraction and classification.
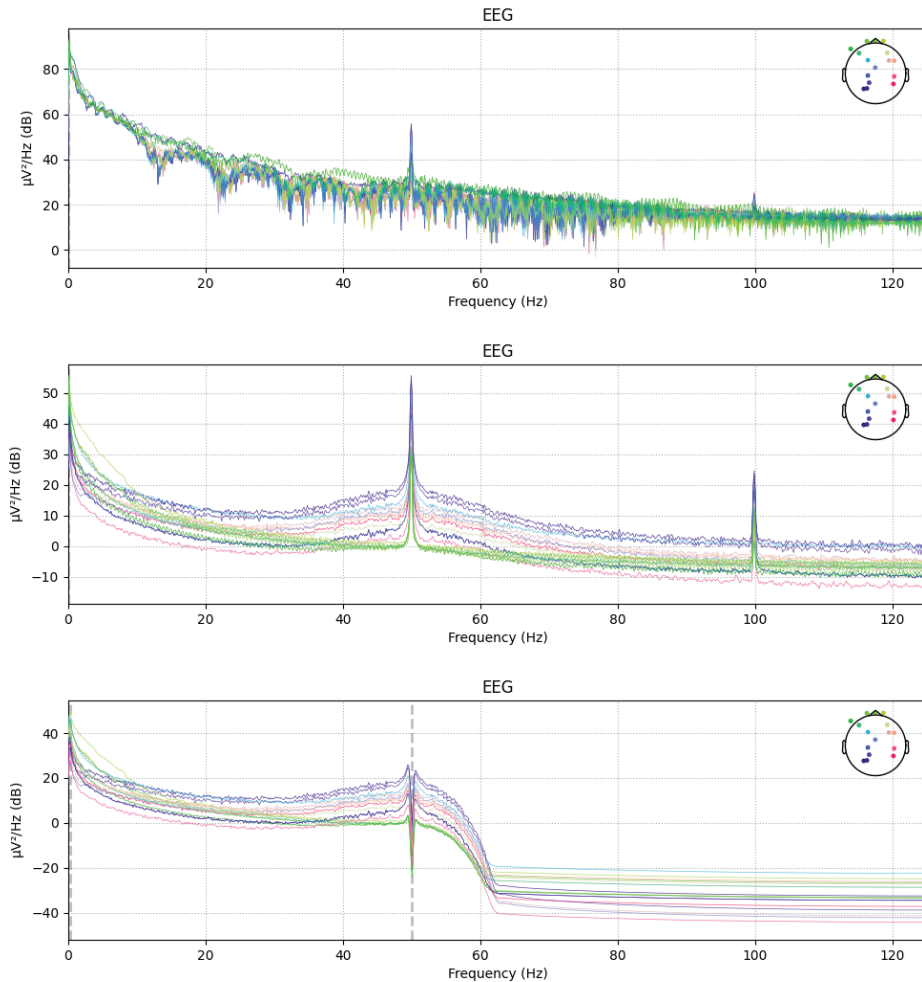
Figure 4.1: The PSD of participant 11's original signal (top), the PSD of only emotion-elicited EEG data, and the PSD of the EEG signal after applying a notch filter and a bandpass filter

4.1.3   Ocular Artefact Removal with ICA

Physiological noises and motion artefacts, such as ocular (eye) artefacts and muscle contractions, can also be captured by the brain and thus contaminate the obtained EEG signals. Based on a literature review on artefact removal from EEG signals conducted by Jiang et al. (2019), blind source separation algorithms, especially Independent Components Analysis (ICA), are shown to be the most popular techniques among researchers. One advantage of ICA is the ability to decompose the EEG signal into different individual components.

The separation process makes it convenient for artefact identification and removal (Mennes et al., 2010). It is quite simple to identify the ocular artefacts by visualizing each ICA component with a topography map. They are most visible around the front of the head, among the front electrodes above the eyes. In addition, the phase spectral density of the components has a peak at the low-frequency end of the spectrum (Newman et al., 2021). Figure 4.2 is an example showing common properties of an ocular artefact. In this thesis, ICA is used to remove the ocular artefacts and preserve the brain activity of interest. The correction is done separately for each subject to overcome individual differences between them. The used method is Picada (Ablin et al., 2018) due to its fast convergence.
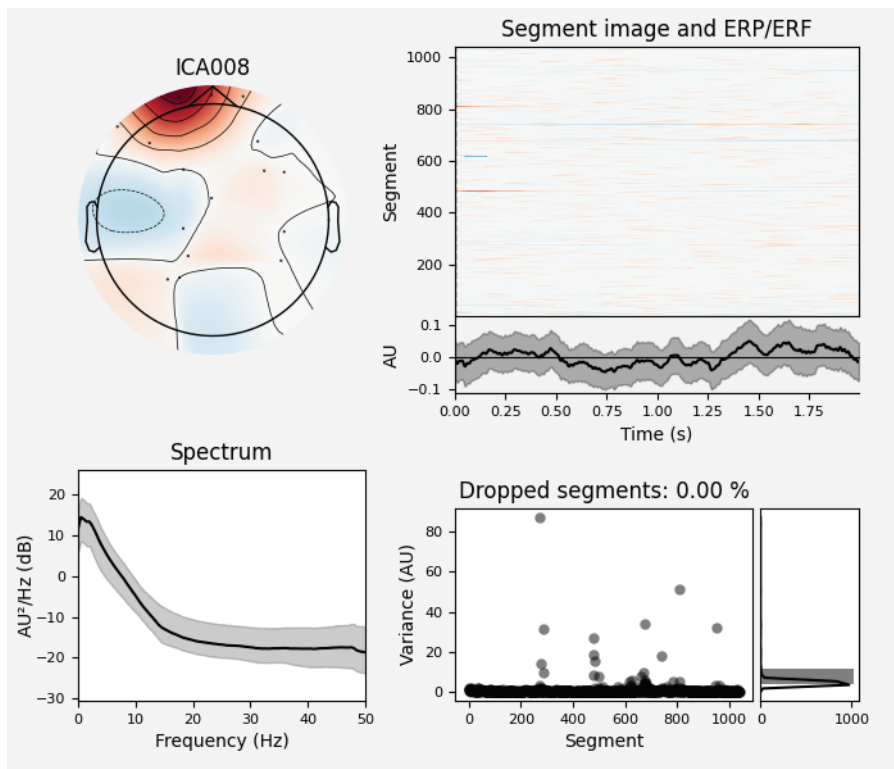
Figure 4.2: Properties of ocular artefacts in an ICA component. Clockwise from top left: topography map, raster plot, epoch variance and phase density spectrum

4.2   Feature Extraction Techniques

The feature extraction process can simplify the model and reduce overfitting (Liu et al., 2021). Feature computation is calculated in two different ways: directly on each of the epochs (channel-wise) and from signals decomposed with a five-order Butterworth band-pass filter into the five frequency bands: $\delta, \theta, \alpha, \beta$, and γ. Six different main methods for feature extraction are proposed in this thesis, both from five sub-bands and channel-wise (Table 4.1).

Table 4.1: Details of feature extraction methods

Feature extraction method	Segment size	Feature dimension
Channel-wise (CW) [DE, HM, HC, KFD, HFD]	40 seconds	(52,16)
	20 seconds	(104, 16)
	10 seconds	(208, 16)
	5 seconds	(416, 16)
Five sub-bands (SB) [DE, HM, HC, KFD, HFD, PSD]	40 seconds	(52, 80)
	20 seconds	(104, 80)
	10 seconds	(208, 80)
	5 seconds	(416, 80)

4.2.1   Differential Entropy

Differential entropy (DE) is an extension idea of Shannon entropy and is used to measure the complexity of a continuous random variable. It has been recognized to be suitable for emotional recognition (Duan et al., 2013; Shen et al., 2020). The calculation formula is defined as

$$DE = h(X) = -\int_X f(x)log(f(x))dx \stackrel{\text{Gauss distribution}}{=} y\frac{1}{2}log(2\pi e\sigma^2), \qquad (4.1)$$

when X is the random variable and $f(x)$ is the probability density function of X.

4.2.2   Fractal Dimension

Fractal dimension (FD) is estimated based on an underlying high degree of the geometric structure of the signal (Stancin et al., 2021). Since EEG signals are nonlinear by nature, FD can measure the signals' complexity and irregularity. Two well-known approaches are the Katz fractal dimension (KFD) and the Higuchi fractal dimension (HFD) (Yuvaraj et al., 2023; Topic and Russo, 2021).

Katz Fractal Dimension

KFD is based on the ratio of the total length of a curve and is used to estimate the FD of a time series. It is calculated as the sum of the distances between two successive points, normalized by the mean distance of successive points (Katz, 1988):

$$KFD = \frac{log(L)}{log(d)} = \frac{log(N)}{log(N) + log(\frac{d}{L})},$$ (4.2)

where L denotes the curve length, d represents the distance between the two successive points, and N is the number of time samples in the EEG epoch.

Higuchi Fractal Dimension

Higuchi's method depends on the length of the irregular curve from a time series consisting of a set of points. The curve's length is calculated and averaged across samples. This process is repeated for different scales and plotted in a graph. The best-fitted slope of the graph is the HFD of the time series (Higuchi, 1988). For a finite set of N time series samples $X(N) = X(1), X(2), X(3), ..., X(N)$, a newly constructed time-series signal is defined as

$$X_i^j = X(i), X(i+j), ..., X\left(i + \left[\frac{N-i}{j}\right] \cdot j\right),$$ (4.3)

where $m = 1, 2, ..., j$ indicates the initial time and j is the interval time. HFD is then defined as

$$HFD = \frac{<L(j)>}{log(j)}, \quad L_i(j) = \frac{\sum_{k=1}^{[\frac{N-i}{j}]} |X(i+kj) - X(i+k-1)| \cdot (N-1)}{k \cdot [\frac{N-i}{j}]}$$ (4.4)

where $L(j)$ is the length of the curve.

4.2.3   Hjorth Parameters

Hjorth parameters are used for time-domain analysis and give an insight into statistical properties of the EEG signal (Hjorth, 1970). These parameters are activity (HA), mobility (HM) and complexity (HC), and are defined as follows:

$$\text{activity} = \delta_0^2; \text{ mobility} = \sqrt{\frac{\delta_1^2}{\delta_0^2}}; \text{ and complexity} = \sqrt{\frac{\delta_2^2/\delta_1^2}{\delta_1^2/\delta_0^2}},$$ (4.5)

where δ_0^2 is the variance of EEG signal (X_t), δ_1^2 is the first derivative of X_t and δ_2^2 is the second derivative of X_t. Activity corresponds to the variance of the signal, while mobility and complexity are based on the variance of the derivatives of the signal (Stancin et al., 2021). Only HM and HC will be used in this thesis as they are more commonly used (Joshi and Ghongade, 2022; Topic and Russo, 2021).

4.2.4 Power Spectral Density

Power spectral density (PSD) shows the average energy distribution per unit of time over different frequency bands. To estimate the PSD of signals, the original signal is turned into a power spectrum that changes with frequency (Liu et al., 2021). The function is defined as

$$P(\omega) = \sum_k = -\infty^\infty r(k)e^{-j\omega k},\tag{4.6}$$

where $r(k)$ is autocorrelation function. In this thesis, the default frequency band values from the Python library *MNE-features*[2]:

$$[\delta,\ \theta,\ \alpha,\ \beta,\ \gamma] = [0.5,\ 4,\ 8,\ 13,\ 30,\ 50]$$

are used in the computation. In addition, the Welch method is used for estimation.

4.2.5 Combinatorial features

Combining different signal properties gives a new set of features, which can further improve the model performance by feeding the classifiers more information. Qin et al. (2019) attempts to combine correntropy spectral density (CSD) and PSD with promising results. The results demonstrate that the combination outperforms CSD and PSD separately. Jacob et al. (2021) showed that combining different features can reduce the weaknesses of individual features while integrating the strength.

Features from the time domain, frequency domain, and also nonlinear features have been combined and explored in this work. The feature combinations explored are as follows:

- HC and HM (channel-wise, only time-domain features)
- DE and PSD (sub-bands, only frequency-domain features)
- HFD and KFD (channel-wise, only nonlinear features)
- DE (sub-bands) and HM (channel-wise)
- All features: HC, HM, HFD, KFD (channel-wise) and DE, PSD (sub-bands)

---

[2]https://mne.tools/mne-features/index.html

4.3   Classification Methods

The selected features are sent to classifiers to evaluate the classification performance. Popular machine learning classifiers are support vector machine (SVM) and K-nearest neighbour (KNN). A relatively simple multilayer perceptron (MLP), proposed in the specialization project, is also used for emotion classification. The classifiers SVM, KNN, and MLP from the project are used to test their generalization to new, unseen data. Table 4.2 summarises the parameter details used in classifiers. All the models are subject-independent, which means they are trained and tested across subjects.

Table 4.2: Details of parameters used in the different classifiers

Classfiers	Parameter details
KNN	K = 5 Train: 80% — test: 20% OR 5-fold CV and 10-fold CV
SVM	Kernel: Radial basis function (RBF) Decision function: One-vs-One (ovo) Train: 80% — test: 20% OR 5-fold CV and 10-fold CV
MLP	Structure with two hidden layers (300 and 500 nodes) Dropout rate: 0.2 Activation functions: ReLU and softmax Train: 60% — test: 20% — validation: 20% OR 5-fold CV and 10-fold CV

4.3.1   K-Nearest Neighbour

K-nearest neighbour (KNN) is a supervised classifier (Figure 4.3). The prediction is made by using a majority vote and some distance measurements between the given test sample and k training samples (Sreeshakthy et al., 2016). The most frequently represented label around the point is used to label a given data point. Following Zheng and Lu (2015) and Asadur Rahman et al. (2020), K = 5 will be used in this thesis.

4.3.2   Support Vector Machine

The goal of a support vector machine (SVM) is to find an optimal hyperplane that maximizes the separation between classes (Figure 4.4). The distance from the hyperplane to the nearest data point on both sides should be maximized (Liu et al., 2021). It is a powerful algorithm and is commonly used as a classifier for extracted features. Recent studies have

shown that the radial basis function (RBF) is the best performer, especially when compared to the other available kernel functions of SVM, e.g. linear, sigmoid, and polynomial (Yao et al., 2021; Topic and Russo, 2021; Yuvaraj et al., 2023). Thus, the RBF kernel will be used in this thesis.
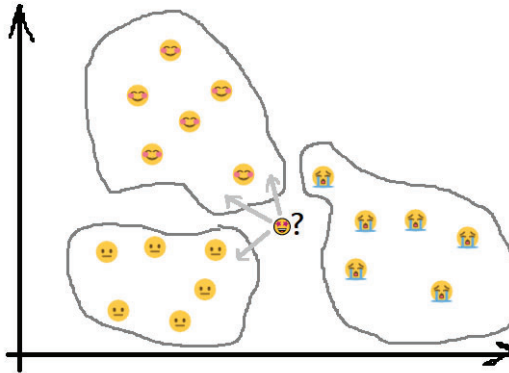
Figure 4.3: An illustration of the K-nearest neighbour method (specialization project)
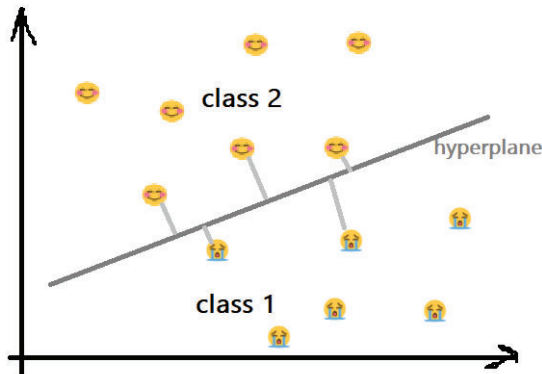
Figure 4.4: An illustration of the support vector machine method (specialization project)

4.3.3   Multilayer Perceptron

Multilayer perceptron (MLP) is the simplest version of DNN consisting of only a few numbers layers (Figure 4.5). Generally, it has only an input layer, an output layer, and one hidden layer in between. However, MLP may have more than one hidden layer (Yang et al., 2018). It is widely used as a tool in various classification problems, including emotion classification.

For MLP, the best model of the MLP architectures from the author's specialization project is tested. MLP consists of two hidden layers with 500 and 300 nodes. A dropout regularization of 0.2 is also applied. The activation function is rectified linear unit (ReLU), and the output is fed into a softmax classifier for emotion recognition. The loss function and the optimizer are set to the categorical cross-entropy loss and the Adam optimizer, respectively. Inspired by Kumar and Molinas (2022), the number of layers and nodes are selected based on the trial and experimentation method.
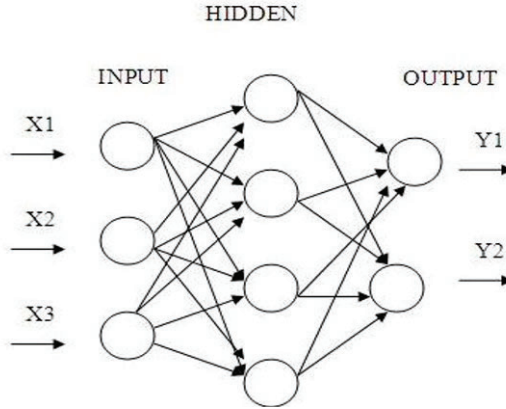
Figure 4.5: A demonstration of a basic multilayer perceptron (Sreeshakthy et al., 2016)

4.4   Evaluation Methods

To examine the robustness and stability of a model, it is important to evaluate how well it recognizes emotions in terms of LV/HV and LA/HA. All classification methods are evaluated by the same metrics, namely accuracy and F1-score, in addition to standard deviation and K-fold cross-validation (CV). For a binary classification problem, these metrics evaluate the models based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are visually represented using a confusion matrix (Table 4.3). A confusion matrix summarizes how many predictions are correct and incorrect for each class in a matrix form (Sokolova et al., 2006). It can help in understanding the model's behaviour and trade-offs between evaluated metrics. In this work, negative classes correspond to LV and LA, whereas positive classes correspond to HV and HA.

Table 4.3: Confusion matrix for binary classification. Green and pink depicts instances predicted as negative (TN+FN) and positive (TP+FP), respectively

	Predicted negative	**Predicted positive**
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

4.4.1   Accuracy

Accuracy is one of the most used empirical measures and evaluates the overall correctness of a machine learning model. It provides a general overview of how well the model is performing. However, it does not provide information about the distribution of correct labels across classes (Sokolova et al., 2006). Mathematically, accuracy is calculated as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.7}$$

4.4.2   F1-Score

The F1-score shows the harmonic mean of precision and sensitivity (Equation 4.8). It ranges between 0 and 1, where a higher score indicates better performance. F1-score is useful for unbalanced classes when there is a significant difference between the number of instances in each class (Sokolova et al., 2006).

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8}$$

where $Precision = \frac{TP}{TP+FP}$ reflects the model's ability to identify positive instances without many false positives, and $Recall = \frac{TP}{TP+FN}$ is the ratio of correctly predicted positive instances to the total number of positive instances,

4.4.3   Standard Deviation

Standard deviation (σ) measures the variation in a dataset and how spread out the values are from the average value. A high standard deviation indicates large differences in performance among individuals. The mathematical formula is given as

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N - 1}} \tag{4.9}$$

where x_i is each individual sample, $\bar{x}$ is the average of the samples, and N is the total number of samples.

4.4.4   K-Fold Cross-Validation

K-folds cross-validation (CV) is an effective method to evaluate the classification methods, especially when the data samples are limited. In this thesis the choice of $K = \{5, 10\}$ is applied due to its popularity among researchers (Tao et al., 2023; Xiao et al., 2022; Topic and Russo, 2021; Shen et al., 2020; Yang et al., 2018). The value of K corresponds to the number of iterations and the amount of equally-sized folds of the dataset, and must not be

confused with the K used in KNN. During each iteration, one fold is used for testing, and the remaining folds are considered for training. Only some of the best-performing models are verified with K-fold CV. The average performance of accuracy and average F1-score over K iterations are reported.

4.5   Software Used

Most of the code produced for this thesis is implemented with Python. One advantage of using Python as a programming language is the availability of extensive libraries useful for signal analysis and machine learning.

The following Python libraries and modules have been used:

- PsychoPy software[3]: Present the stimuli to the participants

- MNE library[4]: Preprocessing of raw EEG signals, including filtering (notch and bandpass) and artefact removal (ICA)

- MNE-Features module[5]: Calculate feature extraction

- Scikit-Learn[6]: Implementation of traditional machine learning methods (SVM and KNN)

- Keras[7] and Tensorflow[8] libraries: Implementation of deep learning method (MLP)

The Unicorn Hybrid Black headset and its software environment are used for EEG signal acquisition. For recording purposes and synchronization of the EEG systems with the stimuli presentation, the Lab streaming layer[9] (LSL) is used. In Appendix A, a link to the code used in this thesis is available.

---

[3]https://www.psychopy.org/
[4]https://mne.tools/stable/index.html
[5]https://mne.tools/mne-features/api.html
[6]https://scikit-learn.org/stable/
[7]https://keras.io/
[8]https://www.tensorflow.org/
[9]https://labstreaminglayer.org/

5

Experimental Results

The chapter presents the classification performances from the experiment design and setup presented in Chapter 3 by using the methods from Chapter 4. The average accuracy and the average F1-score across all participants are reported. The best-performing feature combinations and classifiers are evaluated using 5-fold and 10-fold CV. Confusion matrices and topography maps from participants are also presented.

5.1 Classification Performance

Best-performing classification results for each individual feature classified with SVM, KNN and MLP are presented here. So the poorest reported scores here are among the model performances. All models are subject-independent and are trained on data from multiple subjects. The classification is binary assigned with the labels high/low arousal or high/low valence. The average accuracy of a model is the average of the performance accuracies of all 19 participants. The average value of all the participants' F1-scores is also reported to take unbalanced classes into account.

5.1.1 Individual Features

The individual features consist of DE, HM, HC, KFD, HFD, and PSD extracted channel-wise and from sub-bands. Table 5.1 and Table 5.2 show only the average accuracies and average F1-scores for the best-performing subject-independent models for each feature, respectively. A more detailed overview of the classification results can be found in Appendix D and Appendix E.

Table 5.1: Details, average accuracies and average F1-scores of individual features for the overall best-performing models for each feature for high/low arousal

Arousal (calm-excited)					
Feature	**Method**	**Classifier**	**Epoch size**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	5sec	67.607 ± 11.207	72.410 ± 19.229
HM	Channel-wise	SVM	5sec	66.792 ± 10.961	72.817 ± 17.836
HC	Channel-wise	SVM	10sec	66.507 ± 15.467	70.355 ± 22.493
KFD	Channel-wise	SVM	5sec	66.040 ± 11.255	72.531 ± 17.180
HFD	Channel-wise	SVM	5sec	67.607 ± 11.284	73.167 ± 18.716
PSD	Five sub-bands	KNN	5sec	65.727 ± 11.655	71.159 ± 15.914

The highest average accuracy of 67.607% (± DE: 11.207%; ± HFD: 11.284%) for high/low arousal is observed for DE and HFD extracted channel-wise from 5-second long epochs using the SVM classifier. The average F1-scores for the mentioned models are 72.410% ± 19.229% and 73.167% ± 18.716%, respectively. The lowest average accuracy of 65% ± 11.665% is observed for PSD on 5-second long epochs. The lowest average F1-score of 70.335 ± 22.493% are observed for the channel-wise extracted HC on 10-second long epochs.

For high/low valence, the average accuracies and F1-scores are, in general, lower than for arousal. The highest average accuracy of 61.779% ± 10.483% applies to KFD, whereas the poorest average accuracy of 59.649% ± 10.483% applies to HC. Both features are extracted channel-wise and are classified with SVM, although the epoch size of KFD is five seconds. while HC's epoch size is ten seconds. DE has the highest average F1-score for the valence of 71.361 ± 20.360. However, the feature is calculated on five sub-bands, not channel-wise directly from the epochs. The lowest average F1-score of 61.887% ± 31.425% is observed for PSD on five sub-bands using SVM and epoch size of ten seconds.

Table 5.2: Details, average accuracies and average F1-scores of individual features for the overall best-performing models for each feature for high/low valence

Valence (negative-positive)					
Feature	**Method**	**Classifier**	**Epoch size**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Five sub-bands	SVM	5sec	61.596 ± 11.207	71.361 ± 20.360
HM	Channel-wise	SVM	5sec	61.216 ± 6.403	67.838 ± 15.717
HC	Channel-wise	SVM	10sec	59.649 ± 10.483	67.076 ± 17.658
KFD	Channel-wise	SVM	10sec	61.779 ± 9.031	67.901 ± 15.358
HFD	Channel-wise	SVM	5sec	60.714 ± 6.070	66.104 ± 17.379
PSD	Five sub-bands	SVM	10sec	60.902 ± 9.763	61.887 ± 31.425

5.1.2 Combinatorial Features

Table 5.3 and Table 5.4 present the best-performing subject-independent models for high/low arousal and high/low valence, respectively. Based on feature extraction combinations and classification performances from the specialization project, some features are extracted channel-wise while some are extracted from sub-bands. A complete overview of how the different classifiers performed for all feature combinations can be found in Appendix F. The combination of all individual features has shown the lowest average accuracy of 64.662% ± 15.819% and the lowest average F1-score (< 70%) for the classification of high and low arousal. With 66.729% ± 11.192% as average accuracy, the feature combination of DE and PSD classified with KNN achieved the highest average accuracy. The best average F1-score of 72.793% ± 17.490% is achieved by the feature combination of HFD and KFD classified with SVM.

Table 5.3: Details, average accuracies and average F1-scores of combinatorial features for the overall best-performing model for high/low arousal

Arousal (calm-excited)				
Method	**Classifier**	**Epoch size**	**Average accuracy [%]**	**Average F1-score [%]**
HM [CW], HC [CW]	SVM	10sec	65.664 ± 14.769	71.771 ± 22.301
DE [SB], PSD [SB]	KNN	5sec	66.729 ± 11.192	71.841 ± 14.346
HFD [CW], KFD [CW]	SVM	5sec	65.977 ± 11.933	72.793 ± 17.490
HM [SB], DE [SB]	KNN	5sec	66.541 ± 11.216	71.666 ± 14.533
HM [CW], HC [CW], HFD [CW], KFD [CW], DE [SB], PSD [SB]	SVM	20sec	64.662 ± 15.819	69.391 ± 26.704

Abbrev: SB = sub-bands, and CW = channel-wise

The feature combination for high/low valence with the lowest average F1-score is the combination of HFD and KFD, both extracted channel-wise and classified with SVM. The average F1-score is 66.634% ± 7.307%. The lowest average accuracy of 59.524% ± 7.241% is obtained by the combination of all individual features combined with epoch sizes of five seconds. The highest average accuracy of 62.281% ± 7.787% belongs to the feature combination of HFD and KFD. And the lowest average F1-score among the best-performing models can be observed for the combination of HM and DE with a score of 66.634% ± 7.307%.

Table 5.4: Details, average accuracies and average F1-scores of combinatorial features for the overall best-performing model for high/low valence

Valence (negative-positive)				
Method	**Classifier**	**Epoch size**	**Average accuracy [%]**	**Average F1-score [%]**
HM [CW], HC [CW]	SVM	10sec	60.401 ± 10.232	71.771 ± 22.301
DE [SB], PSD [SB]	KNN	5sec	60.652 ± 4.819	66.634 ± 7.307
HFD [CW], KFD [CW]	SVM	10sec	62.281 ± 7.787	67.441 ± 17.356
HM [SB], DE [SB]	KNN	5sec	60.902 ± 4.961	66.875 ± 7.295
HM [CW], HC [CW], HFD [CW], KFD [CW], DE [SB], PSD [SB]	SVM	5sec	59.524 ± 7.241	67.313 ± 30.932

Abbrev: SB = sub-bands, and CW = channel-wise

5.1.3 K-Fold Cross-Validation

Some of the best-performing models in terms of average accuracy and average F1-score for valence and arousal are verified with 5-fold CV and 10-fold CV. The results in Table 5.5 present the average accuracies of all participants average over 5-folds and 10-folds. In Table 5.6, the average F1-scores are presented.

Table 5.5: 5-fold CV and 10-fold CV on the best-performing models with best-reported feature combinations and epoch sizes. Reports accuracies average over 5-folds and 10-folds

K	Classifier	Features	Epoch size [s]	Arousal (accuracy [%])	Valence (accuracy [%])
5	SVM	DE [SB]	5	66.082 ± 11.597	61.028 ± 4.310
	SVM	HFD [CW]	5	66.400 ± 12.841	60.816 ± 6.220
	SVM	KFD [CW]	5	66.078 ± 12.918	60.007 ± 6.160
	SVM	HFD & KFD [CW]	10	65.131 ± 13.571	59.741 ± 6.255
	KNN	DE & PSD [SB]	5	65.033 ± 13.478	59.490 ± 6.099
10	SVM	DE [SB]	5	65.946 ± 11.820	61.007 ± 4.111
	SVM	HFD [CW]	5	66.152 ± 12.987	60.918 ± 5.791
	SVM	KFD [CW]	5	66.559 ± 12.543	60.495 ± 6.030
	SVM	HFD & KFD [CW]	10	66.281 ± 12.733	60.968 ± 5.306
	KNN	DE & PSD [SB]	5	65.033 ± 13.478	59.490 ± 6.099

Table 5.6: 5-fold CV and 10-fold CV on the best-performing models with best-reported feature combinations and epoch sizes. Reports F1-scores average over 5-folds and 10-folds

K	Classifier	Features	Epoch size [s]	Arousal (accuracy [%])	Valence (accuracy [%])
5	SVM	DE [SB]	5	71.787 ± 13.949	66.532 ± 8.584
	SVM	HFD [CW]	5	72.635 ± 18.540	66.317 ± 17.435
	SVM	KFD [CW]	5	72.327 ± 17.805	66.398 ± 14.681
	SVM	HFD & KFD [CW]	10	70.724 ± 20.665	66.126 ± 15.494
	KNN	DE & PSD [SB]	5	67.041 ± 29.325	62.035 ± 28.774
10	SVM	DE [SB]	5	71.927 ± 13.601	66.175 ± 9.329
	SVM	HFD [CW]	5	71.921 ± 19.312	66.237 ± 17.587
	SVM	KFD [CW]	5	72.874 ± 16.794	66.484 ± 14.827
	SVM	HFD & KFD [CW]	10	72.708 ± 17.709	66.528 ± 16.627
	KNN	DE & PSD [SB]	5	66.990 ± 29.151	61.862 ± 28.730

5.2 Subject-Specific Performance

Due to individual differences between subjects, it can be challenging to get an overview of how good the models' general performances are just by looking at the average accuracies and average F1-scores. Therefore, an overview of the performances of the classifiers for all individual subjects is presented here. Only some features and epoch lengths with the best average performances in accuracy and F1-score are included.

Figure 5.1 shows the accuracy of SVM, KNN and MLP with the individual feature DE and the combinatorial feature with DE and PSD, respectively, as input for all subjects. Participants 11, 13 and 21 are the only participants with a performance lower than 60% for all models when DE is extracted channel-wise. Whereas when DE and PSD are extracted from five sub-bands, all the models of participants 20 and 25 have an accuracy below 60% in addition to the already-mentioned ones. The best model performances (all above 75%), regardless of feature extraction, are obtained by participants 12, 16 and 22.

Figure 5.2 shows the F1-score for all subjects when the feature HFD and the feature combination HFD and KFD are extracted. Participant 13's models on the feature combination HFD and KFD show F1-scores lower than 50%, i.e. random guessing. The same applies to Participant 21's SVM and KNN classifiers. The participants with at least two models above 80% for both individual and combinatorial features are participants 12, 16, and 22.

In Figure 5.3 and Figure 5.4, the high/low valence classification performances in terms of accuracy and F1-score, respectively, for all 19 subjects are presented. The best accuracy is obtained by participants 22 and 24 using SVM for both the individual feature and the combinatorial feature. When the individual feature KFD is extracted channel-wise with five seconds long epochs, participants 13, 14 and 20 have one model each with an accuracy lower than random guessing, namely KNN, MLP and KNN, respectively. Participants 14 and 20 also had the poorest accuracy performance (below 50%) for the feature combination HFD and KFD when MLP is used as the classifier.

The three top F1-scores ($\geq 80\%$) for DE, extracted from five sub-bands decomposed into five seconds long epochs, are obtained by participants 15, 22 and 27 using SVM as the classifier. The lowest F1-score is observed for participant 26 with SVM as the classifier. For the feature combination of HFD and KFD, the highest F1-scores (above 80%) are observed for participants 15, 17, 22, 23 and 27 using SVM. Participants 14, 20 and 26 have one model with an F1-score poorer than 50%, i.e. random guessing.
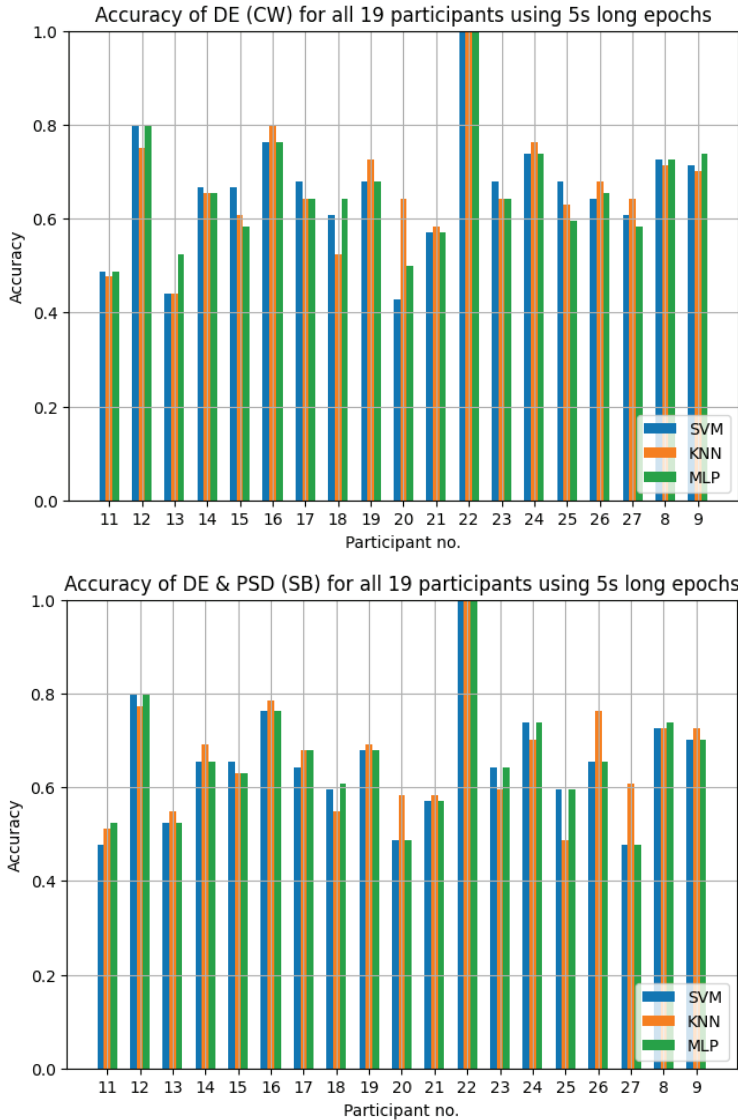
Figure 5.1: Accuracy for all 19 participants using the best-performing individual features (top) and the best-performing combinatorial feature (bottom) classifying high and low arousal
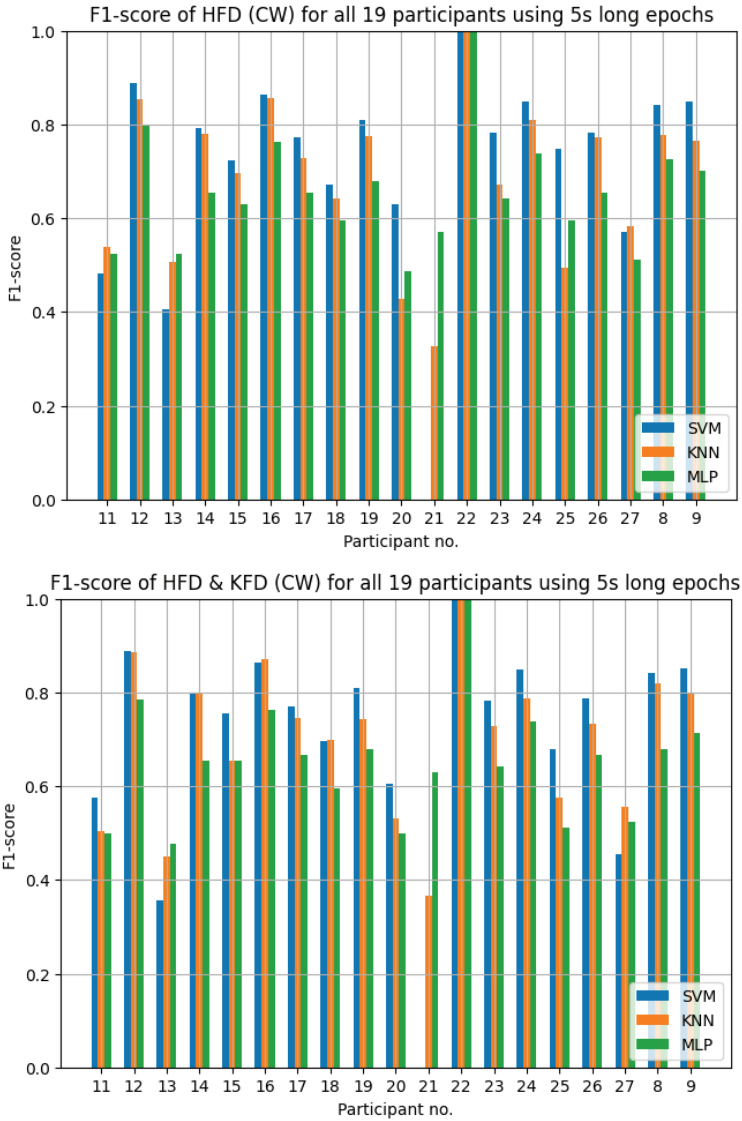
Figure 5.2: F1-score for all 19 participants using the best-performing individual feature (top) and the best-performing combinatorial feature (bottom) classifying high and low arousal
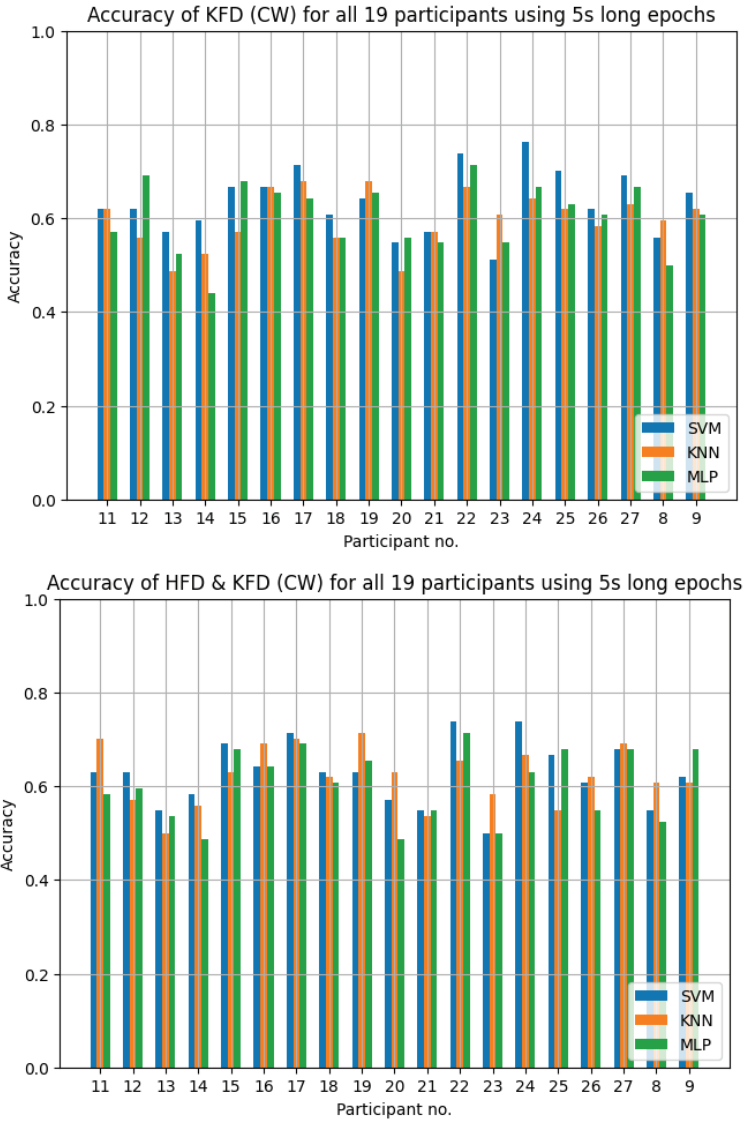
Figure 5.3: Accuracy for all 19 participants using the best-performing individual features (top) and the best-performing combinatorial feature (bottom) classifying high and low valence
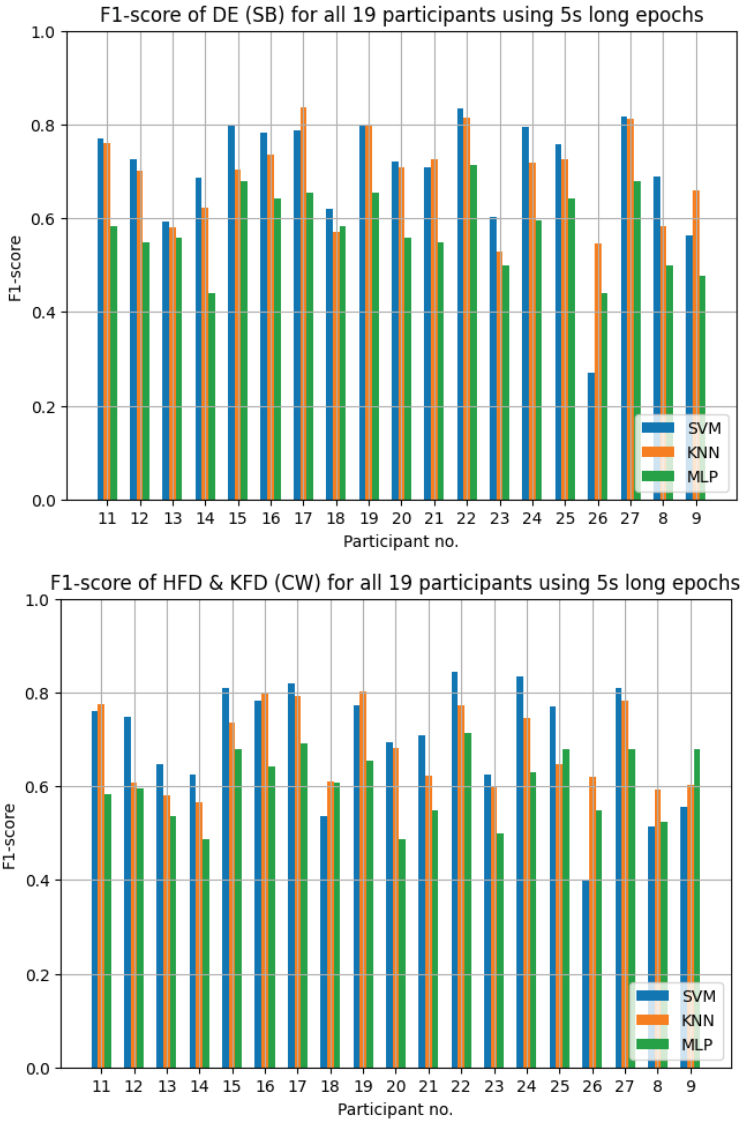
Figure 5.4: F1-score for all 19 participants using the best-performing individual features (top) and the best-performing combinatorial feature (bottom) combinations classifying high and low valence

5.3   Topographic Maps

Topographic maps show the activity in the subjects' brains visually. In Figure 5.5, the weight distribution of different brain regions in five frequency bands is shown for HV/LV and HA/LA. The distribution is activated by the DE feature with sub-bands on participant 12. Figure 5.6, Figure 5.7 and Figure 5.8 show the brain activity of HV/LV and HA/LA without taking frequency bands into account. They show the brain activities of participants 12 and 13. The activation patterns are created by the HFD feature and DE feature with five frequency bands, respectively. No unique pattern is distinguishable in the topographic maps, making it challenging to discriminate between different emotion dimensions. The patterns seem to be more dependent on the participant and frequency bands than whether the rated emotions are high valence/arousal or low valence/arousal.

(a) Delta LA	(b) Theta LA	(c) Alpha LA	(d) Beta LA	(e) Gamma LA
(f) Delta HA	(g) Theta HA	(h) Alpha HA	(i) Beta HA	(j) Gamma HA
(k) Delta LV	(l) Theta LV	(m) Alpha LV	(n) Beta LV	(o) Gamma LV
(p) Delta HV	(q) Theta HV	(r) Alpha HV	(s) Beta HV	(t) Gamma HV
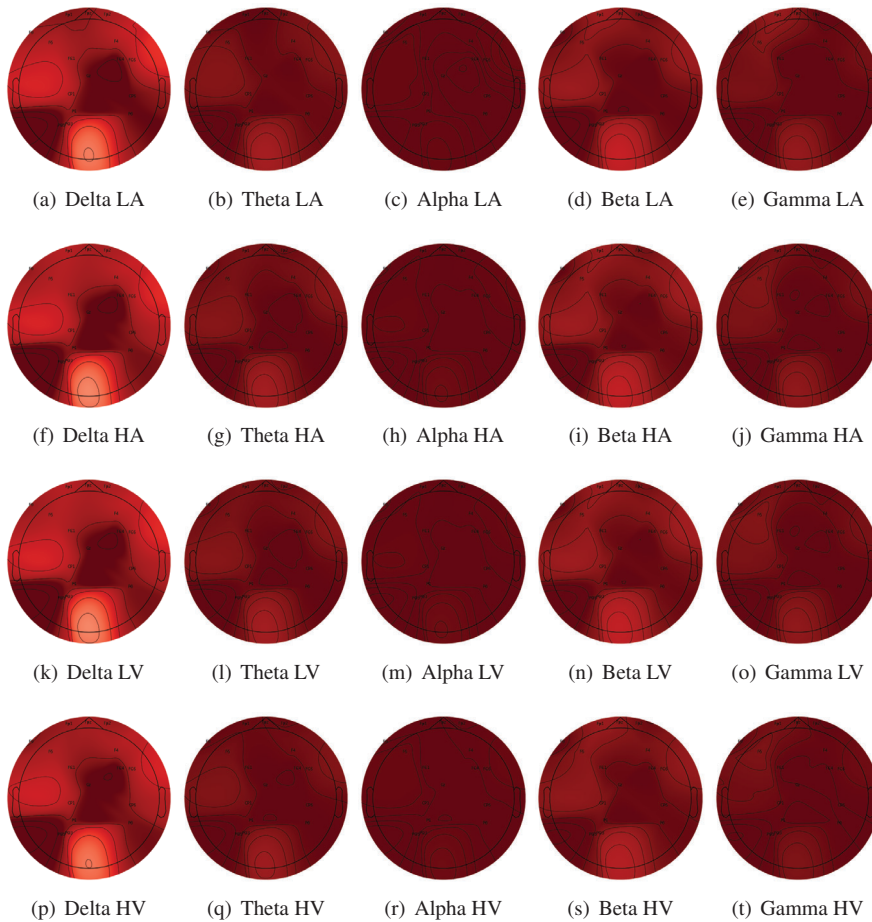
Figure 5.5: Topographic maps of the DE feature with five frequency sub-bands on participant 12

(a) Low valence   (b) High valence   (c) Low arousal   (d) High arousal
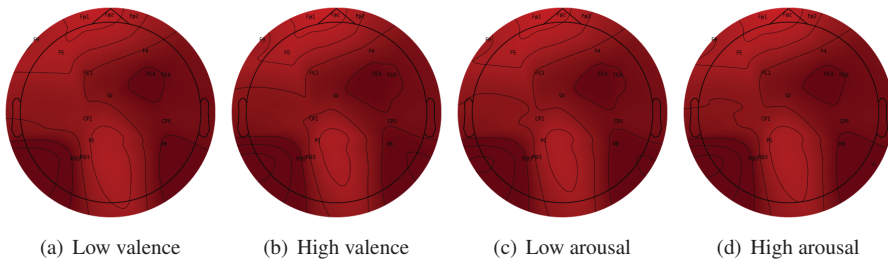
Figure 5.6: Topograpgic maps of the HFD feature on participant 13 for high/low valence and arosual
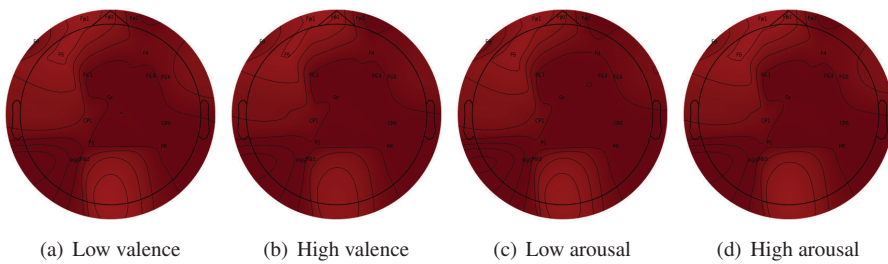
(a) Low valence   (b) High valence   (c) Low arousal   (d) High arousal

Figure 5.7: Topograpgic maps of the HFD feature on participant 12 for high/low valence and arousal

(a) Low valence   (b) High valence   (c) Low arousal   (d) High arousal
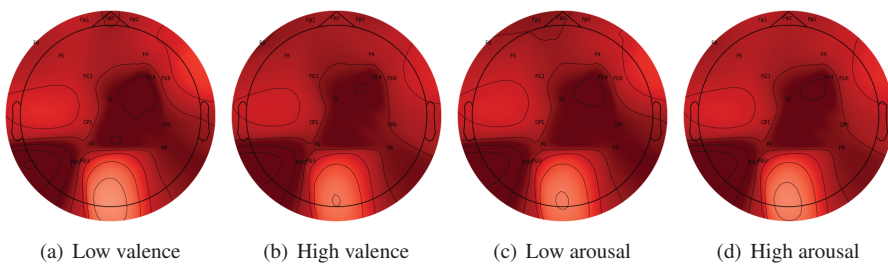
Figure 5.8: Topograpgic maps of the DE feature with five sub-bands on participant 12 for high/low valence and arousal

5.4   Confusion Matrices

The confusion matrices evaluate the models SVM, KNN and MLP for one of the best-performing features, namely DE from five sub-bands, on participants 22 for arousal (Figure 5.9) and 12 for valence (Figure 5.10). Confusion matrices for some of the poorest-performing subjects are also included and are observed in Figure 5.11 for participant 13 and Figure 5.12 for participant 26 for arousal and valence, respectively. Label 0 corresponds to either low valence or low arousal, whereas label 1 corresponds to high valence or high arousal, depending on the binary classification. The confusion matrices are a summation of computed confusion matrices obtained over 5-fold for the respective subjects. It is clear from the confusion matrices of the best-performing subjects and worst-performing subjects that they classify differently.

(a) SVM   (b) KNN   (c) MLP
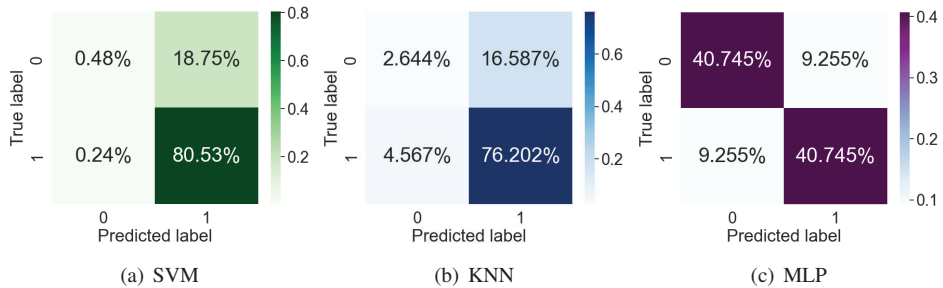
Figure 5.9: Confusion matrices for participant 12 (one of the best) for arousal showing the classification performances when DE from five sub-bands are used as input. Label 0 = LA and label 1 = HA

(a) SVM   (b) KNN   (c) MLP
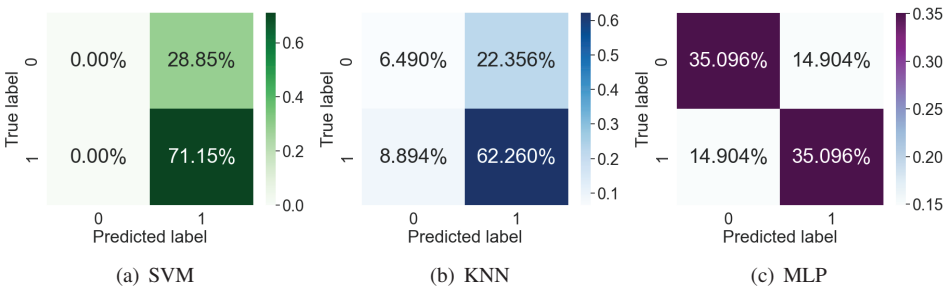
Figure 5.10: Confusion matrices for participant 22 (one of the best) for valence showing the classification performances when DE from five sub-bands are used as input. Label 0 = LV and label 1 = HV

(a) SVM   (b) KNN   (c) MLP
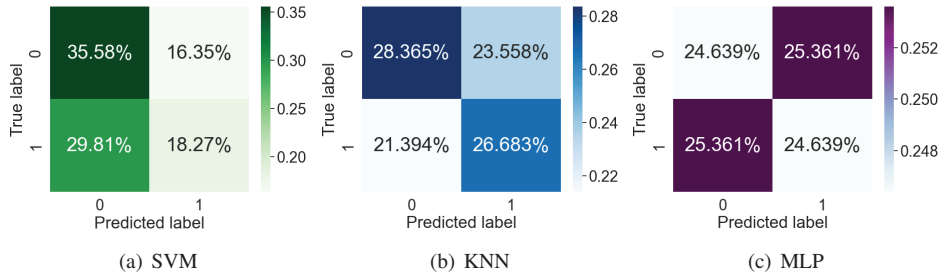
Figure 5.11: Confusion matrices for participant 13 (one of the poorest) for arousal showing the classification performances when DE from five sub-bands are used as input. Label 0 = LA and label 1 = HA

(a) SVM   (b) KNN   (c) MLP
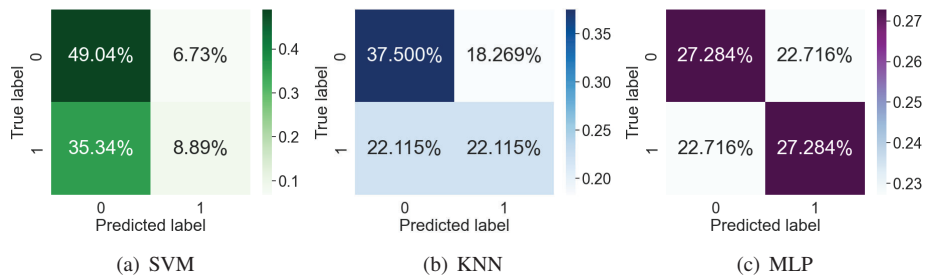
Figure 5.12: Confusion matrices for participant 26 (one of the poorest) for valence showing the classification performances when DE from five sub-bands are used as input. Label 0 = LV and label 1 = HV

6

Discussion

The objective of this chapter is to summarise some of the key findings and make an indepth exploration of the results presented in Chapter 5. The process of data acquisition and preprocessing method will also be discussed here.

6.1   Data Acquistition

Two Unicorn systems with eight dry electrodes on each were used for EEG data acquisition. No impedance measurement was available for the dry system. Thus to ensure a good enough connection, the participants were told to blink their eyes several times and scrunch their faces with the cap on. The recording could start when the signals were significant in the real-time visualization. Some electrode paste was applied on both ear lobes to ensure good contact with the reference electrodes. There was only one cap size available for the participants, so there might be some inconsistency in the electrode position.

LSL was used to time-synchronize the streams from the recording devices, the presenter of movie clips and SAM. From Figure 6.1, it is possible to observe from the first three timestamps of synchronization markers that there are some delays between the two recording systems. The streams are sent through the wireless network, so small delays are expected. During one of the female participant's recording sessions, some unforeseen problems with one of the recording devices arouse, even though the experimenters followed the experimental protocol and setup carefully. As a result, one of the participants has incomplete data and is excluded from this work. The protocol for one experiment lasted about 75 minutes, including small breaks. Some participants have reported that watching movie clips without sound for a long time is demanding and tiring. This might have affected the emotion-elicited EEG data or the SAM rating.

```
[[31533      0      49]  [[31491      0      49]
 [41735      0      50]   [41691      0      50]
 [84001      0      49]]  [83951      0      49]]
```

Figure 6.1: Event markers represent timestamps for the first three events on the Unicorn systems with *Event ID* 49 = 'movie start' and *Event ID* 50 = 'movie end'. From the left: timestamps from device 68 and timestamps from device 69

During the recordings, the participants were informed to limit their motions while watching the movie clips. However, some tiny movement artefacts are expected as it is difficult to sit completely still for a long time. As the recordings happened during office hours some external noises, e.g. conversations in the hallway, construction noise outside, and sounds from doors being opened and closed, are also likely. This could also have some impact on the emotion-elicited data.

6.2 Preprocessing Methods

Before preprocessing the raw EEG data, the emotion-unrelated signals from both devices were removed, and the rest of the signals were joined to a common *mne.io.Raw* object[1]. The signals were then filtered by a notch filter and a bandpass filter. The commonly-used frequency range for the bandpass filter varies among researchers, and there is no standardized cut-off range. According to Zheng and Lu (2015), the bandpass filter should be in the frequency range of $0.3 - 50$ Hz to reduce noise and artefacts, whereas Koelstra et al. (2012) and Miranda-Correa et al. (2021) suggest between $4 - 45$ Hz. Katsigiannis and Ramzan (2018) propose a frequency range of $4 - 30$ Hz to remove most of the ocular artefacts and muscle movements, which are most dominant below 4 Hz and above 30 Hz, respectively. Based on the theory of brain frequency bands, the main sub-bands are located between 0.5 HZ and 30 Hz or more (Sanei and Chambers, 2007). Since this work is based on previous work evaluated on the SEED dataset, the frequency range of the bandpass filter ended up being between $0.3 - 50$ Hz. The ICA method is used to further process the raw EEG signals. The exclusion of ICA components is done manually by looking at each component's properties for each subject. Manual inspection of the ICA components may lead to imprecise artefact removal due to human mistakes.

6.3 Feature Selection

The selection of features is based on the work done in the specialization project. The selection of features plays an important role in EEG-based emotion recognition. The performances of the classifiers can be affected by the choice of features. It is known that

---

[1]https://mne.tools/stable/generated/mne.io.Raw.html

different features respond differently based on whether they are extracted directly from epochs (channel-wise) or with sub-bands. Both individual features and a combination of different features are investigated. The individual features include DE, HM, HC, KFD, HFD, and PSD. They were all extracted channel-wise and from five frequency bands, except for PSD which are only extracted from sub-bands. Combinations of features within the same domain, i.e. time domain, frequency domain, and domains with nonlinear characteristics. Feature combinations across domains are also investigated.

The obtained results indicate that the individual features HFD and KFD, and the combination of them, extracted channel-wise from the EEG signal are observed to be quite suitable for emotion recognition. Both have a good average performance in terms of accuracy and F1 score for valence and arousal. In addition, the performance of the feature DE with and without sub-bands is also quite good. which corresponds well with the findings by Zheng and Lu (2015) and Katsigiannis and Ramzan (2018). It seems like channel-wise feature extraction tends to be superior compared to features from sub-bands. The reason could be that the channel-wise extracted features might be able to capture more relevant information about emotional states. However, the decomposition into frequency bands with a five-order Butterworth filter might cause information loss and the introduction of signal distortion. The filter might not sufficiently preserve all frequency content and the details of the original signal. The sizes of the epochs may also have an impact on the feature extraction. Certain features may need longer epochs to capture meaningful patterns or properties of the EEG signal. By observing the results, epoch sizes of five seconds seem to give the highest performance. Another potential reason could be the lack of baseline correction. A baseline is an EEG signal without stimuli present and typically expresses the underlying brain activity patterns (Yang et al., 2018). With such correction, unwanted temporal drifts and noise in EEG signals can be reduced, and improve classification performance. A combination of features with low performance might affect the accuracy and F1-score poorly as well.

6.4   Classification Performance

The models used for classification are general models using data from all participants and are evaluated in previous work on the SEED dataset. In SEED, the emotions are labelled according to three classes: negative, neutral and positive, whereas in this work, the classification is binary: HV/LV or HA/LA. From the results, SVM seems to be superior compared to KNN and MLP. And the models are more likely to classify HA/LA correctly than HV/LV. However, all methods perform on average in terms of accuracy and F1-score better than random guessing, i.e. accuracy and F1-score above 50%.

The standard deviations on metrics for all models are notably high on both individual and combinatorial features for valence and arousal. They vary from $\pm 6.070\%$ to $\pm 15.467\%$ for average accuracy and from $\pm 11.027\%$ to $\pm 26.704\%$ for average F1-score. This could indicate inconsistency and lack of reliability in the results. High variability in the results can be caused by numerous factors, including noise and unbalanced classes. Dry elec-

trodes usually provide lower signal quality compared to wet electrodes due to high contact impedance, thus being more noisy. Unbalanced classes occur when there is an uneven distribution of classes within a dataset. This seems to be the case since the average F1-scores are better than the average accuracies for all cases. For the 9-point SAM scale, the threshold was placed in the middle, at 4.8 inspired by Betella and Verschure (2016). This could likewise have led to unbalanced classes, considering that individuals have different responses to the same stimuli and hence give different ratings on the SAM scale (Suhaimi et al., 2020). Figure 6.2 shows the number of instances in each class. To cope with the unbalanced dataset problem, F1-score is reported along with accuracy. In addition, 5-fold and 10-fold CVs are used to verify the best-performing features and classifiers. The 5-fold CV and 10-fold CV obtained slightly better average accuracies and average F1-scores compared to a single test-train split, thus suggesting that the models are more reliable and have better generalization potentials, despite high standard deviations.
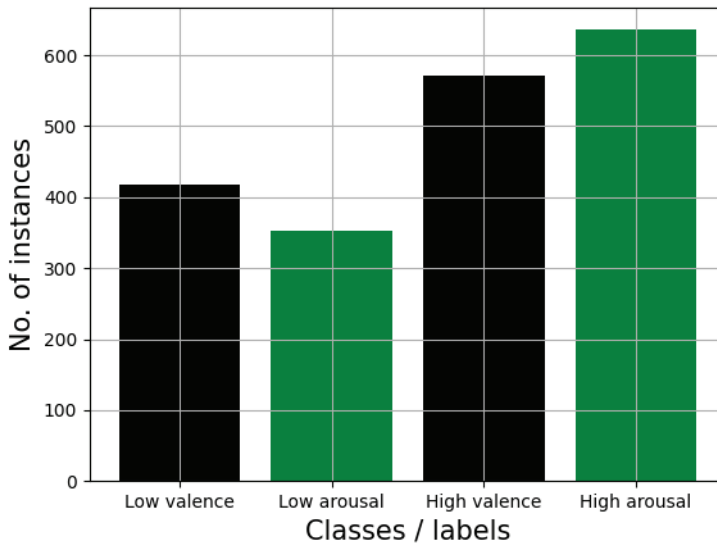
Figure 6.2: Number of instances for each class rated by all 19 participants

6.4.1 Significant Difference in Performance

T-test can be used to determine if there is a significant difference between the performance of two classifier models. It is a convenient way to conclude whether the difference in performance between two classifier models is just a coincidence, i.e. $p > 0.05$. The t-test was performed on the accuracies and F1-scores obtained with SVM, KNN and MLP for the best-reported features, which have also been verified with 5-fold CV and 10-fold CV. For arousal, none of the compared models is significantly different in performance. However,

depending on the feature extraction, there are significant differences in performance ($p <$ 0.05) between some of the compared classifiers for valence. Significant differences in F1-scores can be found between SVM and MLP for the feature combination of DE, HFD, KFD, DE & PSD, and the feature combination of HFD & KFD, in addition to between MLP and KNN for the feature combination of DE, HFD, DE & PSD, and the feature combination of HFD & KFD.

6.4.2 Analysis of Confusion Matrices

Participants 12 and 22 have one of the best performances overall for arousal (Figure 5.9) and valence (Figure 5.10), respectively. The confusion matrices reveal that the classifiers SVM and KNN are better at identifying HV and HA correctly. However, the percentages of LV/LA classified as HV/HA suggest a significant amount of misclassification existing. SVM and KNN also fail to correctly classify LV and LA. None or almost none of the instances that belong to HV were misclassified as low valence. The rates for arousal have the same tendencies and are fairly low. However, none or almost none of the instances that belong to LV/LA were correctly classified as such either. It seems like SVM and KNN are not predicting any instances of low valence or low arousal, and only classify most of them into HV and HA classes. For MLP on LA/HA and LV/HV, the predictions are balanced for HV/LV and HA/LA. The background of this behaviour could be that MLPs struggle to learn the subtle differences and patterns in data, resulting in similar predictions for both classes. As mentioned earlier, the newly-created dataset might be unbalanced and the models fail to address this.

Participants 13 and 26 are the participants that had one of the poorest classification performances for valence and arousal, respectively. The confusion matrices for arousal (Figure 5.11) and valence (Figure 5.12) differ somewhat from the ones from the better-performing subjects. The SVM model is better at identifying LV/LA than HV/HA, although the model tends to incorrectly predict the HV/HA class as the LV/LA class quite often. For KNN and MLP, the predictions are more balanced between HV/LV and HA/LA. However, the FP and FN values are close to the TP and TN values. Ideally, it should be high values for TP and TN, and low values for FP and FN. Based on the confusion matrices, the models are not ideal and might suffer from overfitting instead of generalizing well on new, unseen data.

6.4.3 Analysis of Topographic Maps

The topographic maps of participants 12 and 13 during the influence of different dimensions of emotions, show rather small differences in-between emotions. There is no unique pattern that could be used to discriminate between different emotion dimensions. Although, it is possible to observe that high/low valence and high/low arousal activate slightly different parts of the brain. Figure 5.5 indicates that the patterns of brain activations are slightly more distinct between the five frequency bands than between emotions.

This addresses the difficulties in classifying emotions correctly and could explain some of the challenges with predicting properly in the confusion matrices (Zhang et al., 2021). There might be some explanation for this issue, including unbalanced instances of each class, or the subject might have struggles to express the elicited properly. However, it is important to have in mind that the brain patterns will most likely not be equivalent for all cases for all subjects within the same emotional dimension. The patterns from two individuals with feature extraction of HFD shown in Figure 5.6 and Figure 5.7 is an example of that. The fact that one emotion dimension does not truly correspond to one region of the brain is familiar in the literature. Emotions rather evoke complex arrays of networks in the brain, than a single region. The decoding analysis conducted by Horikawa et al. (2020) showed that emotions are represented through more complex configurations involving multiple networks instead of simple one-to-one mappings between specific emotions and brain regions. This is consistent with findings from Hamann (2012).

6.5 Subject Variability

The collected EEG data are from multiple participants, and they are all used to train and test different models to find the best-performing subject-independent model. However, based on the results, there are huge variabilities in the model performances depending on the classifier and the chosen feature extraction. The quality of EEG signals varies among individuals and also within individuals. According to Zheng and Lu (2015), factors such as educational and social background can affect the experience with emotions. Some people might even have difficulties reaching the outer edge of the emotion dimension using only movie clips with no sounds, i.e. cannot elicit the expected emotional responses to each movie clip. This creates outliers that behave differently than the majority of the participants. The method of labelling emotion-elicited EEG data rely on the participants' subjective self-assessment. Different participants may interpret and rate emotions differently and create inconsistencies in the labelled data. Some may also report their emotional experiences wrong.

Emotions are subjective, thus it is difficult to make a generalized model to decode them. Based on the accuracy and F1-score for all 19 subjects for valence and arousal in Figure 5.1, Figure 5.2, Figure 5.3 and Figure 5.4, the variance in the metrics are distinct. Some participants even have models performing worse than random guessing, i.e. less than 50% for binary classification, which can have a bad impact on the average performance. The general models assume the same distribution for training and testing subsets. However, with outliers present in the dataset, they might introduce noise and distort performance. One participant worth taking a closer look at is participant 22, especially for arousal classification. The accuracy and F1-score for all models are 100%, which deviates from the majority. With a closer examination, it turns out that the labels for this particular participant are unbalanced with only ratings for high arousal. Hence, participant 22 is never exposed to other classes than high arousal during training. This could explain why the HA/LA classification performed better than the HV/LV classification, as the metrics are based on the average of all 19 subjects' performances.

6.6   Comparison with Other Works

The performance of the proposed SVM, KNN and MLP models evaluated on the newly-created dataset are also compared with the results obtained in the specialization project, other publicly available datasets, and state-of-the-art methods. It may be difficult to compare the obtained results with previous works since the methodologies and experimental protocols vary a lot (Avetisyan et al., 2016). The criteria for the subject selection and the used EEG devices might differ. In addition, some researchers even use the same terminology to assess different evaluation metrics, e.g. accuracy, average accuracy, F1-score, the accuracy of individuals, etc.

6.6.1   Comparison with Specialization Project

In the specialization project, the conclusion was that the HM feature gave the best accuracies overall and is a good choice relating to emotion recognition. Furthermore, the features of HFD and KFD were concluded to not be the most suitable features, especially on the SEED dataset using the proposed models. However, in this work, the features of HFD and KFD are quite suitable for emotion recognition. For the classifiers, MLP did not perform as well as it did for the SEED dataset. This might be because MLP is not generalized enough and suffer from overfitting. The models may also have acquired knowledge unique to the SEED dataset, which may not transfer well to the newly-created dataset.

There are several differences between the work done in the specialization project and this work, that could explain the classification performance gap. First of all, the experimental protocols are different. SEED uses 15 movie clips of around four minutes to evoke positive, negative and neutral emotions, while in this work, 52 movie clips of 40 seconds have been used to evoke emotions in the continuous valence-arousal dimension. Even though both protocols use movie clips as their main source of stimuli, SEED uses movie clips with sound. In this work, the clips are non-auditory. The EEG device and the number of channels used also differ with the 62-channel ESI NeuroScan system vs. two 8-channel Unicorn systems. The change in the epoch size could explain the performance. Four different sizes of epochs, namely 5, 10, 20 and 40 seconds were investigated, while in the previous work on SEED, only one-second epochs were used. In addition, there are differences in sampling frequency and duration of the whole experiment.

6.6.2   Comparison with Other Datasets

Table 6.1 shows the results reported by related emotion datasets using EEG for feature extraction. From this work, two of the best models, SVM with HFD and SVM with DE, are compared with the best-performing methods from the datasets: AMIGOS, DEAP, DREAMER, IDEA, and SEED. There are many individual differences between the mentioned datasets, for instance, if they report with average accuracy or average F1-score. However, evaluating the proposed dataset against the others might give some insights into

its characteristics, strengths, and maybe limitations.

Results for the newly-created dataset are consistent with results reported for other datasets. The performance on average F1-scores for valence and arousal are significantly better for the proposed work than the others. For arousal, especially, the difference in the F1-scores is in general more than 13 percentage points better. Using average accuracy as the performance metric, SEED and IDEA are shown to be superior in classifying valence. The reason could be that they don't classify high and low valence continuously like the other databases but classify discrete emotions, i.e. negative and positive emotions. In addition, the classification models BiLSTM and DBN are more complex and efficient than SVM and Gaussian Naïve Bayes (GNB).

Table 6.1: Performance comparison with other datasets

Dataset	Method	Average accuracy [%]		Average F1-score [%]	
		Valence	Arousal	Valence	Arousal
DEAP (2012)	GNB on PSD	57.60	62.00	56.30	58.30
SEED (2015)	DBN on DE	86.08	-	-	-
DREAMER (2018)	SVM on PSD	62.49	62.17	51.84	57.67
IDEA (2022)	BiLSTM on MD-DE	88.57	-	-	-
AMIGOS (2021)	GNB on PSD	-	-	56.40	57.70
Proposed work	SVM on HFD	60.82	66.40	66.32	72.64
	SVM with DE	61.03	66.08	66.53	71.79

The choice of stimuli varies between movie clips with or without sounds, music videos, and for IDEA also mathematical problems and songs. The time of how long the subjects are exposed to stimuli differs as well. In the proposed work, each movie clip lasts for about 40 seconds before self-assessment, while for SEED, the movie clips last for about four minutes each. DEAP uses 40-second-long music videos, and AMIGOS uses a mix of short and long videos. Depending on the subject's background they might react differently to the stimuli as some may need more or less time to elicit the expected emotions.

Preprocessing raw EEG signal before feature extraction and classification is crucial to avoid noise-contaminated signals and remove unwanted artefacts and undesired data from a continuous stream. Most of the datasets filter raw EEG signals to get the frequencies of interest often somewhere between $0 - 50$ Hz. Furthermore, there are different practices on how to decompose into different frequency bands, most of them use the Butterworth filter but the order varies between three and six. EEG data from DEAP, IDEA and AMIGOS are decomposed into four sub-bands, $\theta, \alpha, \beta, \gamma$. Whereas SEED and the newly-created database are decomposed into five in total. The same four sub-bands as DEAP, including δ, are used. DREAMER only provide three frequency bands: θ, α and β. Nevertheless, it is difficult to conclude which dataset is the so-called best performing. Many publicly available datasets have already been preprocessed to obtain favourable results, and the exact details of how the data is touched are often unknown.

6.6.3 Comparison with Other Models

Nowadays, most models used for EEG-based emotion recognition evaluated on well-known datasets have exceptional performance. DE and PSD with sub-bands are common features extracted from the signal, and lately also the Hjorth parameters and fractal dimensions. Most of the state-of-the-art models consist of complex architectures, often including deep learning with several layers and with four or three dimensions, e.g. 4D-aNN (Xiao et al., 2022), 4D-CRNN (Shen et al., 2020), and 3D-CNN (Yang et al., 2018). These types of models are often capable to capture more complex patterns and subtle differences within EEG data, and thus achieve higher performance. Simpler models like MLP and SVM have also been examined and evaluated on SEED and DEAP. Kumar and Molinas (2022) achieved accuracies closer to the state-of-the-art methods, same with Asadur Rahman et al. (2020). The models used in this work are based on previous work done by the author. The performance of the models was evaluated on SEED and got reasonable performances. KNN and MLP were similar or better than some of the state-of-the-art studies, while the SVM performed just a few percentage points below. The reason why these models perform poorly now compared to previous work could include noisy data since dry electrodes are used, unbalanced classes, number and position of electrodes, and choice of stimuli. The split of data samples varies between single split train and test (Liu et al., 2020; Zheng et al., 2019), K-fold CV (Tao et al., 2023; Xiao et al., 2022; Asadur Rahman et al., 2020) and LOVO (Yao et al., 2021), which might also have an impact on the performance.

7

Conclusion

A dataset for emotions elicited by non-audio movie clips is presented in this thesis with experimental design and data collection process. The dataset includes EEG recordings from 20 participants watching 52 movie clips. Two Unicorn devices with 16 electrodes in total are used with a sampling rate set to 250 Hz. Each participant rated the elicited emotions from the movie clips according to the SAM scales for valence, arousal and dominance at the end of each trial.

The collected EEG signals were further preprocessed with a notch filter at 50 Hz and 100 Hz to suppress the power line noise. A bandpass filter between $0.3 - 50$ Hz was also applied to remove artefacts and other noises. ICA was also used to remove ocular artefacts individually from all participants before splitting data into epochs. Four different epoch sizes were tested: 5, 10, 20, and 40 seconds. Some data were further decomposed into five frequency bands of $\delta, \theta, \alpha, \beta$ and γ with a five-order Butterworth bandpass filter. From the epochs with and without sub-bands features from different domains were examined. The feature extraction consisted of HM, HC, KFD, HFD, PSD, DE, and different combinations of them.

In the specialization project, the HM feature with MLP was recognized as the best suitable feature for emotion recognition after being evaluated on SEED and compared with state-of-art methods. HFD and KFD features classified positive, negative, and neutral emotions poorly. The feature extraction methods and models used in the specialization project were evaluated on the newly-created dataset to test for generalization. The evaluation concludes that HFD and KFD are suitable for emotion recognition, and not as poor as predicted in the previous work. Moreover, MLP might be overfitted to SEED and struggle to differentiate the subtle differences in emotions. SVM and KNN fail to predict low valence/arousal for the best-performing models correctly, and the error rates of high valence/arousal predictions are too high. For the poor-performing participants, their confusion matrices seem balanced but with a low prediction performance. The prob-

lem arouses most likely due to unbalanced classes, as there are more instances of high valence/arousal than low valence/arousal available for training.

There are also no unique patterns that are distinct in the topographic maps as emotions are most likely not only evoking single regions of the brain. The models are also general and include data from all participants, including outliers that may affect the general performance. However, the average performance in terms of accuracy and F1-score compared to other publicly available datasets seems consistent despite individual differences in emotion labels, choice of stimuli and classifier, feature extraction, and the number of channels used for EEG recording. There is no conclusion made for which model is best performing. The standard deviation of the average accuracies and the average F1-scores for all models is high, and there is a lot of subject variability between models and in-between valence and arousal discrimination. Emotion classification is challenging, and getting high accuracy is generally more difficult on data from own experiment than publicly available datasets.

7.1 Future Work

In the future, the dataset will hopefully allow researchers to gain a deeper insight into EEG-based emotion recognition by testing different feature extraction and models on the dataset. The dataset is imperfect, so more research scenarios need to be tested and evaluated. If time, more data from new participants should be collected to overcome the problem of unbalanced classes and the limitation of available training data for certain classes. Data collection is time-consuming, so there might be an idea to look into oversampling techniques to create a more balanced dataset by generating samples from the minority class. Transfer learning might also be the key to solving the situation, and should be further explored.

Some other topics that should be further explored are channel reduction, baseline correction and more preprocessing of the raw EEG signals. Channel reduction can make EEG experiments more practical since there will be fewer electrodes to maintain, and less computing power is needed to preprocess and analyse the EEG data. Moctezuma et al. (2022) and Zheng and Lu (2015) suggest that as few as 6-12 channels still get desired data and achieve high performance. Irrelevant channels might introduce artefacts and noise in the system, thus lowering performance (Zheng et al., 2019). Less preprocessing will also be needed if there are fewer channels. Baseline correction is also a topic worth diving into as research has shown that it can improve recognition accuracy (Yang et al., 2018).

Bibliography

Abadi, M. K., Subramanian, R., Kia, S. M., Avesani, P., Patras, I., and Sebe, N. (2015). DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Transactions on Affective Computing*, 6(3):pages 209–222.

Ablin, P., Cardoso, J.-F., and Gramfort, A. (2018). Faster Independent Component Analysis by Preconditioning With Hessian Approximations. *IEEE Transactions on Signal Processing*, 66(15):pages 4040–4049.

Al-Nafjan, A. N., Hosny, M. I., Al-Ohali, Y., and Al-Wabil, A. (2017). Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review. *Applied Sciences*, 7.

Asadur Rahman, M., Foisal Hossain, M., Hossain, M., and Ahmmed, R. (2020). Employing PCA and t-statistical approach for feature extraction and classification of emotion from multichannel EEG signal. *Egyptian Informatics Journal*, 21(1):pages 23–35.

Avetisyan, H., Bruna, O., and Holub, J. (2016). Overview of existing algorithms for emotion classification. Uncertainties in evaluations of accuracies. *Journal of Physics: Conference Series*, 772(1).

Betella, A. and Verschure, P. F. M. J. (2016). The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PloS one*.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25:pages 49–59.

Carvalho, S., Leite, J., Galdo-Alvarez, S., and Goncalves, O. (2012). The Emotional Movie Database (EMDB): A Self-Report and Psychophysiological Study. *Applied psychophysiology and biofeedback*, 37.

Chatrian, G. E., Lettich, E., and Nelson, P. L. (1985). Ten Percent Electrode System for Topographic Studies of Spontaneous and Evoked EEG Activities. *American Journal of EEG Technology*, 25(2):pages 83–92.

Chen, J. H. Y. and Mehmood, R. M. (2020). A Critical Review on State-of-the-Art EEG-Based Emotion Datasets. In *Proceedings of the 1st International Conference on Advanced Information Science and System*, AISS '19, New York, NY, USA. Association for Computing Machinery.

Duan, R.-N., Zhu, J.-Y., and Lu, B.-L. (2013). Differential Entropy Feature for EEG-based Emotion Classification. *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84.

Ekman, P. (1992). Are there basic emotions? *Psychological review*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`. Online; accessed 13 March 2023.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in cognitive sciences*.

Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):pages 277–283.

Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):pages 306–310.

Horikawa, T., Cowen, A. S., Keltner, D., and Kamitani, Y. (2020). The Neural Representation of Visually Evoked Emotion Is High-Dimensional Categorical, and Distributed across Transmodal Brain Regions. *iScience*.

Jacob, J. E., Chandrasekharan, S., Nair, G. K., Cherian, A., and Iype, T. (2021). Effect of combining features generated through non-linear analysis and wavelet transform of EEG signals for the diagnosis of encephalopathy. *Neuroscience Letters*, 765:pages 136–269.

Jiang, X., Bian, G.-B., and Tian, Z. (2019). Removal of Artifacts from EEG Signals: A Review. *Sensors (Basel, Switzerland)*.

Joshi, V. M. and Ghongade, R. B. (2022). IDEA: Intellect database for emotion analysis using EEG signal. *Journal of King Saud University - Computer and Information Sciences*, 34(7):pages 4433–4447.

Kam, J. W., Griffin, S., Shen, A., Patel, S., Hinrichs, H., Heinze, H.-J., Deouell, L. Y., and Knight, R. T. (2019). Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system with wet electrodes. *NeuroImage*, 184:pages 119–129.

Katsigiannis, S. and Ramzan, N. (2018). DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):pages 98–107.

Katz, M. J. (1988). Fractals and the analysis of waveforms. *Computers in Biology and Medicine*, pages 145–156.

Keelawat, P., Thammasan, N., Numao, M., and Kijsirikul, B. (2021). A Comparative Study of Window Size and Channel Arrangement on EEG-Emotion Recognition Using Deep CNN. *Sensors*, 21(5).

Klem, G. H., Lüders, H. O., Jasper, H. H., and Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*.

Klonowski, W. (2009). Everything you wanted to ask about EEG but were afraid to get the right answer. *Nonlinear biomedical physics*.

Knyazev, G. (2013). Eeg correlates of self-referential processing. *Frontiers in Human Neuroscience*, 7.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):pages 18–31.

Kumar, M. and Molinas, M. (2022). Human emotion recognition from EEG signals: model evaluation in DEAP and SEED datasets. In *21st International Conference of the Italian Association for Artificial Intelligence*, Udine, Italy.

Liu, H., Zhang, Y., Li, Y., and Kong, X. (2021). Review on Emotion Recognition Based on Electroencephalography. *Frontiers in Computational Neuroscience*, 15.

Liu, J., Wu, G., Luo, Y., Qiu, S., Yang, S., Li, W., and Bi, Y. (2020). EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder. *Frontiers in Systems Neuroscience*, 14.

Mennes, M., Wouters, H., Vanrumste, B., Lagae, L., and Stiers, P. (2010). Validation of ICA as a tool to remove eye movement artifacts from EEG/ERP. *Psychophysiology*, 47(6):pages 1142–1150.

Miranda-Correa, J. A., Abadi, M. K., Sebe, N., and Patras, I. (2021). AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing*, 12(2):pages 479–493.

Moctezuma, L. A., Abe, T., and Molinas, M. (2022). Two-dimensional CNN-based distinction of human emotions from EEG channels selected by multi-objective evolutionary algorithm. *Scientific Reports*.

Newman, A. J., Godfrey, D., and Post, R. (2021). Data Science for Psychology and Neuroscience — in Python - for the course NESC 3505 Neural Data Science. `https://neuraldatascience.io/intro.html`. Online; accessed 28 April 2023.

Oostenveld, R. and Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, 112(4):pages 713–719.

Pane, E. S., Wibawa, A. D., and Pumomo, M. H. (2018). Channel Selection of EEG Emotion Recognition using Stepwise Discriminant Analysis. In *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, pages 14–19.

Pathirana, S., Asirvatham, D., and Johar, G. (2018). A Critical Evaluation on Low-Cost Consumer-Grade Electroencephalographic Devices. In *2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)*, pages 160–165.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89:pages 334–350.

Qin, X., Zheng, Y., and Chen, B. (2019). Extract EEG Features by Combining Power Spectral Density and Correntropy Spectral Density. In *2019 Chinese Automation Congress (CAC)*, pages 2455–2459.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personally and Social Psychology*.

Sanei, S. and Chambers, J. (2007). *Introduction to EEG*, chapter 1, pages 1–34. John Wiley & Sons, Ltd.

Schomer, D. L. and Lopes da Silva, F. H. (2017). *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press.

Shen, F., Dai, G., Lin, G., Zhang, J., Kong, W., and Zeng, H. (2020). EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cognitive Neurodynamics*, 14.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In Sattar, A. and Kang, B.-h., editors, *AI 2006: Advances in Artificial Intelligence*, volume Vol. 4304, pages 1015–1021.

Sreeshakthy, M., Preethi, J., and Dhilipan, A. (2016). A Survey On Emotion Classification From Eeg Signal Using Various Techniques and Performance Analysis. *International Journal of Information Technology and Computer Science(IJITCS)*, pages 19–26.

Stancin, I., Cifrek, M., and Jovic, A. (2021). A Review of EEG Signal Features and their Application in Driver Drowsiness . *Sensors (Basel, Switzerland)*.

Suhaimi, N. S., Mountstephens, J., and Teo, J. (2020). EEG-Based Emotion Recognition: A State-of-the-Art Review of Current Trends and Opportunities. *Computational intelligence and neuroscience*.

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., and Chen, X. (2023). EEG-Based Emotion Recognition via Channel-Wise Attention and Self Attention. *IEEE Transactions on Affective Computing*, 14(1):pages 382–393.

Topic, A. and Russo, M. (2021). Emotion recognition based on EEG feature maps through deep learning network. *Engineering Science and Technology, an International Journal*, 24(6):pages 1442–1454.

Xiao, G., Shi, M., Ye, M., Xu, B., Chen, Z., and Ren, Q. (2022). 4D attention-based neural network for EEG emotion recognition. *Cognitive Neurodynamics*.

Yang, G., Jiao, R., Jiang, H., and Zhang, T. (2020). Ground Truth Dataset for EEG-Based Emotion Recognition With Visual Indication. *IEEE Access*, 8:pages 188503–188514.

Yang, Y., Wu, Q., Fu, Y., and Chen, X. (2018). Continuous Convolutional Neural Network with 3D Input for EEG-Based Emotion Recognition. In Cheng, L., Leung, A. C. S., and Ozawa, S., editors, *Neural Information Processing*, pages 433–443. Springer International Publishing.

Yao, L., Wang, M., Lu, Y., Li, H., and Zhang, X. (2021). EEG-Based Emotion Recognition by Exploiting Fused Network Entropy Measures of Complex Networks across Subjects. *Entropy*.

Yuvaraj, R., Thagavel, P., Thomas, J., Fogarty, J., and Ali, F. (2023). Comprehensive Analysis of Feature Extraction Methods for Emotion Recognition from Multichannel EEG Recordings. *Sensors*, 23.

Zhang, X., Du, T., and Zhang, Z. (2021). EEG Emotion Recognition Based on Channel Attention for E-Healthcare Applications. In Lokoč, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., and Patras, I., editors, *MultiMedia Modeling*, pages 159–169. Springer International Publishing.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. (2019). EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Transactions on Cybernetics*, 49(3):pages 1110–1122.

Zheng, W.-L. and Lu, B.-L. (2015). Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):pages 162–175.

Appendix

A GitHub Repository With Source Codes

The code excerpts can be found as a part of the GitHub repository *WavesResearch*. The link to the codes used in Chapter 4 is available at `https://github.com/wavesresearch/emotion_unicorn_Rose/`. A request for permission might be needed.

B Overview of Some Relevant Datasets

Dataset	Stimuli	Modality	No. of subjects	No. of channels	Feature extraction	Classifiers	Emotions
AMIGOS 2021	Movie clips	ECG,EEG GSR	40	14	HR, PSD, SR++	GNB	Arousal, valence, control, familiarity, liking, basic emotions
DEAP 2012	Music videos	BVP, EEG, EMG, EOG, GSR	32	32	Derivation, HR, mean, PSD, SR, STD ++	GNB	Arousal, valence, liking, dominance, familiarity
DECAF 2015	Movie clips	ECG, tEMG, hEOG, MEG, NIR++	23	16	DCT, PSD, motions	SVM	Arousal, valence, dominance
DREAMER 2018	Movie clips	ECG, EEG	23	16	PSD	SVM	Arousal, valence, dominance
EMDB 2012	Movie clips	HR, SCL	32	-	-	-	Arousal, valence, dominance
IDEA 2022	Movie clips, math problems, songs	ECG, EEG, EMG	14	24 (32 in total)	HA, HM HC, PSD, MD-DE, RASM, DASM	BiLSTM, MLP	Positive, negative
SEED 2015	Movie clips	EEG	15	62	DE	DBN, KNN, SVM	Positive, negative, neutral

C Information Letter And Consent Form

Are you interested in taking part in the research project

"David and Goliath: single-channel EEG unravels its power through adaptive signal analysis: EEG-based emotion recognition study"?

_____ _____

We would like to invite you to take part in a research project where the main purpose is to develop models that can decode human emotions through the analysis of EEG signals by learning from experiments designed for emotion elicitation. In this letter we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project

We are interested in understanding how human emotions function and promote effective communication among individuals and human-to-machine information exchange. Different emotions can activate the same brain location, or conversely, a single emotion can activate several brain structures. Therefore, there is no simple mapping between affective states and specific brain structures. We will investigate this using electroencephalography (EEG) signals, which is a safe and non-invasive technique to record brain signals. This project aims at creating an emotion-inducing database to help developing models that can decode human emotions through EEG signal analysis. The outcome of the study could be a helpful tool in the diagnosis of depression, PTSD, and other mental disorders.

Who is responsible for the research project?
Norwegian University of Science and Technology (NTNU) is the institution responsible for the project.

Why are you being asked to participate?
You have been invited to take part in this research because you are healthy and over 18 years of age. Please avoid participating if you have neurological diseases or use strong medicine or drugs.

What does participation involve for you?
If you chose to take part in the project, this involves participating in one session with data collection. One session lasts about 75 minutes. 35 of these minutes are for collection of EEG signals from your brain using the Mentalab system (https://mentalab.com/). During the session, you will watch 52 non-auditory movie clips from 6 different categories: horror, erotic, social positive content, social negative content, scenery and object manipulation. After each movie clip, you will be asked to rate your elicited emotions following the SAM (Self-Assessment Manikin) scale description. SAM is a pictorial assessment technique that measures emotional reactions on three dimensions: valence, arousal and dominance. The valence dimension ranges from pleasure to displeasure, the arousal dimension ranges from excited to relaxed, and the dominance dimension ranges from submissiveness to dominance. You will get a thorough explanation before the recordings.

Participation in this study will take approximately 2 hours of your time. You will not be given information about which movie clips you will be watching, as this may affect the results. We do not anticipate you to experience negative feelings when responding to items in this study, however, some of the movie clips might portray potentially shocking scenes. Your participation in this study is completely voluntary. Should you decide to discontinue participation or decline to answer any specific part of the study, you may do so without penalty.

It is important to mention that we are using a device developed for recording brain signals in wet conditions. This means that we will apply electrode cap gel to your hair to increase conductivity between electrode and skin. But the gel is easily washed out with water and shampoo. We will also clean the areas of the scalp where the electrodes are placed with isopropyl alcohol. After the experiment we will ask you some question about how you feel (e.g., are you relaxed, tired or bored) and some feedback questions about the experiment (e.g., duration, procedure, and equipment).

Participation is voluntary
Participation in the project is voluntary. If you chose to participate, you can withdraw your consent at any time without giving a reason. All information about you will then be made anonymous. There will be no negative consequences for you if you chose not to participate or later decide to withdraw. Additionally, there are no risks associated with an EEG test. The test is non-invasive, painless, and safe.

Your personal privacy – how we will store and use your personal data
We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act).
- To protect your privacy and confidentiality, PI Professor Marta Molinas and co-PI Dr Andres Soler and are going to have access to the personal data.
- In addition, we will replace your name and contact details with a code. The list of names, contact details and respective codes will be stored separately from the rest of the collected data, we will store the collected data on a computer protected by the Norwegian University of Science and Technology security systems
- Other group members of the research project will have access just to collected data that has been de-identified
- No personal information will appear in any publication of the research project. The data will be reported in a way that will not identify you.

What will happen to your personal data at the end of the research project?
The project is scheduled to end on December 31, 2024. The personal data will be deleted and destroyed, including any digital recordings at the end of the project. However, we would like to make the recorded electroencephalographic data collected in this study available to other researchers after the study is completed. For this, the researcher will remove any identifying information. Researchers of future studies will not ask your permission for each new study. The other researcher will not have access to your name and other information that could potentially identify you nor will they attempt to identify you.

Your rights
So long as you can be identified in the collected data, you have the right to:
- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

What gives us the right to process your personal data?
We will process your personal data based on your consent.

Based on an agreement with Norwegian University of Science and Technology, NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?
If you have questions about the project, or want to exercise your rights, contact:
- Norwegian University of Science and Technology via Marta Molinas, by mail: (marta.molinas@ntnu.no) or phone: +47 94287670
- Our Data Protection Officer: Thomas Helgesen at Norwegian University of Science and Technology, by mail: (thomas.helgesen@ntnu.no)
- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17.

Yours sincerely,

Marta Molinas /Andres Soler
(Researcher/supervisor)

Consent form

I have received and understood information about the project: "David and Goliath: single-channel EEG unravels its power through adaptive signal analysis" and the study "Reconstruction of brain activity during motor imaginary and motor execution tasks" and have been given the opportunity to ask questions. I give consent:

- ☐ to participate in the questionnaire of the study.
- ☐ to participate in the recording of electroencephalographic data.
- ☐ to my recorded data as well as its processed outcomes can be published and shared for scientific purposes in anonymized form.

I give consent for my personal data to be processed until the end date of the project, approx. December 31, 2024

Signature: _____ Date: _____

D   Detailed Average Accuracies And Average F1-Score of Arousal Classification For Individual Features

Arousal Classification Performance - 5 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for arousal using train-test subsets and five seconds long epochs

Feature	Method	Classifier	Average accuracy [%]	Average F1-score [%]
Arousal (calm-excited) - 5 seconds epochs - individual features				
DE	Channel-wise	SVM	67.607 ± 11.207	72.410 ± 19.229
	Channel-wise	KNN	64.787 ± 11.801	70.801 ± 10.239
	Channel-wise	MLP	65.476 ± 11.651	65.476 ± 11.651
	Five sub-bands	SVM	66.979 ± 10.580	71.361 ± 20.360
	Five sub-bands	KNN	66.729 ± 11.192	71.841 ± 14.347
	Five sub-bands	MLP	64.975 ± 12.380	64.975 ± 12.380
HM	Channel-wise	SVM	66.792 ± 10.961	72.817 ± 17.836
	Channel-wise	KNN	65.351 ± 11.046	71.313 ± 12.250
	Channel-wise	MLP	64.787 ± 12.645	64.787 ± 12.645
	Five sub-bands	SVM	64.724 ± 12.625	67.101 ± 30.845
	Five sub-bands	KNN	63.409 ± 11.237	68.918 ± 15.382
	Five sub-bands	MLP	64.850 ± 12.494	64.850 ± 12.494
HC	Channel-wise	SVM	65.727 ± 11.709	72.778 ± 15.685
	Channel-wise	KNN	60.276 ± 13.228	66.522 ± 15.219
	Channel-wise	MLP	64.474 ± 12.897	64.474 ± 12.897
	Five sub-bands	SVM	64.724 ± 12.817	70.465 ± 23.489
	Five sub-bands	KNN	62.155 ± 13.302	68.524 ± 16.414
	Five sub-bands	MLP	63.910 ± 13.580	63.910 ± 13.580
KFD	Channel-wise	SVM	66.040 ± 11.255	72.531 ± 17.180
	Channel-wise	KNN	62.155 ± 12.007	68.513 ± 13.838
	Channel-wise	MLP	64.223 ± 12.072	64.223 ± 12.072
	Five sub-bands	SVM	64.787 ± 13.073	70.008 ± 24.075
	Five sub-bands	KNN	63.847 ± 11.605	70.038 ± 13.970
	Five sub-bands	MLP	62.531 ± 13.459	62.531 ± 13.459
HFD	Channel-wise	SVM	67.607 ± 11.284	73.167 ± 18.716
	Channel-wise	KNN	64.850 ± 11.428	71.144 ± 13.513
	Channel-wise	MLP	64.724 ± 12.411	64.724 ± 12.411
	Five sub-bands	SVM	64.724 ± 12.453	67.101 ± 30.425
	Five sub-bands	KNN	62.469 ± 11.574	68.390 ± 15.273
	Five sub-bands	MLP	65.132 ± 11.988	65.132 ± 11.988
PSD	Five sub-bands	SVM	65.789 ± 11.877	66.650 ± 30.923
	Five sub-bands	KNN	65.727 ± 11.655	71.159 ± 15.914
	Five sub-bands	MLP	64.724 ± 12.625	64.724 ± 12.625

Arousal Classification Performance - 10 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for arousal using train-test subsets and ten seconds long epochs

Arousal (calm-excited) - 10 seconds epochs - individual features				
Feature	**Method**	**Classifier**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	66.667 ± 14.197	70.589 ± 24.018
	Channel-wise	KNN	61.654 ± 16.532	67.758 ± 19.085
	Channel-wise	MLP	63.283 ± 16.710	63.283 ± 16.710
	Five sub-bands	SVM	64.662 ± 15.436	67.936 ± 25.901
	Five sub-bands	KNN	61.028 ± 15.209	66.674 ± 19.519
	Five sub-bands	MLP	62.657 ± 17.243	62.657 ± 17.243
HM	Channel-wise	SVM	65.915 ± 14.614	71.928 ± 20.864
	Channel-wise	KNN	64.286 ± 12.549	71.634 ± 13.674
	Channel-wise	MLP	64.662 ± 13.889	64.662 ± 13.889
	Five sub-bands	SVM	64.16 ± 15.922	63.867 ± 34.842
	Five sub-bands	KNN	63.283 ± 14.446	68.063 ± 19.471
	Five sub-bands	MLP	64.035 ± 16.029	64.035 ± 16.029
HC	Channel-wise	SVM	64.286 ± 15.873	70.355 ± 22.493
	Channel-wise	KNN	62.030 ± 14.141	69.032 ± 16.828
	Channel-wise	MLP	62.030 ± 14.666	62.030 ± 14.666
	Five sub-bands	SVM	64.035 ± 16.166	63.867 ± 34.842
	Five sub-bands	KNN	60.902 ± 15.086	66.050 ± 21.082
	Five sub-bands	MLP	66.667 ± 13.113	66.667 ± 13.113
KFD	Channel-wise	SVM	65.664 ± 14.468	71.445 ± 20.999
	Channel-wise	KNN	61.404 ± 16.430	68.813 ± 17.741
	Channel-wise	MLP	64.536 ± 15.119	64.536 ± 15.119
	Five sub-bands	SVM	64.411 ± 16.06	68.565 ± 25.647
	Five sub-bands	KNN	60.150 ± 15.509	66.086 ± 18.639
	Five sub-bands	MLP	63.784 ± 16.64	63.784 ± 16.640
HFD	Channel-wise	SVM	64.912 ± 16.339	70.525 ± 23.532
	Channel-wise	KNN	62.406 ± 15.112	68.881 ± 17.066
	Channel-wise	MLP	65.038 ± 14.785	65.038 ± 14.785
	Five sub-bands	SVM	64.160 ± 15.705	63.867 ± 34.368
	Five sub-bands	KNN	63.784 ± 13.741	69.751 ± 16.607
	Five sub-bands	MLP	64.912 ± 14.974	64.912 ± 14.974
PSD	Five sub-bands	SVM	64.662 ± 15.147	65.354 ± 31.746
	Five sub-bands	KNN	65.163 ± 14.937	70.715 ± 19.040
	Five sub-bands	MLP	64.160 ± 15.922	64.160 ± 15.922

Arousal Classification Performance - 20 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for arousal using train-test subsets and 20 seconds long epochs

Arousal (calm-excited) - 20 seconds epochs - individual features				
Feature	**Method**	**Classifier**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	62.657 ± 17.189	70.134 ± 18.337
	Channel-wise	KNN	60.902 ± 18.953	66.625 ± 20.607
	Channel-wise	MLP	60.652 ± 17.521	60.652 ± 17.521
	Five sub-bands	SVM	62.907 ± 16.232	68.929 ± 19.244
	Five sub-bands	KNN	56.892 ± 18.813	62.959 ± 22.872
	Five sub-bands	MLP	62.657 ± 16.208	62.657 ± 16.208
HM	Channel-wise	SVM	63.409 ± 16.652	70.182 ± 18.535
	Channel-wise	KNN	61.153 ± 16.439	66.446 ± 18.734
	Channel-wise	MLP	58.647 ± 19.637	58.647 ± 19.637
	Five sub-bands	SVM	64.912 ± 15.730	69.588 ± 26.744
	Five sub-bands	KNN	57.143 ± 17.022	63.570 ± 21.016
	Five sub-bands	MLP	62.155 ± 16.759	62.155 ± 16.759
HC	Channel-wise	SVM	62.657 ± 17.041	68.864 ± 19.515
	Channel-wise	KNN	57.644 ± 17.453	65.864 ± 17.529
	Channel-wise	MLP	58.396 ± 17.084	58.396 ± 17.084
	Five sub-bands	SVM	64.662 ± 15.085	70.911 ± 22.207
	Five sub-bands	KNN	58.396 ± 14.792	61.169 ± 23.119
	Five sub-bands	MLP	65.915 ± 14.912	65.915 ± 14.912
KFD	Channel-wise	SVM	62.155 ± 16.302	69.022 ± 18.168
	Channel-wise	KNN	61.905 ± 17.388	66.664 ± 18.966
	Channel-wise	MLP	57.644 ± 19.173	57.644 ± 19.173
	Five sub-bands	SVM	60.902 ± 17.999	64.224 ± 27.418
	Five sub-bands	KNN	59.148 ± 17.045	65.903 ± 18.872
	Five sub-bands	MLP	57.143 ± 20.015	57.143 ± 20.015
HFD	Channel-wise	SVM	63.158 ± 15.458	69.958 ± 19.768
	Channel-wise	KNN	60.401 ± 17.173	65.600 ± 17.709
	Channel-wise	MLP	61.404 ± 16.334	61.404 ± 16.334
	Five sub-bands	SVM	64.912 ± 15.516	69.588 ± 26.380
	Five sub-bands	KNN	59.900 ± 15.682	65.898 ± 19.507
	Five sub-bands	MLP	65.288 ± 15.417	65.288 ± 15.417
PSD	Five sub-bands	SVM	63.910 ± 16.212	66.472 ± 27.860
	Five sub-bands	KNN	60.150 ± 16.047	64.877 ± 20.399
	Five sub-bands	MLP	65.163 ± 15.475	65.163 ± 15.475

Arousal Classification Performance - 40 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for arousal using train-test subsets and 40 seconds long epochs

Arousal (calm-excited) - 40 seconds epochs - individual features				
Feature	**Method**	**Classifier**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	66.029 ± 16.554	71.393 ± 23.265
	Channel-wise	KNN	64.593 ± 14.500	70.737 ± 16.121
	Channel-wise	MLP	62.201 ± 15.530	62.201 ± 15.530
	Five sub-bands	SVM	66.507 ± 14.549	73.459 ± 15.791
	Five sub-bands	KNN	64.593 ± 16.289	69.817 ± 19.606
	Five sub-bands	MLP	63.636 ± 14.533	63.636 ± 14.533
HM	Channel-wise	SVM	64.593 ± 15.715	69.007 ± 23.862
	Channel-wise	KNN	61.722 ± 15.023	69.925 ± 14.970
	Channel-wise	MLP	57.895 ± 18.235	57.895 ± 18.235
	Five sub-bands	SVM	65.072 ± 15.530	67.620 ± 31.445
	Five sub-bands	KNN	61.244 ± 18.641	69.664 ± 18.166
	Five sub-bands	MLP	66.029 ± 13.498	66.029 ± 13.498
HC	Channel-wise	SVM	66.507 ± 15.467	69.264 ± 28.488
	Channel-wise	KNN	58.373 ± 15.545	64.679 ± 21.574
	Channel-wise	MLP	56.938 ± 17.628	56.938 ± 17.628
	Five sub-bands	SVM	65.072 ± 15.530	67.620 ± 31.445
	Five sub-bands	KNN	60.766 ± 23.089	65.958 ± 27.506
	Five sub-bands	MLP	63.636 ± 17.142	63.636 ± 17.142
KFD	Channel-wise	SVM	61.722 ± 18.822	65.743 ± 27.348
	Channel-wise	KNN	61.722 ± 17.029	67.908 ± 19.804
	Channel-wise	MLP	58.852 ± 16.699	58.852 ± 16.699
	Five sub-bands	SVM	63.636 ± 15.152	66.166 ± 25.815
	Five sub-bands	KNN	59.330 ± 16.140	64.336 ± 24.069
	Five sub-bands	MLP	57.895 ± 15.807	57.895 ± 15.807
HFD	Channel-wise	SVM	64.115 ± 18.048	65.66 ± 29.858
	Channel-wise	KNN	62.679 ± 15.715	69.953 ± 14.797
	Channel-wise	MLP	63.636 ± 14.213	63.636 ± 14.213
	Five sub-bands	SVM	65.072 ± 15.318	67.620 ± 31.018
	Five sub-bands	KNN	62.919 ± 17.220	69.269 ± 20.794
	Five sub-bands	MLP	65.311 ± 16.077	65.311 ± 16.077
PSD	Five sub-bands	SVM	66.029 ± 15.700	68.836 ± 29.188
	Five sub-bands	KNN	64.115 ± 17.000	69.543 ± 23.073
	Five sub-bands	MLP	65.550 ± 15.023	65.550 ± 15.023

E   Detailed Average Accuracies And Average F1-Score of Valence Classification For Individual Features

Valence Classification Performance - 5 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for valence using train-test subsets and five seconds long epochs

Valence (negative-positive) - 5 seconds epochs - individual features				
Feature	Method	Classifier	Average accuracy [%]	Average F1-score [%]
DE	Channel-wise	SVM	61.596 ± 11.207	66.914 ± 15.263
	Channel-wise	KNN	59.837 ± 5.624	65.488 ± 8.747
	Channel-wise	MLP	60.714 ± 6.687	60.714 ± 6.687
	Five sub-bands	SVM	60.088 ± 10.580	71.361 ± 20.360
	Five sub-bands	KNN	60.652 ± 4.819	66.634 ± 7.307
	Five sub-bands	MLP	57.519 ± 9.188	57.519 ± 9.188
HM	Channel-wise	SVM	61.216 ± 6.403	67.838 ± 15.717
	Channel-wise	KNN	57.707 ± 6.585	63.900 ± 9.754
	Channel-wise	MLP	59.461 ± 5.242	59.461 ± 5.242
	Five sub-bands	SVM	59.336 ± 7.584	60.137 ± 32.193
	Five sub-bands	KNN	56.579 ± 5.575	62.717 ± 9.907
	Five sub-bands	MLP	59.962 ± 6.690	59.962 ± 6.690
HC	Channel-wise	SVM	60.276 ± 5.444	66.58 ± 15.823
	Channel-wise	KNN	58.584 ± 6.430	64.174 ± 10.757
	Channel-wise	MLP	60.840 ± 5.352	60.840 ± 5.352
	Five sub-bands	SVM	59.085 ± 7.594	64.902 ± 21.880
	Five sub-bands	KNN	53.947 ± 6.521	60.646 ± 9.815
	Five sub-bands	MLP	59.336 ± 7.584	59.336 ± 7.584
KFD	Channel-wise	SVM	60.777 ± 6.128	67.467 ± 14.49
	Channel-wise	KNN	56.015 ± 6.429	62.04 ± 9.737
	Channel-wise	MLP	59.023 ± 6.922	59.023 ± 6.922
	Five sub-bands	SVM	59.837 ± 7.714	64.589 ± 22.675
	Five sub-bands	KNN	56.516 ± 5.824	62.038 ± 9.772
	Five sub-bands	MLP	58.459 ± 7.327	58.459 ± 7.327
HFD	Channel-wise	SVM	60.714 ± 6.070	66.104 ± 17.379
	Channel-wise	KNN	57.206 ± 5.477	62.644 ± 10.140
	Channel-wise	MLP	60.526 ± 6.463	60.526 ± 6.463
	Five sub-bands	SVM	59.336 ± 7.481	60.137 ± 31.755
	Five sub-bands	KNN	56.861 ± 6.104	63.119 ± 10.450
	Five sub-bands	MLP	58.709 ± 8.220	58.709 ± 8.220
PSD	Five sub-bands	SVM	59.962 ± 6.523	62.957 ± 26.875
	Five sub-bands	KNN	58.584 ± 6.993	64.427 ± 11.027
	Five sub-bands	MLP	59.962 ± 6.690	59.962 ± 6.690

Valence Classification Performance - 10 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for valence using train-test subsets and ten seconds long epochs

Valence (negative-positive) - 10 seconds epochs - individual features				
Feature	**Method**	**Classifier**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	58.772 ± 11.309	62.668 ± 20.927
	Channel-wise	KNN	55.764 ± 12.303	60.643 ± 15.936
	Channel-wise	MLP	59.273 ± 11.333	59.273 ± 11.333
	Five sub-bands	SVM	58.145 ± 10.775	62.728 ± 21.172
	Five sub-bands	KNN	57.268 ± 7.549	63.457 ± 11.972
	Five sub-bands	MLP	58.145 ± 10.775	62.728 ± 21.172
HM	Channel-wise	SVM	60.526 ± 10.478	67.181 ± 17.151
	Channel-wise	KNN	57.018 ± 10.423	62.771 ± 14.316
	Channel-wise	MLP	61.404 ± 8.823	61.404 ± 8.823
	Five sub-bands	SVM	60.276 ± 10.258	60.952 ± 32.878
	Five sub-bands	KNN	56.767 ± 7.827	63.261 ± 12.200
	Five sub-bands	MLP	57.895 ± 12.168	57.895 ± 12.168
HC	Channel-wise	SVM	59.649 ± 10.483	67.076 ± 17.658
	Channel-wise	KNN	53.885 ± 10.232	59.813 ± 14.861
	Channel-wise	MLP	59.900 ± 9.565	59.900 ± 9.565
	Five sub-bands	SVM	58.271 ± 11.456	62.575 ± 28.239
	Five sub-bands	KNN	55.890 ± 8.561	61.568 ± 12.615
	Five sub-bands	MLP	60.150 ± 10.494	60.15 ± 10.494
KFD	Channel-wise	SVM	61.779 ± 9.031	67.901 ± 15.358
	Channel-wise	KNN	56.767 ± 8.220	63.469 ± 12.706
	Channel-wise	MLP	62.030 ± 8.565	62.030 ± 8.565
	Five sub-bands	SVM	58.396 ± 10.659	61.675 ± 25.772
	Five sub-bands	KNN	52.757 ± 10.477	59.147 ± 14.271
	Five sub-bands	MLP	51.629 ± 11.141	51.629 ± 11.141
HFD	Channel-wise	SVM	60.276 ± 10.679	65.991 ± 19.761
	Channel-wise	KNN	55.890 ± 9.503	62.880 ± 14.320
	Channel-wise	MLP	60.025 ± 10.274	60.025 ± 10.274
	Five sub-bands	SVM	60.276 ± 10.118	60.952 ± 32.43
	Five sub-bands	KNN	57.957 ± 7.432	64.202 ± 11.78
	Five sub-bands	MLP	60.276 ± 10.118	60.276 ± 10.118
PSD	Five sub-bands	SVM	60.902 ± 9.763	61.887 ± 31.425
	Five sub-bands	KNN	57.018 ± 7.795	63.827 ± 11.714
	Five sub-bands	MLP	61.028 ± 9.398	61.028 ± 9.398

Valence Classification Performance - 20 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for valence using train-test subsets and 20 seconds long epochs

Valence (negative-positive) - 20 seconds epochs - individual features				
Feature	**Method**	**Classifier**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	54.637 ± 12.023	60.813 ± 20.333
	Channel-wise	KNN	52.632 ± 8.905	57.901 ± 14.402
	Channel-wise	MLP	58.647 ± 9.531	58.647 ± 9.531
	Five sub-bands	SVM	52.632 ± 12.546	59.729 ± 20.137
	Five sub-bands	KNN	52.632 ± 12.646	55.508 ± 19.203
	Five sub-bands	MLP	58.145 ± 11.727	58.145 ± 11.727
HM	Channel-wise	SVM	55.138 ± 11.033	63.189 ± 16.497
	Channel-wise	KNN	53.634 ± 10.390	60.368 ± 13.988
	Channel-wise	MLP	52.381 ± 11.224	52.381 ± 11.224
	Five sub-bands	SVM	58.396 ± 9.503	61.895 ± 28.413
	Five sub-bands	KNN	51.378 ± 8.777	59.073 ± 11.796
	Five sub-bands	MLP	54.887 ± 13.125	54.887 ± 13.125
HC	Channel-wise	SVM	54.887 ± 13.596	60.597 ± 22.067
	Channel-wise	KNN	53.133 ± 14.997	57.282 ± 18.902
	Channel-wise	MLP	50.627 ± 12.62	50.627 ± 12.620
	Five sub-bands	SVM	58.396 ± 8.815	61.613 ± 28.219
	Five sub-bands	KNN	53.885 ± 10.17	60.556 ± 13.692
	Five sub-bands	MLP	55.89 ± 11.318	55.890 ± 11.318
KFD	Channel-wise	SVM	58.396 ± 10.019	63.085 ± 17.520
	Channel-wise	KNN	54.637 ± 12.833	60.285 ± 13.422
	Channel-wise	MLP	51.128 ± 14.183	51.128 ± 14.183
	Five sub-bands	SVM	52.882 ± 13.552	58.209 ± 26.724
	Five sub-bands	KNN	55.388 ± 15.73	60.328 ± 18.549
	Five sub-bands	MLP	54.135 ± 13.857	54.135 ± 13.857
HFD	Channel-wise	SVM	52.381 ± 14.019	58.008 ± 24.742
	Channel-wise	KNN	51.378 ± 8.920	55.489 ± 16.193
	Channel-wise	MLP	56.391 ± 9.291	56.391 ± 9.291
	Five sub-bands	SVM	58.396 ± 9.374	61.895 ± 28.026
	Five sub-bands	KNN	51.880 ± 9.182	58.808 ± 12.040
	Five sub-bands	MLP	55.514 ± 11.887	55.514 ± 11.887
PSD	Five sub-bands	SVM	57.393 ± 9.184	61.044 ± 28.005
	Five sub-bands	KNN	51.880 ± 13.365	57.965 ± 16.657
	Five sub-bands	MLP	56.140 ± 11.177	56.140 ± 11.177

Valence Classification Performance - 40 Seconds Long Epochs - Individual

Average accuracies and average F1-scores on individual features for arousal using train-test subsets and 40 seconds long epochs

Valence (negative-positive) - 40 seconds epochs - individual features				
Feature	**Method**	**Classifier**	**Average accuracy [%]**	**Average F1-score [%]**
DE	Channel-wise	SVM	55.024 ± 11.529	60.661 ± 23.246
	Channel-wise	KNN	54.545 ± 13.209	59.739 ± 18.281
	Channel-wise	MLP	55.981 ± 13.300	55.981 ± 13.300
	Five sub-bands	SVM	53.11 ± 12.950	58.465 ± 23.553
	Five sub-bands	KNN	47.368 ± 17.819	54.293 ± 20.857
	Five sub-bands	MLP	55.981 ± 10.170	55.981 ± 10.170
HM	Channel-wise	SVM	54.545 ± 12.121	64.058 ± 16.601
	Channel-wise	KNN	55.502 ± 13.852	62.738 ± 14.128
	Channel-wise	MLP	50.239 ± 13.676	50.239 ± 13.676
	Five sub-bands	SVM	54.545 ± 11.338	61.066 ± 28.004
	Five sub-bands	KNN	57.416 ± 10.074	64.505 ± 11.958
	Five sub-bands	MLP	52.632 ± 11.961	52.632 ± 11.961
HC	Channel-wise	SVM	54.545 ± 12.121	64.058 ± 16.601
	Channel-wise	KNN	55.502 ± 13.852	62.738 ± 14.128
	Channel-wise	MLP	50.239 ± 13.676	50.239 ± 13.676
	Five sub-bands	SVM	54.545 ± 11.338	61.066 ± 28.004
	Five sub-bands	KNN	57.416 ± 10.074	64.505 ± 11.958
	Five sub-bands	MLP	52.632 ± 11.961	52.632 ± 11.961
KFD	Channel-wise	SVM	55.502 ± 10.451	63.554 ± 15.173
	Channel-wise	KNN	46.890 ± 12.950	54.748 ± 15.559
	Channel-wise	MLP	56.459 ± 13.409	56.459 ± 13.409
	Five sub-bands	SVM	53.110 ± 11.444	59.440 ± 23.617
	Five sub-bands	KNN	54.545 ± 12.494	63.343 ± 13.224
	Five sub-bands	MLP	56.459 ± 12.339	56.459 ± 12.339
HFD	Channel-wise	SVM	49.761 ± 15.263	56.338 ± 25.100
	Channel-wise	KNN	48.325 ± 10.074	55.245 ± 14.073
	Channel-wise	MLP	53.589 ± 12.456	53.589 ± 12.456
	Five sub-bands	SVM	54.545 ± 11.184	61.066 ± 27.623
	Five sub-bands	KNN	55.263 ± 11.840	62.316 ± 14.853
	Five sub-bands	MLP	53.110 ± 11.867	53.110 ± 11.867
PSD	Five sub-bands	SVM	50.718 ± 11.466	54.441 ± 29.801
	Five sub-bands	KNN	54.067 ± 13.372	58.229 ± 16.827
	Five sub-bands	MLP	53.110 ± 11.839	53.110 ± 11.839

F Detailed Average Accuracies And Average F1-Score of Valence And Arousal Classification For Features Combinations

The best-performing epoch sizes for each feature combination and classification methods for valence
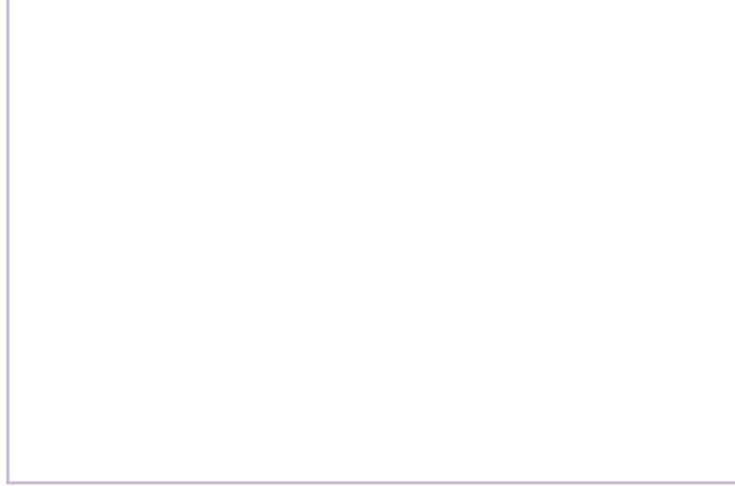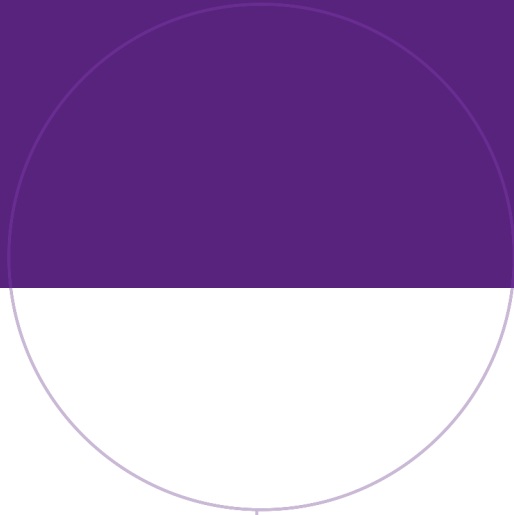
Valence (negative-positive) - combinatorial features				
Features	Classifier	Epoch size	Average accuracy	Average F1-score
HM [CW], HC [CW]	SVM	10sec	60.401 ± 10.232	71.771 ± 22.301
	KNN	10sec	58.396 ± 6.392	64.030 ± 10.900
	MLP	10sec	59.649 ± 9.864	59.649 ± 9.864
DE [SB], PSD[CW]	SVM	5sec	59.461 ± 7.333	67.043 ± 30.851
	KNN	5sec	60.652 ± 4.819	66.634 ± 7.307
	MLP	10sec	59.900 ± 8.409	59.900 ± 8.409
HFD [CW], KFD[CW]	SVM	10sec	62.281 ± 7.787	67.441 ± 17.356
	KNN	10sec	57.707 ± 6.366	63.738 ± 8.894
	MLP	10sec	59.398 ± 10.896	59.398 ± 10.896
DE [SB], HM [CW]	SVM	5sec	60.276 ± 6.06	65.237 ± 32.056
	KNN	5sec	60.902 ± 4.961	66.875 ± 7.295
	MLP	10sec	60.15 ± 10.343	60.15 ± 10.343
HM [CW], HC [CW], HFD [CW],KFD[CW], DE [SB],PSD [SB]	SVM	5sec	59.524 ± 7.241	67.313 ± 30.932
	KNN	5sec	58.772 ± 5.968	64.655 ± 8.521
	MLP	5sec	57.456 ± 7.548	57.456 ± 7.548

Abbrev: SB = sub-bands, and CW = channel-wise

The best-performing epoch sizes for each feature combination and classification methods for arousal

Arousal (calm-excited) - combinatorial features				
Features	Classifier	Epoch size	Average accuracy	Average F1-score
HM [CW], HC [CW]	SVM	10sec	65.664 ± 14.769	71.771 ± 22.301
	KNN	10sec	62.406 ± 13.737	69.458 ± 16.160
	MLP	5sec	64.348 ± 11.855	64.348 ± 11.855
DE [SB], PSD [SB]	SVM	5sec	65.977 ± 11.933	67.043 ± 30.851
	KNN	5sec	66.729 ± 11.192	71.841 ± 14.346
	MLP	5sec	64.098 ± 12.842	64.098 ± 12.842
HFD [CW], KFD[CW]	SVM	5sec	65.977 ± 11.933	72.793 ± 17.490
	KNN	10sec	62.782 ± 14.937	70.067 ± 16.199
	MLP	5sec	64.599 ± 12.629	64.599 ± 12.629
DE [SB], HM [CW]	SVM	40sec	65.072 ± 15.530	67.620 ± 31.445
	KNN	5sec	66.541 ± 11.216	71.666 ± 14.533
	MLP	40sec	66.029 ± 14.796	66.029 ± 14.796
HM [CW], HC [CW], HFD [CW], KFD[CW], DE [SB], PSD [SB]	SVM	20sec	64.662 ± 15.819	69.391 ± 26.704
	KNN	10sec	62.531 ± 14.348	68.874 ± 17.248
	MLP	5sec	65.100 ± 12.056	65.100 ± 12.056

Abbrev: SB = sub-bands, and CW = channel-wise