Master's thesis

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

Embla Celine Stengel Neverlien

Decoding Human Emotions From Video-Elicited EEG Responses With Simple Machine Learning Techniques

With The Design of Experiments and Data Collection

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

NTNU
Norwegian University of
Science and Technology

Embla Celine Stengel Neverlien

Decoding Human Emotions From Video-Elicited EEG Responses With Simple Machine Learning Techniques

With The Design of Experiments and Data Collection

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

Abstract

The potential of EEG-based automatic emotion recognition is immense, from improvement in the treatment of mental health and mental disorders to advancements in the entertainment domain or improving the knowledge of emotional processing in humans. In my specialization project last fall, two quite simple machine-learning models were proposed for emotion recognition. Other researchers have obtained very positive results for EEG-based emotion recognition using complex deep-learning models. The models proposed in the specialization project were comparable with regard to the performance of the more complex models.

The proposed models utilize the features Hjorth mobility and Hjorth complexity in the time and power spectrum domain, frequency band energy, and differential entropy. One of the proposed models has a support vector machine classifier, while the other uses logistic regression. In both models, frequency sub-bands and baseline correction was applied to the data.

This master's thesis is a continuation of the work in my specialization project, with the aim of further improving my models and ensuring their generalization qualities. The hope for this project is to further simplify the complexity of the models by exploring channel reduction using the Non-dominated Sorting Genetic Algorithm (NSGA-II). A new dataset containing EEG recordings of participants that watch emotionally activating stimuli will be collected in this project. There are only a few well-known datasets of this kind available at this time. Creating a new dataset that can become publicly available is therefore a great benefit for the research community. The new dataset will be used to check the generalization properties of the proposed models by analyzing the model's performance on unseen data.

Sammendrag

Potensialet til EEG-basert automatisk gjenkjenning av følelser er enormt, fra å kunne forbedre behandlingen av mentale lidelser og utvide kunnskapen om følelsesprossesering hos mennesker til fremskritt innen underholdning eller å fasililitetere for bedre læringsmiljøer. I mitt spesialiseringsprosjekt høsten 2022 ble det foreslått to relativt enkle maskinlæringsmodeller for følelsesgjenkjenning. Andre forskere har tidligere oppnådd svært positive resultater for EEG-basert følelsesgjenkjenning ved hjelp av komplekse modeller som bruker dyp læring. Ytelsen til modellene som ble foreslått i mitt spesialiseringsprosjekt var sammenlignbar med ytelsen til de beste dyp læring modellene foreslått av andre.

Mine foreslåtte modeller benytter seg av egenskapene Hjorth-mobilitet og Hjorth-kompleksitet i tids- og frekvensdomenet, energi i frekvensbåndet, og differensiell entropi. En av de foreslåtte modellene bruker en "support vector machine"-klassifisering, mens den andre bruker logistisk regresjon med L_2-regularisering. I begge modellene er frekvensundergrupper og grunnlinjekorreksjon anvendt på dataen.

Denne masteroppgaven er en fortsettelse av arbeidet i mitt spesialiseringsprosjekt, med mål om å videre forbedre mine modeller og sikre deres generaliseringskvaliteter. Håpet for dette prosjektet er å ytterligere forenkle kompleksiteten til modellene ved å utforske kanalreduksjon ved hjelp av Non-dominated Sorting Genetic Algorithm (NSGA-II). Et nytt datasett som inneholder EEG-opptak av deltakere som ser emosjonelt aktiverende stimuli, vil bli samlet inn i dette prosjektet. Det er for tiden bare noen få velkjente datasett av denne typen tilgjengelig. Å opprette et nytt datasett som kan bli offentlig tilgjengelig er derfor til stor fordel for forskersamfunnet. Det nye datasettet vil bli brukt til å teste generaliseringsegenskapene til de foreslåtte modellene ved å analysere modellens ytelse på usett data.

Preface

The work presented in this report is part of the TTK4900 - Engineering Cybernetics, Master's Thesis at the Norwegian University of Science and Technology and is worth 30 SP (credits). The work in this thesis is a continuation of the work presented in TTK4550 - Engineering Cybernetics, Specialization Project.

The master's thesis gives students an opportunity to work independently on a large project with everything that entails from creating a project plan, executing that plan, adapting to unforeseen problems and changes, and reporting the work done in a timely and tidy manner. Students also get to exchange knowledge with each other and obtain invaluable knowledge and guidance from supervisors.

The work presented in this project is the result of everything I have learned throughout my five-year journey in Cybernetics and Robotics. While not every aspect of my degree directly influenced this specific project, each course has broadened my knowledge and improved my ability to apply cybernetic methods to solve complex problems. During the specialization project and master's thesis, I obtained valuable experience in biomedical engineering and its immense potential, specifically in EEG-based machine learning methods for medical tools.

Working alongside my exceptionally motivated supervisors on this topic that can have a profound impact on the world has been greatly rewarding. I hope that my work someday can be a small stepping stone for proper emotion recognition tools that can be used in medical treatment. The work I have done over the past year has not only validated my passion for this field but has also solidified my wish to contribute to the advancement of cybernetic engineering in the medical field through my career.

I must give a big thank you to my supervisor, Marta Molinas, and co-supervisor, Mohit Kumar, for important guidance and support throughout this research journey. In particular, I want to express my gratitude to Marta Molinas for her trust in me and for providing me with the opportunity to engage in this highly intriguing project. She has helped me grow both academically and as a person. Additionally, I would like to extend my appreciation to Andres Soler for stepping in as a co-supervisor during the data collection phase. His expertise and support were invaluable during this critical stage.

I would also like to acknowledge the exceptional collaboration of my fellow student, Rose Lu, during the data collection process. Her impeccable cooperation ensured a smooth progression of the project and she was a great support during setbacks. Lastly, I must give a big thank you to my partner, family, and friends for their unwavering support throughout this process.

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Description
HCI	Human-computer interaction
EEG	Electroencephalogram
HC	Hjorth complexity in time domain
HCS	Hjorth complexity in power spectrum domain
HM	Hjorth mobility in time domain
HMS	Hjorth mobility in power spectrum domain
DE	Differential entropy
FBE	Frequency band energy
SVC	Support vector machine classifier
LR	Logistic regression
k-NN	k-Nearest neighbor
NSGA	Non-dominated sorting algorithm
ICA	Independent component analysis
PCA	Principal component analysis
EMDB	Emotional movie database
SCL	Skin conductance level
HR	Heart rate
SAM	Self-assessment manikin
LSL	Lab Streaming Layer
PSD	Power spectral density

1

Introduction

1.1 Problem Description

This thesis aims to improve the prediction accuracy of the EEG-based emotion recognition models proposed in the specialization project [18] by reducing the number of electrodes. With a decrease in the number of channels, the complexity of the model will decrease further. This thesis will also investigate the generalization properties of the proposed models by running them on brand-new data.

1.2 Motivation

Today's focus on human-computer interaction (HCI) is an essential part of the improvement of existing and new technology. Computers and other technological devices are integrated into our daily lives to improve life quality in minor and major ways. In order to extract the maximum potential from computers, the communication between the user and the computer needs to be seamless. Automatic emotion recognition implemented on a computer could improve and even revolutionize HCI.

As humans, we are not always the best at recognizing our own emotional needs. Emotion recognition systems could help provide useful information in order to expand our understanding of ourselves. By studying computational emotion decoding, we can increase our knowledge of how emotions are processed in the brain. There are many possible use cases for this type of technology. A big driving factor is its potential within the field of healthcare. An emotion recognition device could be used to monitor and treat patients with depression, help communicate with and treat patients who are not able to communicate in ordinary ways, help recognize early signs of post-traumatic stress disorder (PTSD), etc. It could be a tool for developing good teaching environments or workspaces where the well-being of all parties is better preserved. And it could be used as a safety mechanism during tasks where the emotional state of a person might affect the risks involved, for example, while driving.

Electroencephalogram (EEG) is a noninvasive, simple method for recording brain signals. From the outside, emotions might present in different ways depending on the person and the culture, but an EEG-based emotion recognition system presents a more universal way of decoding emotions as everybody processes emotions in the brain. Therefore, the possibilities with EEG-based automatic emotion recognition devices are enormous if we are able to create simple, robust, and wearable systems.

An essential part of a simple and wearable system is to develop a simple algorithm that does not need high computational power. To make the device wearable in more everyday scenarios, it should be quite small and comfortable. In order to achieve this, it would be beneficial to minimize the number of electrodes being used. It is also important that the system is able to adapt to different people, so any models created must have good generalization properties across devices and participants.

There are some very well-known datasets in EEG-based emotion recognition research like the SEED (Zheng and Lu [24]] and DEAP (Koelstra et al. [11]) datasets. Many recent papers propose models that achieve great accuracy when working on these datasets. When a dataset has been worked on for too long, there is always the fear of overfitting. A new dataset will provide more data from a different set of participants that experience different kinds of emotional stimuli. By having multiple datasets, researchers can check the generalization properties of

their models further.

1.3 Related Work

As mentioned, the work presented in this master thesis is a continuation of the work in the specialization project [18]. In that work, the features Hjorth complexity in time domain (HC) and as derived from the power spectrum (HCS), Hjorth mobility in time domain (HM) and as derived from the power spectrum (HMS), differential entropy (DE) and frequency band energy (FBE) were explored. The effect of decomposing the signal into sub-bands theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz), and gamma (31-45 Hz) was explored. Information from baseline recordings was also used to correct features in each trial. The project showed that decomposing the signal into frequency sub-bands and later subtracting baseline information from the trial data was very beneficial to the emotion recognition model. The best-performing model consisted of the features HC, HM, DE, FBE, and HMS in combination with the support vector machine classifier (SVC). This model achieved accuracies of 96.50% for high/low valence predictions and 96.71% for high/low arousal predictions with 5-fold cross-validation. Another proposed model utilized all six features with L_2 regularized logistic regression used as the classifier. This model achieved accuracies of 96.26% for valence and 96.61% for arousal, also with 5-fold cross-validation. Related works that justify the choices made in the specialization project can be found in Chapter 1 in [18].

Studies have shown that brain activity in different emotional states behaves very differently from each other and that the activity is individual for each person (Huang et al. [9], Kragel et al. [13], Lindquist et al. [14], Zilio [25]). These individual differences make it so that there is no global optimal selection of electrodes to use. To obtain the best possible solutions, the choice of utilized electrodes should be optimized for each person. This is exactly what Gannouni et al. [7] aimed for in their paper. They improved the performance of the emotion recognition model by implementing an adaptive channel selection method.

Pane et al. [19] used Stepwise Discriminant Analysis (SDA) for channel selection purposes in EEG-based emotion recognition. The 62-channel SEED dataset (created by Zheng and Lu [24]) was utilized in that paper. Pane et al. attempted to locate the 3, 4, 7, and 15 most important channels in each frequency band. The highest accuracy was obtained using 15 channels in each frequency band. They concluded that you can in fact increase the prediction accuracies by decreasing the number of channels. They also concluded that the alpha, beta, and gamma bands are most important for emotion classification.

Moctezuma et al. [17] used the Non-dominated sorting algorithm (NSGA-II) for channel reduction with their two-dimensional CNN model on the well-known DEAP dataset (Koelstra et al. [11]). Using NSGA-II they achieved close to 100% accuracy for some participants using 8 channels for high/low arousal classification and two channels for high/low valence classification.

1.4 Structure of the Report

Chapter 1 has introduced the problem description and motivation for the work in this master's thesis. Related work on which this thesis is built can be read in section 1.3. In chapter 2 the theory behind the methods used in this thesis is explained. The materials used in this project can be found in chapter 3. Chapter 3 also describes the methods in this work. A summary of important results of the specialization thesis and all results of this thesis can be found in chapter 4. The results are further discussed in chapter 5. Finally, the work is summarized and some conclusions are drawn in chapter 6.

2 Theory

2.1 Non-Dominated Sorting Algorithm

The Non-Dominated Sorting Algorithm (NSGA) is an evolutionary sorting algorithm created to solve multi-objective optimization problems, proposed by Srinivas and Deb [21]. NSGA was a computationally expensive algorithm with cost $O(MN^3)$ where M and N is the number of objectives and the population size, respectively. Additionally, the algorithm does not utilize elitism properties and calls for a specified sharing parameter in order to ensure diversity in the population. An improved version of the algorithm was developed by Deb et al. [5] to overcome these obstacles, NSGA-II. With the second version the computational cost is reduced to $O(MN^2)$, the elitism approach is introduced, and the need for setting extra sharing parameters is removed. The elitism approach compares the current population to the best non-dominated solutions found so far. NSGA-II is one of the most widely used algorithms for multi-objective optimization.

The algorithm consists of three steps; non-dominated sorting, crowding distance storing, and tournament mating selection. These steps are repeated until a specified termination criteria is reached.

Firstly the algorithm initializes a population of a specified size N containing N random individuals with different genes. In this project, the genes contain what electrodes are utilized. The M specified objective functions are calculated for each of the N individuals. The two objectives of this project are to maximize the model accuracy and simultaneously minimize the number of electrodes utilized. The non-dominated sorting is performed next. Each solution is put into a rank where rank 1 contains the solutions that are not dominated by any other individuals, rank 2 contains solutions that are dominated by one other solution, and so on. The solution X_1 dominates X_2 if X_1 is no worse than X_2 for all objectives and X_1 is strictly better than X_2 in at least one objective.

The individuals for the next generation are chosen according to their rank and will make up a mating pool. First, the solutions of rank 1 are included, followed by rank 2, and so on until a mating pool of size N is reached. If a full rank can not fit in the mating pool without the population exceeding N, crowding distance sorting will be performed to select what individuals in that rank to include in the mating pool.

Crowding distance provides an estimate of the density of solutions around a given solution. Solutions that are less dense are preferred. The algorithm for crowding distance sorting works the following way. First, the solutions are sorted in ascending order in terms of the objective function value. The crowding distance value, which is the average distance between the two neighboring solutions, is calculated for each solution. The upper and lower boundary solutions are given an infinite crowding distance value to always be included in the mating pool for the next generation. This process is repeated for all objective functions. A solution's final crowding distance value is found by adding the crowding distance values of all the objectives for that solution. The solutions with the highest crowding distance are chosen to be included in the mating pool until the population of size N is reached. Crowding distance sorting is done to ensure a big spread in the solutions.

From the mating pool, some off-spring individuals are created by means of binary tournament selection, crossover, and mutation. In the tournament selection, two random individuals are compared in terms of their rank and crowding distance. The individual with the lower rank is passed on to be a parent. The individual with the higher crowding distance is chosen if their rank is equal. This process is repeated with two new random individuals to find the second parent. Crossover is performed between the two parents in order to combine their genes. This results in two off-spring individuals, one off-spring having half of each parent's genes and the other having the remaining half's. Additionally, a specified type of mutation of the genes will occur with a specified probability

p_m.

This process will repeat until a termination criteria is reached. In summary, the pseudo-code for the process is:

1. Set hyperparameters: population size N, mutation probability p_m, crossover rate p_c, and termination parameters.
    - N: the size of the population in each generation.
    - p_m: the probability that an individual will undergo mutation.
    - p_c: the probability that two individuals will swap bits.
    - The termination parameters can be defined many ways, but is defined as a set number of generations in this project.
2. Randomly initialize a population of size N.
3. Non-dominated sorting of the individuals into rank.
4. Crowding distance sorting to ensure spread in the solutions.
5. Binary tournament selection in order to select the preferred parents.
6. Crossover and polynomial mutation to create off-spring.
7. Step 3-6 is repeated in a manner that always keeps the mating pool at size N until the termination criteria are met.

The goal of NSGA-II is to find the Pareto optimal solutions, i.e. the solutions in which you can not improve one objective function without degrading one or more other objective values. The algorithm can not differentiate between these Pareto optimal solutions in terms of which is better to solve your problem. The multi-objective optimization problem presented in this project will maximize the accuracy of the emotion recognition model while minimizing the number of electrodes utilized. This will present a trade-off scenario between the number of electrodes and the obtainable model accuracy.

2.2 Independent Component analysis

The objective of the independent component analysis (ICA) is to separate a mixture of signals into its pure signal sources. The method is often explained with the *cocktail party problem* where you wish to separate the speech of one person from the room noise. After separating the signal sources one might inspect the information contained in each of these sources, and consider keeping or disregarding that signal. This way ICA can be used for dimensionality reduction or to filter out artifacts.

There are many algorithms that perform ICA in different ways. The one used in this project is the FastICA algorithm, which was developed by Hyvärinen and Oja [10]. This algorithm for ICA works by maximizing the non-Gaussianity. This is a measure of how far a random variable is from being Gaussian distributed. It can be measured by the Kurtosis and negative entropy. The independent source signals can be found by maximizing the Kurtosis. The Kurtosis (K) of x is defined as in eq. (2.1).

$$K(x) = E[x^4] - 3E[E[x^2]]^2 \tag{2.1}$$

, where x is assumed to be of unit variance, so the equation simplifies to eq. (2.2).

$$K(x) = E(x^4) - 3 \tag{2.2}$$

Kurtosis is very simple to calculate, but it is unfortunately very sensitive to outliers and therefore is not a robust measure of non-Gaussianity.

The negative entropy is, unsurprisingly, related to entropy. Entropy is a measure of the degree of information a random variable gives. A variable that is very unpredictable and unstructured will have high entropy. The entropy (H) for a discrete random variable X is defined as eq. (2.3).

$$H(X) = -\sum_i P(X = a_i) \log P(X = a_i) \tag{2.3}$$

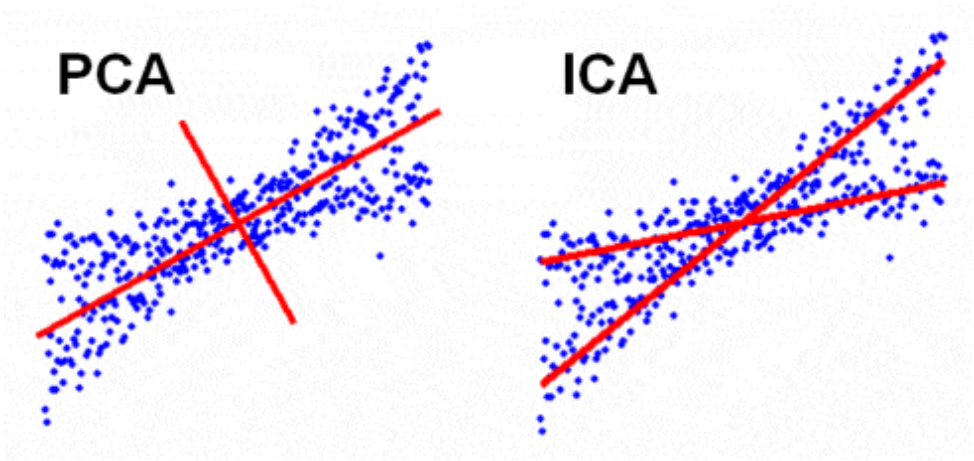


Figure 2.1: PCA (left) maximizes the variation along the axes. ICA (right) minimizes the mutual information across axes. The picture is taken from Unho Choi [23].

, where a_i are the possible values of X. When the definition is generalized for continuous-valued variables or vectors it is usually called differential entropy (H). eq. (2.4) gives the definition of the differential entropy of a random vector **x** with density $f(\mathbf{x})$.

$$H(\mathbf{x}) = -\int f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \tag{2.4}$$

Since a more "random" variable will have higher entropy, a Gaussian variable will have the largest entropy of variables of equal variance. The entropy will be large for the Gaussian distribution and small for variables that are clearly clustered or have a very "spiky" probability density function". The negative entropy J, defined in **??** is used for the non-Gaussianity variable as this will always be non-negative and be zero if and only if **x** has a Gaussian distribution.

$$J(\mathbf{x}) = H(\mathbf{x}_{gauss}) - H(\mathbf{x}) \tag{2.5}$$

, where $\mathbf{x}_{gauss}$ is a Gaussian random variable of the same covariance matrix as **x**. The negative entropy is a very good estimator of non-Gaussianity, but it is very computationally difficult as it requires an estimate of the probability density function. Because of this, the negative entropy is not directly used as a measurement of the non-Gaussianity. Hyvärinen and Oja [10] found that for practically all non-quadratic functions G, the negative entropy of x could be approximated eq. (2.6)

$$J(x) \propto [EG(x) - EG(v)]^2 \tag{2.6}$$

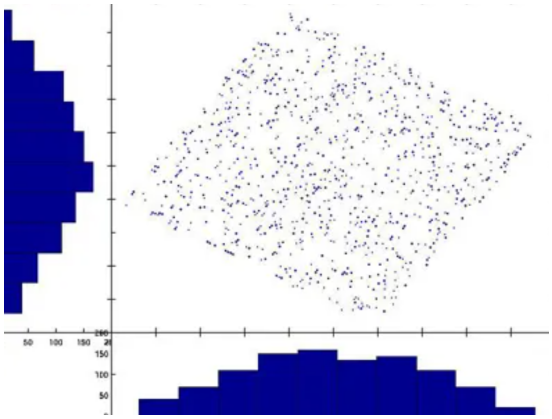
, where v is a standardized Gaussian variable, i.e. it has zero mean and unit variance. Since FastICA works by maximizing the non-Gaussianity, it can not separate Gaussian signals. Therefore, at most one of the mixed signals can be Gaussian distributed.

ICA can also be performed by minimizing the mutual information or with maximum likelihood estimation. Some whitening of the signals (often through Principal component analysis (PCA) is first applied in order to remove correlations before performing ICA.

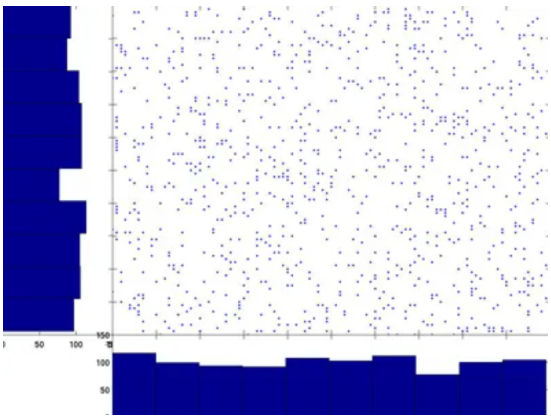
While PCA finds the axes in which the variation of the input data is maximized, ICA minimizes the mutual information. See

In short, ICA works under the assumption that the mixture of two independent random variables will be more Gaussian than the original variables. Visually, ICA can be explained to perform a rotation of the whitened mixture of signals in order to minimize the Gaussian distribution along the axes. This will provide the most independent signals, that are most likely the sources. This is visualized in fig. 2.2.

Because of these properties, ICA is a good tool for locating for example blinking artifacts in EEG signals. Blinking will show up clearly in the EEG signals, but is usually not of interest when analyzing EEG. Eye blinks happens independently of any emotion one might feel. The blinking rate might change somewhat depending on the emotional state, but the epochs one look at during automatic emotion recognition will not be long enough to capture this. Because eye blinks come from an independent source, their signal contribution can be located and removed with ICA.



(a) Whitened linear mixture of two sources. The distribution along both axes looks quite similar to the Gaussian distribution.



(b) After ICA has performed a rotation in order to minimize the Gaussianity of the projection.

Figure 2.2: ICA minimizes the Gaussianity along the axes. The pictures are taken from Arnaud Delrome [1].

2.3 Generalization

The generalization quality of a model is the model's ability to perform well on unseen data. When a machine learning model is created it is very easy to overfit the model to the data it is trained on. The data available might also be "too clean" compared to real-world scenarios, or only capture a few of many instances that might occur. It is therefore important to check the generalization qualities of the model, to ensure that the performance is expected to be about the same in real-world scenarios.

One way to do this is by extracting a test set from the data, that is not used on the model before the final testing of the performance. If the model performance on the test set is about the same as during training and tuning, this is a good indication that the model has learned the underlying patterns for accurate predictions.

In this project, EEG data has been collected. This is explained further in section 3.5. Up until this point, only the preprocessed DEAP dataset has been used. The new data is collected on another system, which is a good test for generalization. It is desirable for the model to work well on all EEG data, no matter the EEG recording system, so that the model can be used on all kinds of systems around the world. Additionally, the new dataset is collected from new participants. It is well-known that emotional processing is very individual, so having more people to test the model on ensures that is it quite robust for these individual differences.

3

Materials and Methods

The model used and further developed in this thesis is built upon the work done in TTK4550 - Engineering Cybernetics, Specialization Project [18] at the Norwegian University of Science and Technology. Section 3.2 summarizes the methods used in that project. For a more in-depth explanation, see Chapter 3 "Materials and Methods" in the report [18].

3.1 DEAP dataset

The dataset utilized in the specialization project was the preprocessed publicly available DEAP dataset Koelstra et al. [11]. The dataset contains recordings with a sampling rate of 512 Hz of 8 peripheral physiological signals and 32 EEG channels. The data was collected during an experiment in which participants would watch 40 one-minute music videos, selected to evoke emotions that would span the arousal-valence plane.

The music videos used as stimuli in the recordings were chosen through analysis of affective tags found on the website Last.fm[1] followed by video highlight detection, and lastly an online self-assessment method.

The preprocessed EEG data has been downsampled to 128Hz and have EOG artifacts removed. A 4.0-45.0Hz bandpass frequency filter has been applied to the signals and the data has been averaged to the common reference.

The EEG channels were placed according to the international 10-20 system. The 32 EEG channels present in the dataset are depicted in fig. 3.1.

During the experiment, the participants would watch a music video followed by a self-assessment of their levels of arousal, valence, dominance, and liking. A five-second baseline recording was collected before every trial, in which three seconds are included in the preprocessed dataset.

3.2 Building the model

The EEG data was imported and segmented into one-second epochs. This resulted in trial data on the structure 2400x32x128 (epoch x channel x data point) and baseline data of size 120x32x128 (epoch x channel x data point). Next, the signals were decomposed into the four frequency sub-bands theta (4-8Hz), alpha (8-14Hz), beta (14-31Hz), and gamma (31-45Hz). The effect of implementing these sub-bands was explored.

Six features were extracted from every epoch in each sub-band. These features were Hjorth complexity calculated in time domain (HC) and from the power spectrum (HCS), Hjorth mobility calculated in the time domain (HM) and from the power spectrum (HMS), frequency-band energy (FBE), and differential entropy (DE). This process produced feature matrices for the trial data of size 2400x128 (epoch x feature). The benefits of baseline correction were investigated in the following way. Baseline features were extracted from the three baseline epochs before each trial. All trial features were baseline corrected by subtracting the average value of the corresponding three baseline features.

Lastly, the effects of combining multiple features in one model were explored with multiple classifiers: Support Vector Machine Classifier (SVC), k-Nearest Neighbor (k-NN), logistic regression (LR) with L_1 and L_2 regularization, decision tree, and random forest. The two best-performing models in the specialization project utilized

[1] `http://last.fm`

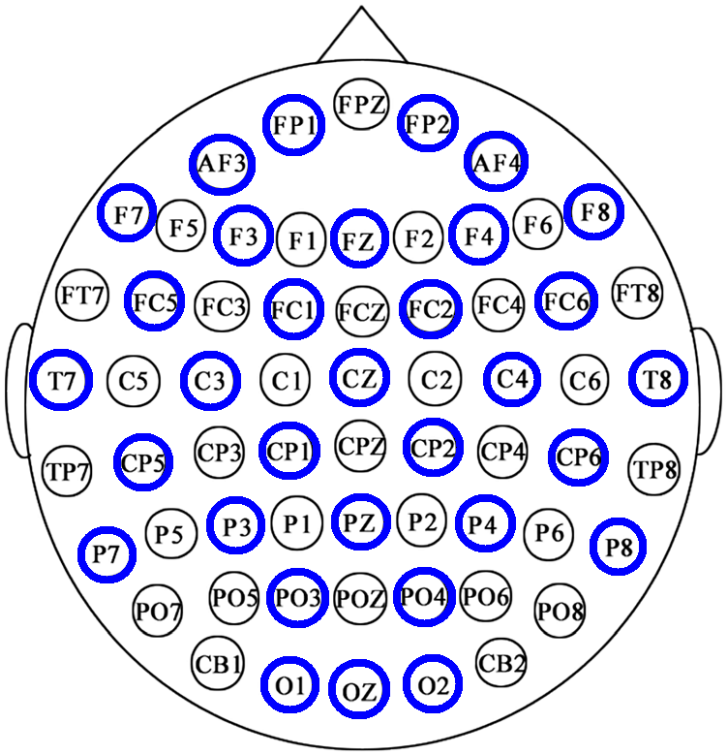


Figure 3.1: Location of the electrodes utilized (marked with blue outline) in the DEAP dataset.

SVC and L_2-regularized LR with multiple features and are listed with more specifics in table 3.1. These models are further investigated in this project. See section 4.1 for important results from the specialization project.

Classifier	Hyperparameter	Features
Support vector machine	Kernel= 'rbf'	HC, HM, DE, FBE, HMS
Logistic regression with penalty L_2	C = 1	HC, HM, DE, FBE, HMS, HCS

Table 3.1: Specification for the two best-performing models from the specialization project.

3.3 Non-dominated Sorting Genetic Algorithm on DEAP dataset

The DEAP dataset consists of EEG recordings from 32 electrodes. In the proposed models five or six types of features are used for classification. Additionally, all signals are decomposed into four frequency bands. This results in 640 features in total for the SVC model and 768 features for the (LR) model. The Non-dominated Sorting Genetic Algorithm (NSGA-II) is applied to the data to see if the same or better performance can be achieved with fewer channels.

The NSGA-II is implemented using the pymoo framework for multi-objective optimization in Python created by Blank and Deb [3]. The objectives of the optimization problem were defined to maximize the accuracy of the model while minimizing the number of utilized electrodes. Different settings were explored for the population and number of generations, and the most important are listed in table 3.2. The NSGA-II is run for high/low arousal and high/low valence classification separately. This is done to examine the difference in electrode placement to achieve good arousal predictions and good valence predictions.

Having a large population size and a large number of generations simultaneously was not possible because of the time constraints of the project. As the algorithm explores many possible solutions, the algorithm is quite computationally heavy. The importance of having a larger population or running for more generations was explored.

The variables in the NSGA-II are defined as an array of 32 binary values, one for each electrode, called $array\_channels$. If the value of the first element is 1, then the first channel is utilized in the model. If it is 0, it is removed from the data. As mentioned, one objective of the NSGA-II was to minimize $\sum array\_channels$.

The accuracy of the model, $model\_accuracy$, is computed for every configuration of the $array\_channels$ in the population. The second objective of the NSGA-II is then to minimize $1 - model\_accuracy$.

The optimization problem is then defined as in eq. (3.1).

$$\begin{aligned} &min\ \ inaccuracy = 1 - model\_accuracy(array\_channels) \\ &min\ \ number\ of\ channels = \sum array\_channels \\ &array\_channels\ \epsilon\ \{0, 1\} \end{aligned} \tag{3.1}$$

,where $i\epsilon[1\ ..\ 32]$ is the index of the i'th channel in $array\_channels$.

Polynomial mutation with probability = $\frac{1}{number\ of\ objectives} = \frac{1}{2}$ was implemented. A rounding repair characteristic was added to the mutation in order to keep the variables in $array\_channels$ binary. The sampling of the population was chosen randomly with values 0 or 1 for all elements in all $array\_channels$. Additionally, a two-point crossover with probability =$\frac{1}{2}$ was performed.

The algorithm was tried without any constraints, and also with a constraint of a defined number of channels. The constrained problem is defined as eq. (3.2):

$$\begin{aligned} &min\ \ inaccuracy = 1 - model\_accuracy(array\_channels) \\ &\sum array\_channels = x \\ &array\_channels[i]\ \epsilon\ \{0, 1\} \end{aligned} \tag{3.2}$$

, where $x\epsilon[0\ ..\ 32]$ defines the number of channels in the solutions. This optimization problem only has one objective, so the probability of the polynomial mutation is changed to 1.

Population size	Number of generations	Constraints
50	50	-
20	120	-
35	70	-
70	35	-
50	50	$\sum array\_channels = 16$
35	70	$\sum array\_channels = 16$
50	50	$\sum array\_channels = 25$
50	50	$\sum array\_channels = 29$

Table 3.2: Settings for population and generation size that were explored in NSGA-II.

Test and validation sets, each of sizes 15%, were extracted randomly from the data before running the NSGA-II. The NSGA-II was run participant-wise and for high/low arousal and high/low valence classification separately but with the same settings. All solutions in the Pareto front were saved for each participant. For the unconstrained problems, the five solutions with the highest accuracies were tested on the test set. For the constrained problems, the first five solutions in the Pareto front were tested using the test set. In both cases, the channel selection that provided the best accuracy was saved alongside the accuracies. Finally, the channel specifications that provided the best accuracy on the test set were run on the validation set. This is done in order to check that the model is not overfitted to the data. The channel selection process is visualized in fig. 3.2. A 5-fold cross-validation was also run over all the data with the same settings, to have results that could be compared to the results obtained in the specialization project. The mean accuracies for each participant were then used to analyze the effect of the channel reduction.

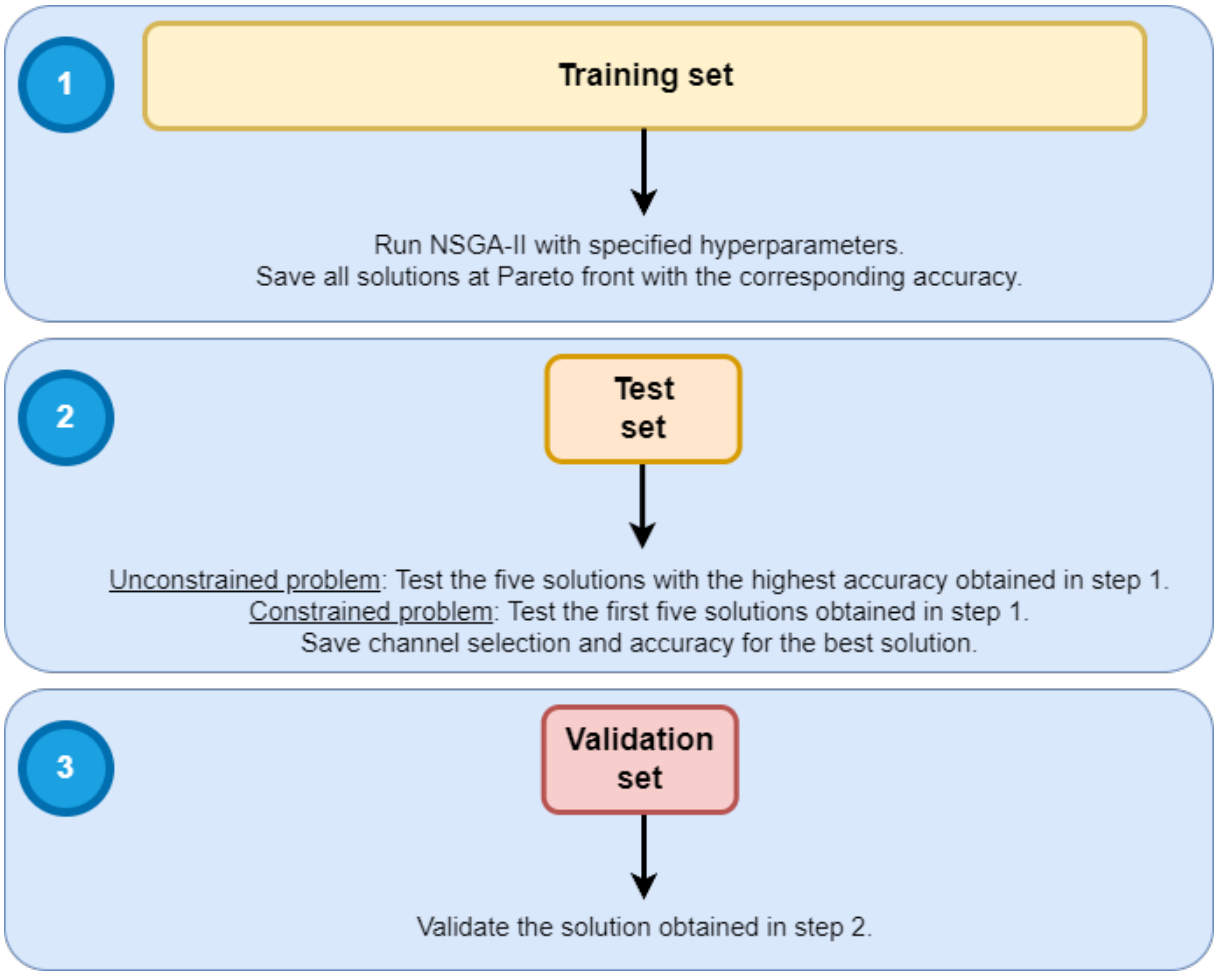


Figure 3.2: The three steps for channel selection with NSGA-II. This process was run participant-wise.

3.4 The Emotional Movie Database

In this project, the Emotional Movie Database (EMDB) is utilized as stimuli during data collection. The database was created by Carvalho et al. [4] in their efforts to create a collection of affective non-auditory movie clips which would evoke emotions that would span the multidimensional space of emotions consisting of valence, arousal, and dominance. The EMDB contains 40-second movie clips in six different genres: horror, erotic, social positive, social negative, scenery, and object manipulation. All categories contain ten videos, except object manipulations which only consist of two movie clips, i.e. the database consist of 52 clips in total.

The study consisted of three steps: (1) selection of movie clips, followed by film clip validation through (2) self-assessment and (3) psychophysiological signals. First, 127 film clips that fulfilled some criteria were selected by researchers. The clips had to portray a stable context, have a continuous presence of any people in the scene, and elicit constant valence throughout the scene (i.e. no scene should evoke both positive and negative feelings). The 40-second film clips were edited to ensure that these criteria were followed as closely as possible. 11 participants rated the film clips using the Self-Assessment Manikin (SAM) method. The SAM rating system describes feelings on a 1-9 scale for arousal, valence, and dominance. Arousal describes the intensity of the feeling, valence describes the pleasantness of the feeling, and dominance is a measurement of whether the participant feels in control of the feeling or not. The SAM scales are pictured in Based on these ratings, 52 movie clips were selected. The movie clips were selected based on (I) the variability in the rating scores, (II) the average valence score being high (close to 9), neutral (close to 5), or low (close to 1), and (III) the average arousal score being high (close to 9) or low (close to 1).

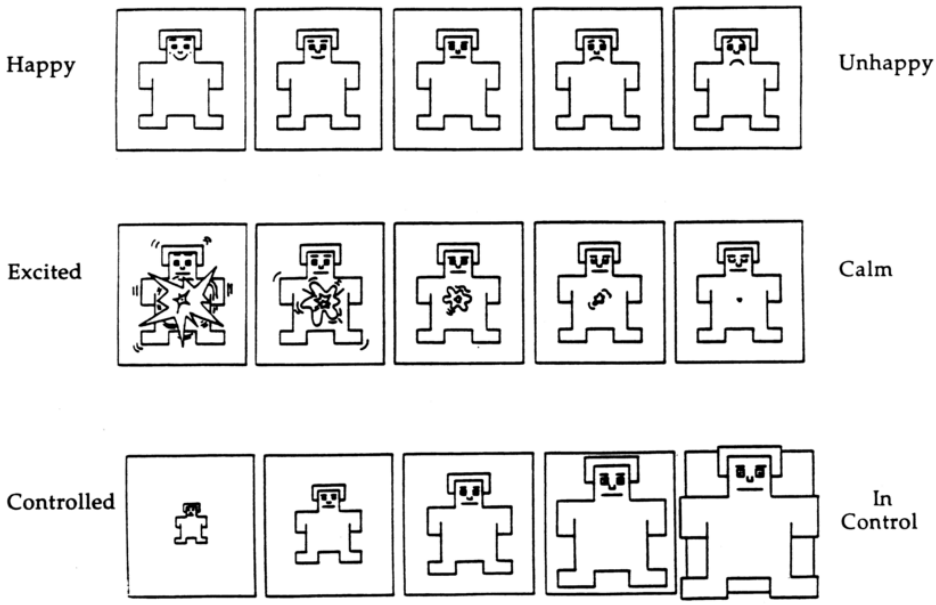


Figure 3.3: The scales in Self-Assessment Manikin. (Upper) valence, (middle) arousal, (lower) dominance.

The horror clips depict people in horrifying, life-threatening situations, the erotic clips showcase heterosexual sex (with no genitalia exposure), the social positive and social negative clips consist of happy and sad/angry social interactions respectively, the scenery clips showcase natural scenery or landscapes and the object manipulation category are clips show a hand pushing small objects on a table.

For phase (2), 113 participants (75 female and 38 male) watched the movie clips. After each movie clip self-assessment using SAM was conducted, in which the participants rated the experience using the SAM method. Additionally, they had to answer the questions ”Have you watched this movie before?” and ”Have you closed your eyes or looked away during the clip presentation?”. If a movie clip had previously been seen by more than 30% of the participants, the clip would be discarded. The movie clips were shown in a pseudo-random order so that no two consecutive clips were from the same category.

Phase (3) was performed to assess the skin conductance level (SCL) and heart rate (HR) response the chosen clips would elicit. SCL and HR were recorded from 32 right-handed healthy participants (16 females, 16 males) who watched the 52 movie clips.

3.5 Data collection

For this part of the project, I collaborated with another Master's student that is working on a similar thesis. The main objective of the data collection is to create a new dataset that can be used for EEG-based emotion recognition. The collected dataset will be used to test the model molded thus far on new unseen data. The performance of the model on this new dataset will give a good indication of its generalization properties.

The participants watched the 52 videos contained in the EMDB. This movie database was chosen as the movies should elicit emotions spanning the arousal/valence/dominance space, see section 3.4 for an explanation. After each video, the participants are asked to rate their emotions using the widely known self-assessment manikin (SAM) scale. EEG is collected during the entire experiment.

Before enlisting participants, they received an information letter attached in **??** . The information letter explained the procedures of the experiments, the purpose of the data collection, information about handling personal information, and the criteria you must fulfill to volunteer as a participant. It also stated that you are free to withdraw at any time without giving a reason, free of penalty. Each participant was compensated with a gift card. After reading the information letter the participants signed a letter of consent.

All participants had to be between the age of 20 and 30 years old. We determined that it would be beneficial to limit the age range as the brain changes throughout a person's lifespan, and we do not want to introduce too many unknown variables. This interval was chosen specifically because it is similar to the ages of the participants in the DEAP and SEED dataset and therefore makes for a good comparison to these well-known datasets. People were

also asked to avoid participating if they have neurological diseases or use strong medicine or drugs as this affects the emotional response.

The data collected in this project was recorded on two devices: the Mentalab Explore+ 32 and the Unicorn Hybrid Black. 7 people (4 female and 3 male) participated in the data collection with the Mentalab system. The average age of these participants was 25,3 with a standard deviation of 1,3. Data from 20 participants (10 female and 10 male) was collected for the Unicorn system. The average age of these participants was 23,8 with a standard deviation of 0,9.

3.5.1 Experimental protocol

The experimental protocol was written by F. Arevalo (M. Molinas 2023, personal communication, 10 February) in Python and run on PsychoPy [2]. Markers (1: fixation-cross start, 2: video start, 3: video end) were added to the original protocol and sent to Lab Recorder via the Python to the Lab Streaming Layer (LSL) interface created mainly by Kothe [12]. The experiment would start with some information about the scales of arousal, valence, and dominance. The participants were specifically asked to give an honest rating of their experienced emotions during the stimuli, without being influenced by what they think they *should* feel. The protocol consists of 52 trials where each trial starts with 8 seconds of cross-fixation. This was followed by one 40-second video. Lastly, the participants would rate emotions on the SAM scale. The trial would end with 15 seconds of rest before the next trial started. The participants were informed not to move during the cross-fixation and the video stimuli. They were allowed to move and ask practical questions during the self-evaluation and break.

The experiment was conducted on a Dell UltraSharp 38" Curved Monitor. The screen was quite reflective, so we dimmed the lights in order to decrease this problem. The participants sat centered in front of the screen, at a distance they felt comfortable (around one meter). The height of the screen was adjusted to 90 degrees in relation to their eyes.

Two EEG recording devices were used in order to collect the data in this project, due to some unexpected problems. For the first seven participants, the Mentalab Explore+ 32 amplifier with wet electrodes was used. The rest of the recording was done using two 8-channel Unicorn Hybrid Black devices with dry electrodes. The setup process for the two systems was quite different, so both processes are explained in below.

The EEG recording devices were connected to the computer via Bluetooth, and the EEG signals were pushed to Lab Streaming Layer (LSL). The Psychopy protocol was also pushed to LSL. Everything was recorded through the Lab Recorder in order to synchronize the signals and the stimuli.

Mentalab Explore+ 32

The BrainCap-SL created by EASYCAP and wet electrodes were utilized with the 32-channel Mentalab Explore+ amplifier. Pictures of the electrodes and the cap can be found in fig. 3.4. The amplifier is pictured in fig. 3.5. A conductive gel must be used with the wet electrodes. The Signa Gel by Parker Labs (fig. 3.6b) was used on all electrodes around the scalp. The reference node was placed on the ear lobe, and for this electrode, the Electro-Gel produced by Electro-Cap International (fig. 3.6a) was used. The placements of the electrodes follow the 10-10 system. The channel to electrode position mapping can be found in table 3.3 and the electrode positions are visualized in fig. 3.8. The cap utilized in this setup had these 32 electrode positions. The electrodes span most of the scalp, which is beneficial for emotion recognition purposes.

[2] `https://psychopy.org/index.html`

(a) Picture of the wet electrodes used with the Mentalab Explore+ 32 system. Picture is taken from Mentalab [16]

(b) The BrainCap-SL is created by EASYCAP. Here it is pictured with the electrodes inserted.

Figure 3.4: Cap and electrodes used in the Mentalab Explore+ 32 setup.

(a)

(b)

Figure 3.5: Pictures of the Mentalab Explore+ 32 amplifier. The pictures are taken from Mentalab [15].

The procedure for putting on the EEG cap was as follows:

1. Measure the distance over the scalp from Nz to Iz and mark the middle of this distance, see fig. 3.7. Measure the distance between the ears across the marked point. The middle of this distance will be the center of the scalp. The cap is placed with the electrode Cz at the center of the scalp.
2. Remove electrodes and prep the skin/scalp with alcohol at each electrode position. Clean the ear lobe with alcohol. Reattach the electrodes.
3. Apply gel to the ear lobe. We experienced that the signals would become better as the gel started to give moisture to the skin, so it was beneficial to do this early.
4. Apply gel to all electrodes on the scalp.
5. Tape the reference electrode to the ear lobe. From our experience, very light pressure on the reference electrode was preferable.
6. Try to optimize and then assess the impedance of the electrodes placed on the forehead. This was a good indicator of what kind of impedance we could expect to achieve on the other electrodes.
7. Improve the impedance on the rest of the electrodes such that they are all within the same range. This was done by moving the gel around and sometimes applying more.

8. Check if we could improve the connection at the reference, as the resistance at this location will affect the impedance measured at all locations.

9. Lastly, we check the signals visually. The participants were asked to sit still and blink a few times and scrunch their faces. This is what we expected to see:

   - The amplitude of the signals where around 100 mV.
   - The blinking showed up in the signals, especially at the channels located at the front of the head.
   - The scrunching of the face showed up in the signals on all channels.

   If the signals behaved as expected at all channels, we started the experiment. If not, we would try to improve the connections between the electrodes and the scalp.

(a) Picture of the conductive Electro-Cap International Electro-Gel used on the reference electrode. Picture taken from bio-medical [2]

(b) Picture of the conductive Parker Labs Signa gel used on the scalp electrodes. Picture taken from TerniMed [22]

Figure 3.6: Pictres of the conducive gels used with the Mentalab Explore+ 32 setup.

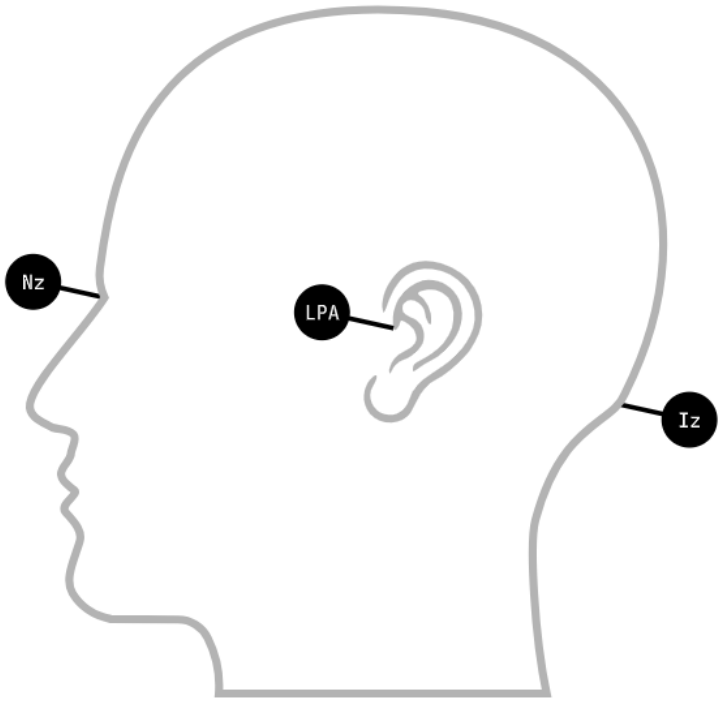


Figure 3.7: Nz is right above the bridge of the nose and Iz is on the back of the head. The distance between the points is measured across the scalp.

Channel no.	Location	Channel no.	Location
1	Fp1	17	CP1
2	F8	18	CP2
3	FT10	19	Pz
4	F4	20	O1
5	FC6	21	Oz
6	T8	22	O2
7	C4	23	Fp1
8	CP6	24	F7
9	P9	25	FT9
10	P4	26	F3
11	AFz	27	FC5
12	Fz	28	T8
13	FC1	29	C3
14	FCz	30	CP5
15	FC2	31	P7
16	Cz	32	P3

Table 3.3: The overview of what EEG channel is connected to each location for the Mentalab Explore+ 32 setup.

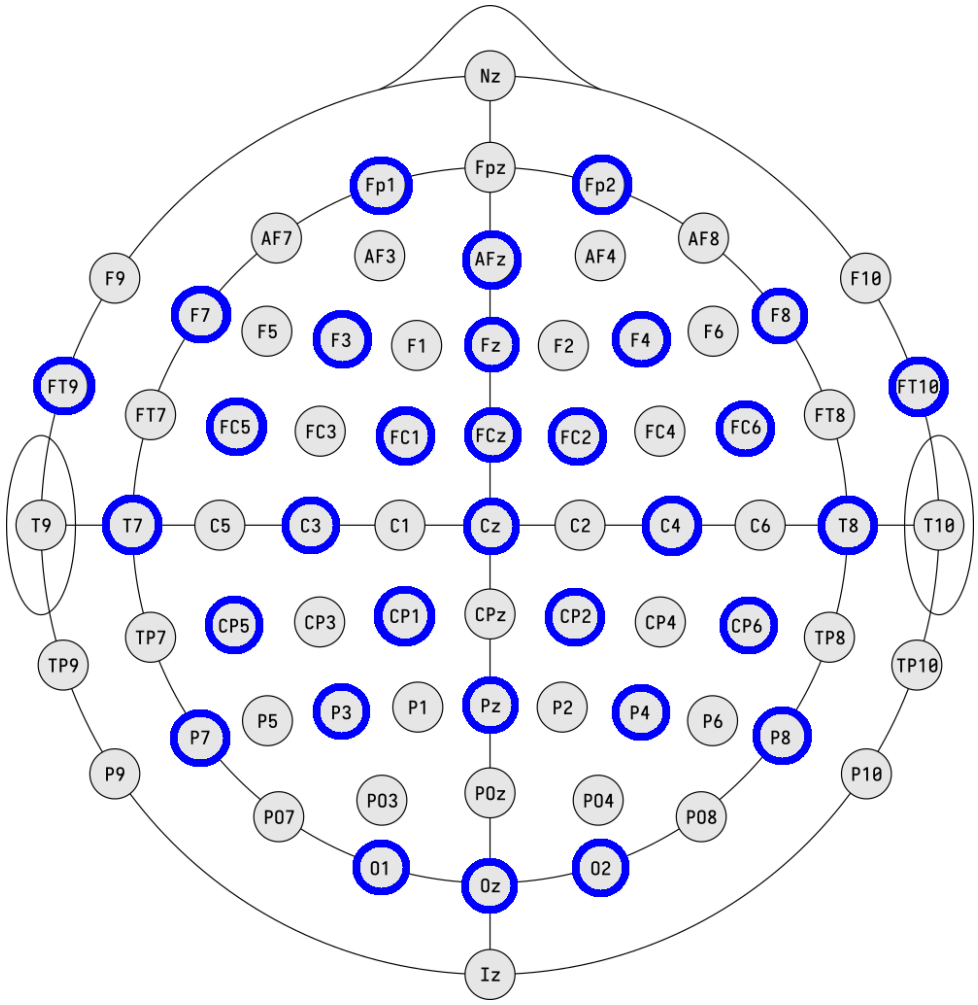


Figure 3.8: The figure shows the extended 10-20 EEG location system. The locations marked in blue are utilized in the experiments with the Mentalab Explore+ 32 setup. Picture source for unedited version [20].

Unicorn Hybrid Black

The Unicorn Hybrid Black EEG headset with hybrid EEG electrodes was utilized for the last 20 participants. The electrodes allow for recording dry or with gel and we opted for dry recordings. The amplifier and electrodes are pictured in fig. 3.10. This system has 8 channels, so we combined two devices in order to record at 16 channels in each experiment. This was done simply by stacking the devices at the back of the head using an elastic. The devices are called '68' and '69'. In order to utilize 16 channels simultaneously, the Unicorn Hybrid Black devices were connected to the g.GAMMAcap^2 created by g.tec. A picture of g.GAMMAcap^2 can be found in fig. 3.11. Both the amplifier and cap are created by g.tec medical engineering. Since the number of electrodes utilized in the experiment decreased from 32 to 16 channels, some related works about channel selection in EEG-based emotion recognition were examined. Pane et al. [19] used stepwise discrimination analysis to find the 15 most important channels for emotion recognition in each frequency sub-band delta, theta, alpha, beta, and gamma in the SEED dataset [24]. As gamma is the most important frequency band to classify emotions, the 15 electrode positions found to be most important in this band were chosen for our experiment. Additionally, the location Cz at the center of the head was included, in order to have 16 channels in total. The electrode positions are visualized in fig. 3.11. The mapping of the channels to electrode positions can be found in table 3.4.

Figure 3.9: The g.GAMMAcap^2 was used with the Unicorn Hybrid Black system in order to record on 16 channels simultaneously. The picture is taken from g.tec medical engineering [6].

(a) Picture of Unicorn Hybrid Black amplifier.

(b) Picture of the hybrid electrodes used with the Unicorn Hybrid Black system and how they are connected to the amplifier.

Figure 3.10: Pictures are taken from g.tec medical engineering [8].

The setup procedure was as follows:

1. Place the cap on the participant's head with Cz at the center of the scalp.
2. Massage all electrodes into the scalp so the teeth on the electrodes can get past the hair and be in contact with the scalp.
3. Clean the earlobes with alcohol, apply conductive gel, and attach the reference electrodes. Each device has one right and one left reference. One reference electrode was placed at the front and one at the back of the earlobe, at both ears.
4. There is no way to check the channel impedances with this system, so the EEG signals were examined visually in the same manner as with the Mentalab Explore+ system. If any channels behaved differently than expected, we tried to achieve better contact between the electrode and the scalp.

5. When the signals looked as expected at all channels, we started the experiment.

Device ’68’		Device ’69’	
Channel no.	Location	Channel no.	Location
1	Fp1	1	FP2
2	Fpz	2	F8
3	F5	3	F4
4	FC1	4	FC6
5	CP1	5	FC4
6	P1	6	CP6
7	PO3	7	P8
8	PO5	8	Cz

Table 3.4: The overview of what EEG channel is connected to each location for the Unicorn Hybrid Black setup.

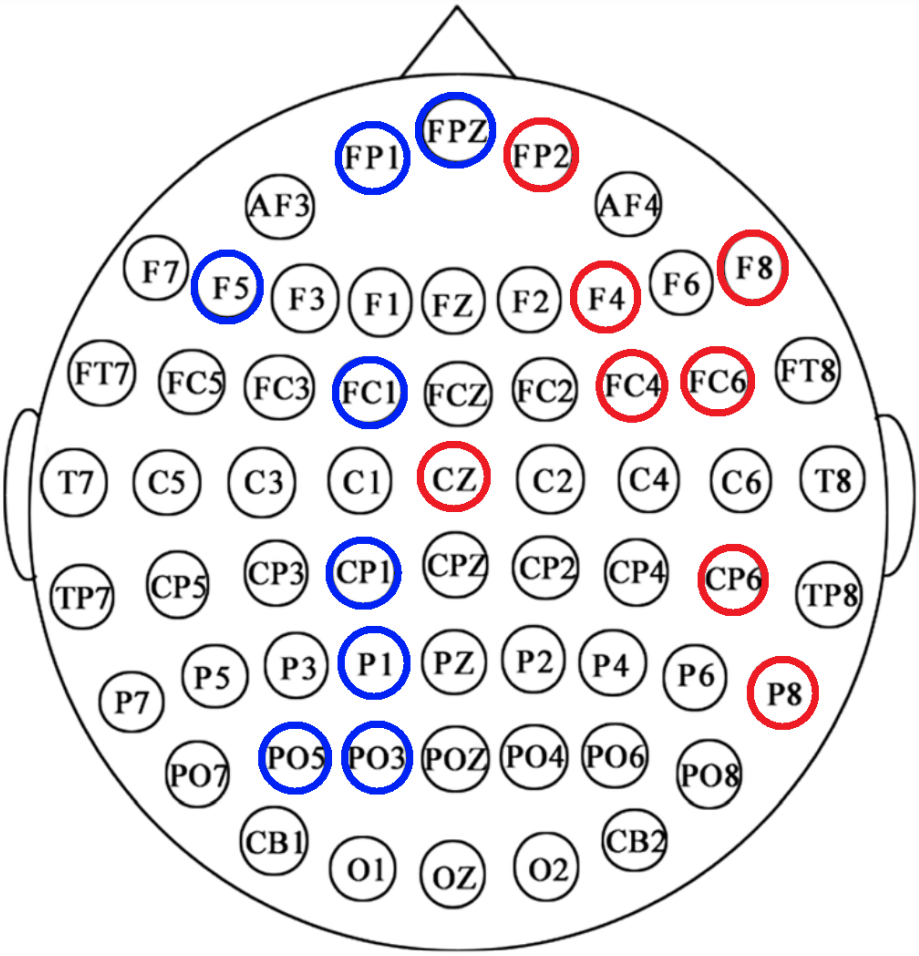


Figure 3.11: The figure shows the extended 10-20 EEG location system. The locations marked in blue are connected to device ’68’ and the electrodes marked in red are connected to device ’69’, with the Unicorn Hybrid Black setup.

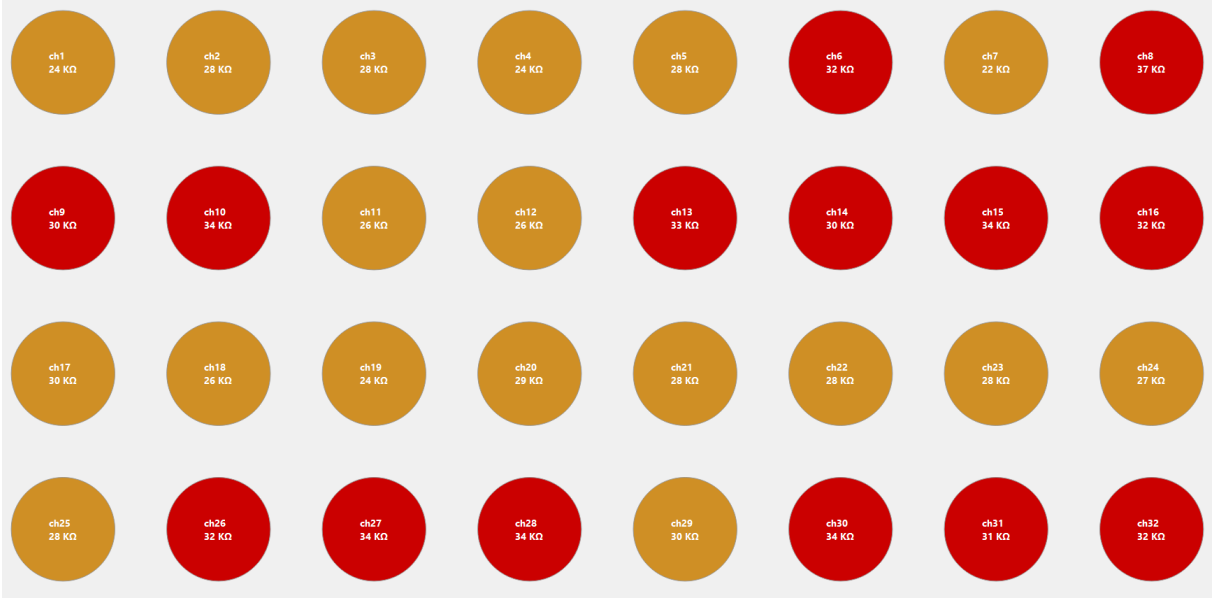
Figure 3.12: Picture of the experimental setup. The lights were turned off before the experiment began to minimize glare on the computer screen. This picture shows the Unicorn Hybrid Black system, but the setup was the same for both devices.

3.5.2 Problems and limitations

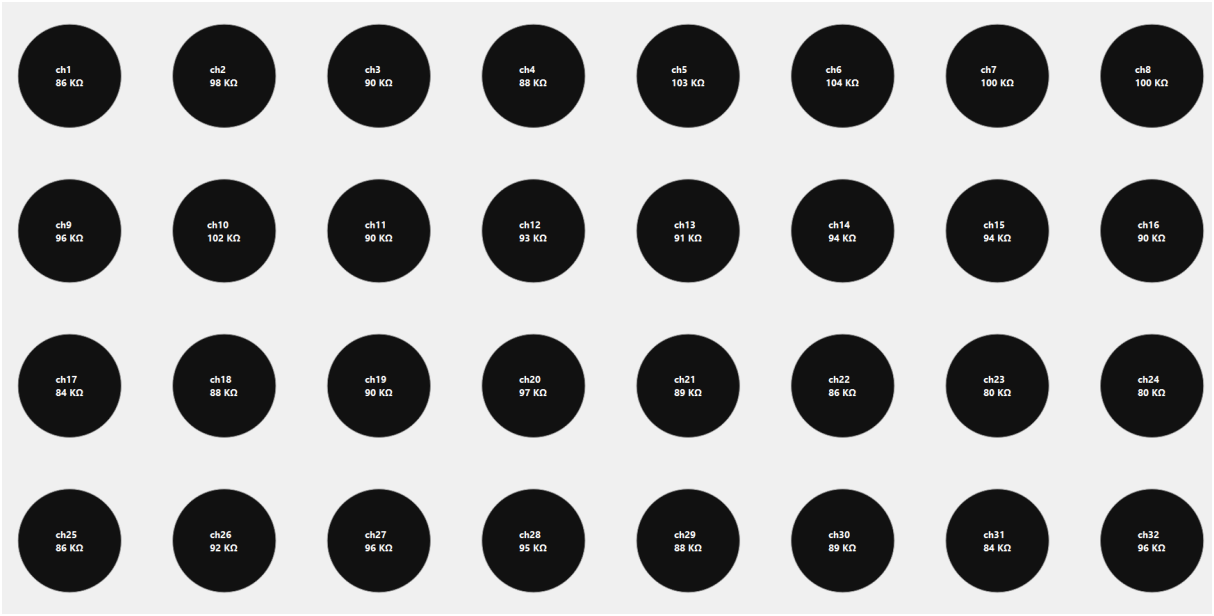
The problems and limitations that affected the data collection are mentioned in this section.

For the Mentalab Explore 32+ system, the setup time before the experiment could begin was 30 minutes. This was due to the tedious steps of prepping the skin, applying the gel, and improving the impedance at all 32 channels and the reference. The 30-minute setup time was not a problem in itself, but it contributed to the participants getting tired of the experiment a bit faster, as the protocol lasted for about another 75 minutes after setup. At the beginning of the experiments, the participants were quite excited and interested, but after approximately 30 minutes most of the participants started to feel bored and found it harder to concentrate on the experiment. The stimuli were only 40-second movie clips out of context without any sounds, so the participants found it a bit difficult to feel really immersed in the stimuli for all other categories than horror or erotic clips. The lights were turned off during the experiments to reduce the glare on the computer screen. This might have also contributed to the participants starting to feel kind of drowsy and more bored.

We had to set the maximum length of the set-up time to 30 minutes in order to keep the participants engaged with the experiment. This, together with the fact that the natural impedance of the skin is participant-dependent, resulted in quite a high impedance in the device with big variation across participants. In fig. 3.13 the impedance on the 32 channels right before the experiment started are shown for participant no. 4 and participant no. 6, i.e. the participants with the lowest and highest impedance. The lowest impedance obtained at any electrode was $22k\Omega$. The average impedance for participant no.4 over all channels was $29{,}5k\Omega$ with a standard deviation of $3{,}5k\Omega$. For participant no. 6 it was $91{,}8 \pm 6,2k\Omega$.



(a) Impedance on all channels for participant no. 4. This participant had the lowest impedance after the set-up time of 30 min.



(b) Impedance on all channels for participant no. 6. This participant had the highest impedance after the set-up time of 30 min.

Figure 3.13: Impedance at all channels before experiments started with Mentalab Explore+ 32 device. The pictures are taken from the Mentalab desktop interface. The circle is black for channels with impedance$\geq 50k\Omega$, red for impedance$\geq 30k\Omega$, and orange for impedance$\geq 20k\Omega$.

The Mentalab Explore 32+ device stopped recording data halfway through the experiments for all the participants. Unfortunately, this error was not detected until seven participants had gone through the experiment. Because of this, the data from these subjects only contain recording for 33-45 trials. The device stopped working altogether during the recordings of the 7th participant. It was concluded that we unfortunately had received a malfunctioning device. Some of the channels also introduced a lot of noise at all frequencies during some trials. Because of these issues, another full dataset of 20 participants was collected with the Unicorn Hybrid Black device.

The set-up time with the Unicorn Hybrid Black device was very short, and only took about 5 minutes. This was beneficial because the participants did not get bored with the experiment as quickly. A drawback with this system was the fact that the cap and the electrodes were not very comfortable. Even though the electrodes used with this device are far from the most uncomfortable electrodes on the market, the participants would complain especially about the electrodes pressing on the forehead towards the end of the experiments. This unpleasantness

might affect the emotions the participants are experiencing, which is somewhat of a drawback. Still, this should not be a problem as long as the feeling is not too distracting and the participants rate the feelings they are actually experiencing.

Another thing to keep in mind is the fact that people are somewhat excited to be part of the experiment, which affects their emotional state. For example, participant no. 14 with the Unicorn Hybrid Black device rated their arousal to be between 6 and 9, i.e. high, for all videos. For this one participant, this was simply dealt with by changing the definition of high arousal to values above 6 instead of 5, like for the other participants. Understandably, the data from each participant is not balanced even though the stimuli were chosen to create close-to-balanced datasets.

For participant no. 3 with the Unicorn Hybrid Black system, one of the EEG amplifiers stopped recording halfway through the experiment. The data collected from this participant was not used in this project.

The Mentalab Explore+ 32 device has wet electrodes, while dry electrodes were used with the Unicorn Hybrid Black device. Wet electrodes as a much higher signal-to-noise ratio compared to dry electrodes. Additionally, the impedance in the Unicorn Hybrid Black device could not be controlled.

3.6 Preprocessing data

3.6.1 Mentalab Explore+ 32

The power spectral density was investigated for each trial. This way the bad channels could be located and removed from the dataset for this participant. Because of this, it varies how many channels are present for each participant in the preprocessed data.

The EEG data and events were extracted. A few things were explored as possibilities in preprocessing:

1. No preprocessing
2. Downsampling (128 Hz)
3. ICA
4. Notch filter (50 and 100 Hz)
5. Downsampling (128 Hz) and notch filter (50 and 100 Hz)

MNE-Python's ICA function is used. The FastICA method is used in this project as it is more stable for not completely independent signals, which is often the case for EEG signals. A process diagram of how it works is shown in fig. 3.14. Before ICA was applied, the breaks during the experiments were annotated.

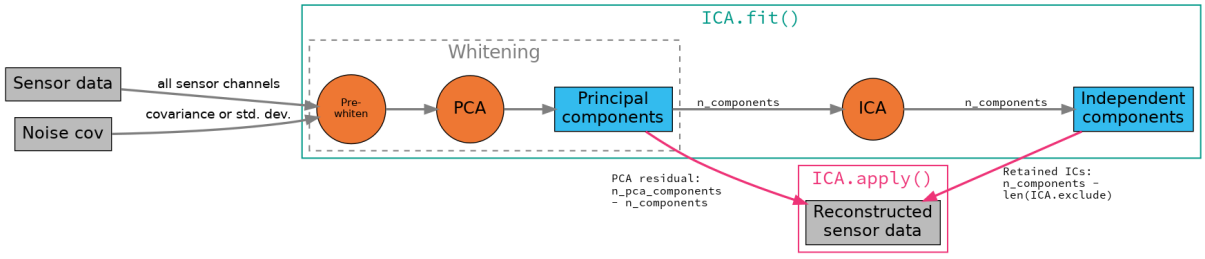


Figure 3.14: The process of the mne.preprocessing.ICA function implemented in MNE-Python.

When ICA was applied to the signal, the independent components signals and topographic maps were examined visually. The component(s) corresponding to eye blinks were tried located and removed. fig. 3.15 shows the topographic map of component 5 for participant no. 7. The activity is located at the forehead, around the eyes. The signal of component 5 is shown in fig. 3.16b. By inspecting the topographic plot and the signal, the component is determined to be eye blinks. In fig. 3.16a and fig. 3.16c the signals at the channels O2, Fp1, F7, FT9, F3, and FC5 are shown before and after removing component 5, respectively. The blinking could be seen most clearly in Fp1, as this channel is located at the forehead.

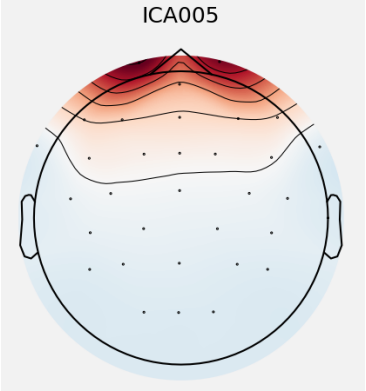
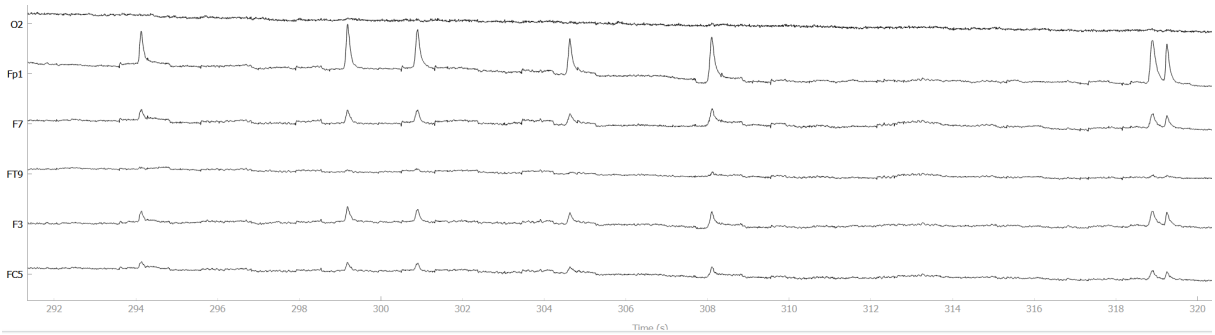
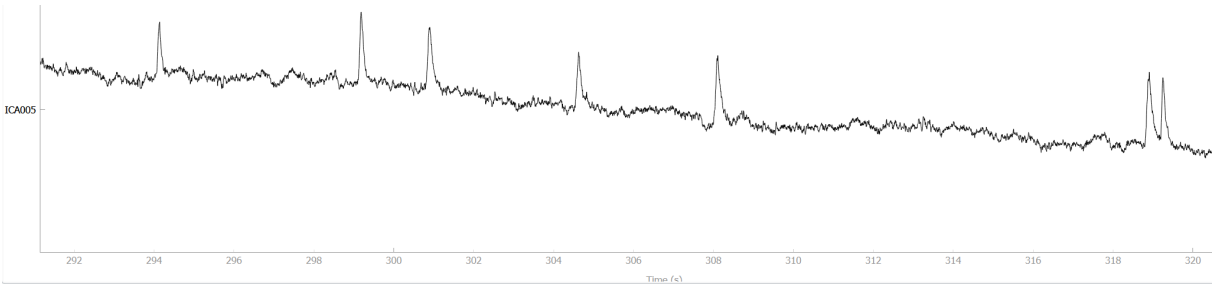


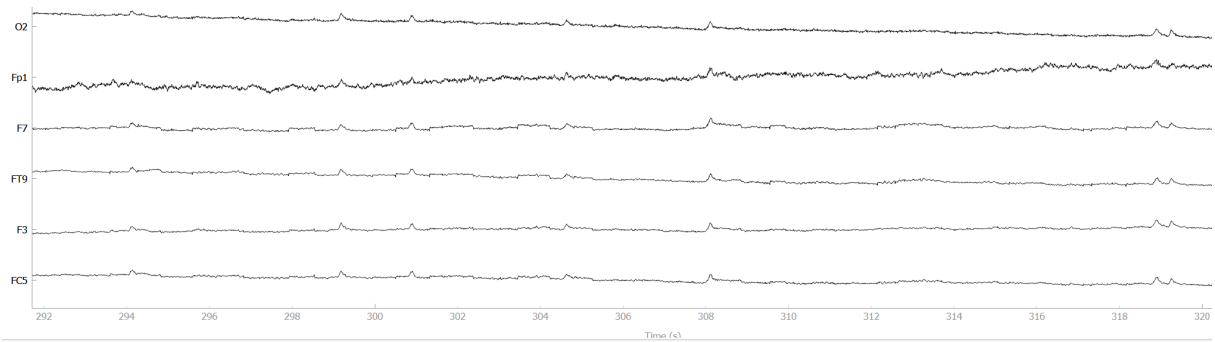
Figure 3.15: Topographic map of component 5 from ICA on participant no. 7. The component is located at the forehead of the participant, the electrodes closest to the eyes. This suggests that the component describes eye artifacts.



(a) Original EEG signals. The amplitude of the Fp1 signals is around $600\mu V$.



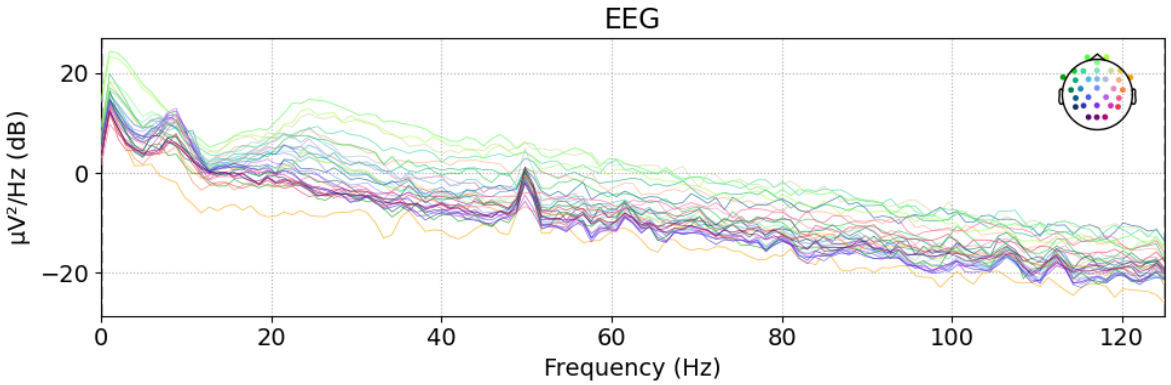
(b) Component 5 found with ICA. See fig. 3.15 for a topographic map of the source.



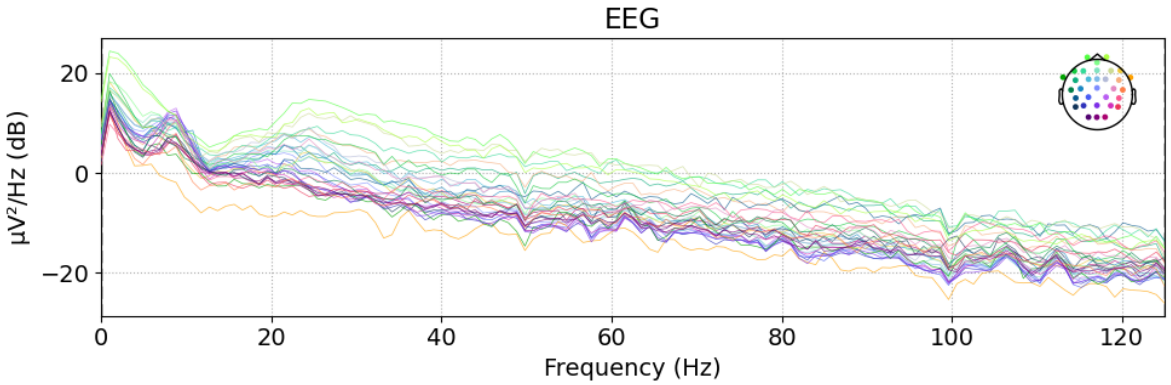
(c) EEG signal after removing component 5. The amplitude of the Fp1 signal is around $230\mu V$.

Figure 3.16: Removing blinking artifacts from EEG recordings of participant no. 7 with ICA. The electrode Fp1 is located on the forehead, i.e. close to the eyes. The blinking is most visible in this channel. The scale of the axes in plots (a) and (c) are the same.

The effect of applying the notch filter of 50 and 100 Hz can be seen in fig. 3.17. The big peak at 50 Hz and the small peak at 100 Hz in the PSD plot in fig. 3.17a are due to the frequency of the power line.



(a) Original power spectral density plot.



(b) Power spectral density after applying a notch filter at 50 and 100 Hz.

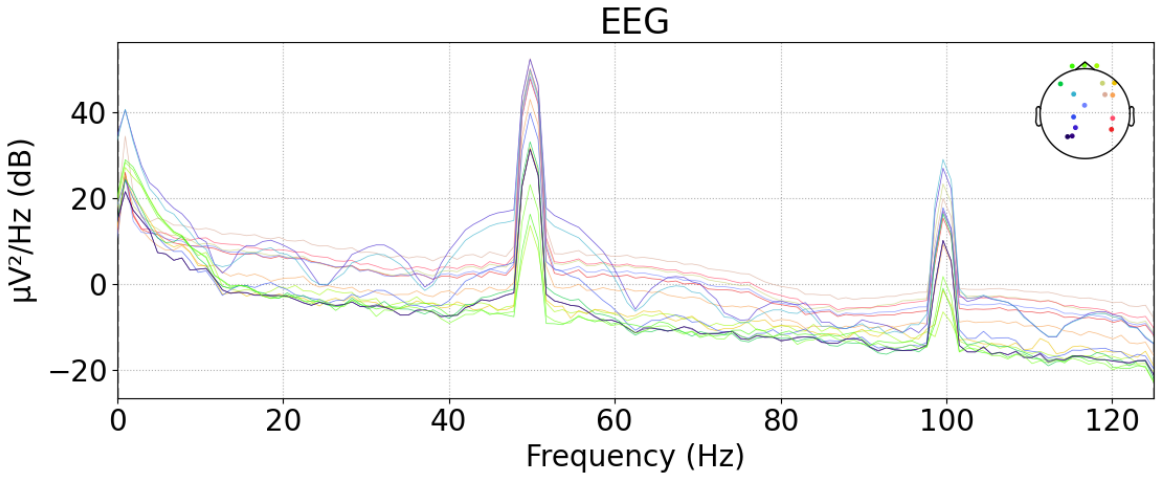
Figure 3.17: The PSD of trial 3 for participant number 6 using the Mentalab Explore+ 32 dataset.

After the preprocessing, 5-sec baseline signals and the 40-sec trials were extracted using the event markers. One second after the ”cross-fixation start” marker was located, the 5-second baseline was extracted. 40 seconds of recording was extracted starting from the ”video start” marker. The intervals are visualized in fig. 3.18. As mentioned in section 3.5.2, the equipment was somewhat malfunctioning, so some trials from the experiments were lost for each participant. Therefore, each participant has a different number of EEG-recorded trials.

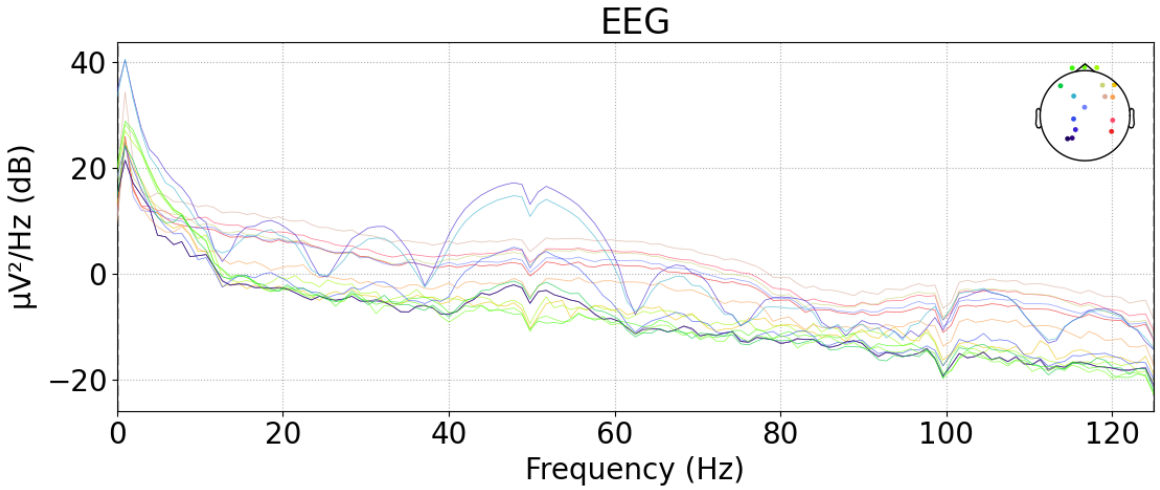
Figure 3.18: The 5-second baseline and the 40-second trial data was extracted using the marker events. The baseline data starts one second after the participants had looked at the cross-fixation for one second. The trial data last the entire length of the stimuli.

3.6.2 Unicorn Hybrid Black

The EEG data recorded on the Unicorn Hybrid Black also includes some signals from the power line. In Norway, the power line frequency is 50 Hz and can clearly be seen as peaks at 50 Hz and 100 Hz in the power spectral



(a) Original power spectral density plot.



(b) Power spectral density after applying a notch filter at 50 and 100 Hz.

Figure 3.19: The PSD of trial 3 for participant number 10 using the Unicorn Hybrid Black dataset. (a) is the original PSD and (b) is the PSD after applying a notch filter.

density (PSD) plot in fig. 3.19a. The noise contributed by the power line can be removed by applying a notch filter at these frequencies, resulting in the PSD plot in fig. 3.19.

The same preprocessing steps as above were also tried for the Unicorn Hybrid Black. With ICA, it was difficult to separate only the eye blinks from all other sources, so the model was not tried on ICA-processed data.

Lastly, the 5-sec baseline signals and 40-sec trials were extracted for each of the preprocessing options.

The frequency of the power line is very apparent in the PSD plots from the Unicorn Hybrid Black data. The big peaks at 50 Hz and 100 Hz can be seen in fig. 3.19b. The effect applying the notch filter of 50 and 100 Hz has on the PSD can be seen in **??**.

3.7 Feature extraction

The feature extraction was done in the same manner as in the project thesis. Below is a summary, see [18] for a closer explanation. The baseline and trial data of size *trial x channels x data points* was segmented into one-second epochs and decomposed into the five frequency sub-bands delta (1-4Hz), theta (4-8Hz), alpha (8-14Hz), beta (14-31Hz), and gamma (31-45Hz). Each sub-band contains data of size *seconds x channels x sample frequency*. Next,

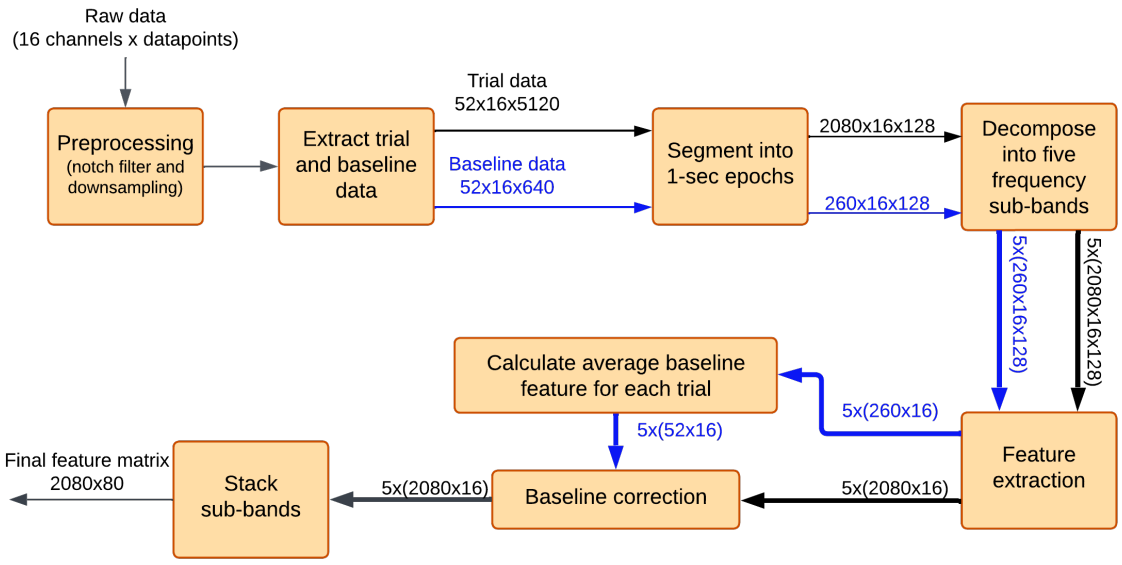


Figure 3.20: The diagram explains the process from raw data to the final feature matrix for one feature. In this instance, the preprocessing is notch filtering and downsampling.

the features HM, HC, HMS, HCS, FBE, and DE were extracted for each epoch. The mean value of the features extracted from the baseline epochs (now called the baseline feature) was calculated for each feature. Baseline correction was done by subtracting the baseline features from all the features (under stimuli) in the corresponding trial. The process is explained visually in fig. 3.20.

The original SVC model proposed in the specialization project utilized only five features: HC, HM, DE, FBE, and HMS. HCS was also explored in this combination of features but for the DEAP dataset, the model performance decreased by 0,1 percentage points for valence and 0,2 percentage points for arousal. The difference in participant-wise accuracy is however not statistically significant, i.e. the p-value is above 0.05. Because of this, the effect of including the HCS feature together with the other five in the SVC model was explored on the new datasets.

3.8 Test generalization qualities of model

Finally, the proposed models could be tested on brand-new, unseen data. Firstly, 20% of the data was extracted in a random manner and kept as a test set. 5-fold cross-validation was used to test the SVC and LR models proposed in the project [18] on the remaining data. The results from the cross-validation were used in an effort to improve the performance of the model by applying more/different preprocessing.

In addition to testing the classification qualities of the model on high/low classification for arousal and valence, the model's prediction qualities on the original labels reported by the participants were also tested. The original labels reported by the participants are discrete values from 1 to 9 for both arousal and valence. The predictions of the original labels could then be mapped to high/low values for arousal and valence. So three ways of classifying the data were tested: (1) predict high/low labels, (2) predict original (discrete 1-9) labels, and (3) predict original (discrete 1-9) labels and convert to high/low afterward.

Finally, the test set was used to validate the performance of the best combinations of preprocessing and model choice on unseen data.

4

Results

4.1 Important results from the specialization project

In the specialization project [18], a model was created for the DEAP dataset. Table 4.1 displays the model performance of 4 cases:

1. Classifying from epochs extracted from original data
2. Classifying from epochs that are decomposed into frequency sub-bands theta, alpha, beta, gamma
3. Classifying from epochs that are baseline corrected
4. Classifying from epochs that are decomposed into frequency sub-bands and are baseline corrected

,for the two best single features, HC and HM. It was clear that implementing frequency sub-bands and baseline correction was very important for the model performance.

Combining multiple features into the model further improved performance. For the SVC model, the use of five features (HC, HM, DE, FBE, HMS) performed best. The same was checked for the LR model, where the use of HCS in addition to the other five features was the best option. In table 4.2 the average accuracies obtained from 5-fold cross-validation on the SVC model with five and six features and the LR model with six features are listed.

Feature	Feature options	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Hjorth complexity	Original data	63.63% ± 6.26%	67.25% ± 7.99%
	Sub-bands	60.85% ± 6.10%	65.60% ± 8.88%
	Baseline removal	77.91% ± 5.83%	79.50% ± 6.26%
	Sub-bands + baseline removal	**88.63%±3.92%**	**89.28%±4.36%**
Hjorth mobility	Original data	65.60% ± 5.68%	68.15% ± 7.78%
	Sub-bands	64.01% ± 5.53%	66.71% ± 8.36%
	Baseline removal	73.85% ± 6.67%	76.68% ± 6.08%
	Sub-bands + baseline removal	**88.38%±4.70%**	**88.54%±5.88%**

Table 4.1: Average accuracies of the proposed SVC model with randomly divided training set (66%) and test set (33%). HC and HM were the two best-performing single features. The table shows the effect implementing sub-bands and baseline removal had on the model performance.

Classifier	Features	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
SVC	HC, HM, DE, FBE, HMS	96.50% ± 2.10%	96.71% ± 2.06%
SVC	HC, HM, DE, FBE, HMS, HCS	96.43% ± 2.08%	96.58% ± 2.12%
LR	HC, HM, DE, FBE, HMS, HCS	96.26% ± 1.41%	96.61% ± 1.54%

Table 4.2: The performance of the proposed models and the six-feature SVC from the specialization project. The reported average accuracy is the average accuracy of all participants after 5-fold cross-validation. The kernel in the SVC model is 'rbf'. The logistic regression is L_2-regularized and has C=1.

4.2 Feature selection

In the original SVC model proposed in the specialization project, only five features were utilized. The average accuracies obtained with 5-fold cross-validation of the 32-channel SVC model on the new datasets with the original five and all six features are listed in table 4.3. Utilizing all six features (HM, HC, HMS, HCS, DE, and FBE) seems to increase the accuracy somewhat. The difference is not big, but since the extra information shows signs that it might improve the model somewhat, the HCS feature was included in the rest of the project. However, the improvement is not statistically significant.

Dataset	Features	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Mentalab Explore+ 32	HC, HM, DE, FBE, HMS	92.89% ± 3.97%	92.79% ± 3.98%
	HC, HM, DE, FBE, HMS, HCS	93.10% ± 3.99%	93.10% ± 3.97%
Unicorn Hybrid Black	HC, HM, DE, FBE, HMS	91.26% ± 3.16%	90.61% ± 4.31%
	HC, HM, DE, FBE, HMS, HCS	91.53% ± 2.97%	91.00% ± 4.20%

Table 4.3: Average accuracies obtained from the SVC model when using five features (like the proposed model in the specialization project) or using all six features on the newly collected datasets.

4.3 Channel Selection

4.3.1 Non-dominated Sorting Algorithm on DEAP dataset

In this section, the results derived from applying the NSGA-II in an attempt to reduce the number of electrodes needed for emotion recognition with the SVC model, are presented. The channel selection is done individually for arousal and valence and participants-wise. Therefore, the number of channels might be varying across participants, so the average number of channels across participants is reported. The column "average total number of channels" in table 4.4-table 4.6 is the total number of channels needed in one device to fulfill the optimal channels selection found for both arousal and valence. The total number of channels will be higher than the number of channels chosen for arousal or valence classification individually, as the channels selection for arousal and valence do not have all channels in common.

The prediction accuracies on the validation set using the unconstrained NSGA-II are shown in table 4.4. Setting both the population size and the number of generations equal to 50 provides the best solution for channel selection. The mean accuracy achieved on the validation set using the unconstrained NSGA-II was 95.35% ± 2.15% and 95.71% ± 2.48% for valence and arousal, respectively. On average, around 11.0-11.5 channels are needed for the classification of arousal and for valence. Many of these channels are however unique for the classification of either valence or arousal. Therefore, the average number of electrodes a device would need to have to achieve this prediction accuracy for both valence and arousal would on average be around 17.5.

The accuracy decreases when the population size is smaller than the number of generations. However, the solution with the lowest number of channels is found with the NSGA-II with population size=20 and a number

of generations=120. In this instance, we are able to obtain a mean accuracy of 95.02% ± 2.69% for high/low valence classification and 95.04% ± 2.60% for high/low valence classification using around 16 electrodes for each participant.

Specifications	Average number of channels ± STD			Valence Average accuracy ± STD	Arousal Average accuracy ± STD
	Valence	Arousal	Total		
Population:50 Generations:50	11.53±2.33	11.09±3.26	17.53±3.71	**95.35% ± 2.15%**	**95.71%±2.48%**
Population:35 Generations:70	10.03±1.96	10.16±2.76	16.16±3.41	95.08%±2.3%	95.51% ± 2.37%
Population:20 Generations:120	9.81±2.73	10.47±2.49	16.13±3.06	95.02% ± 2.69%	95.04% ± 2.60%
Population:70 Generations:35	9.28±2.96	9.88±2.33	15.06±3.25	95.31% ± 2.35%	95.42% ± 2.34%

Table 4.4: Average classification accuracies on the validation set using unconstrained NSGA-II on the DEAP dataset with the SVC model. The average number of channels needed for high/low arousal and valence classification and both at the same time is reported.

For the constrained problem the constraint is on the number of channels used for one type of classification, i.e. valence or arousal. This means that the total number of electrodes used to classify both valence and arousal from the same device will usually be bigger than that constraint. The average accuracy was about the same for all hyperparameter settings, with only one percentage point differentiating between the best and worst average accuracy for valence and arousal classification. This can be seen in table 4.5. In terms of accuracy, the best hyperparameters were again a population size and number of generations equal to 50, and now constraining the problem to use 29 electrodes. The average of the total channels used for this solution would however be 31.25 electrodes, which is basically using all 32 electrodes as original.

One solution only uses on average 22.5 channels in total on average. This solution is found with the population size set to 35 and running this for 70 generations. The average accuracy for valence and arousal respectively is 95.02% ± 2.39% and 96.12% ± 2.08%.

Specifications	Average total number of channels ± STD	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Population:50 Generations:50 Channel constraint:16	23.25±1.66	96.02% ± 2.35%	95.92% ± 2.25%
Population:35 Generations:70 Channel constraint:16	22.53±1.50	95.02% ± 2.39%	96.12% ± 2.08%
Population:50 Generations:50 Channel constraint:25	30.25±1.03	96.60%±2.12%	96.41%±2.43%
Population:50 Generations:50 Channel constraint:29	31.25±0.71	**96.76%± 2.19%**	**96.51%± 2.51%**

Table 4.5: Average classification accuracies on the validation set using constrained NSGA-II on the DEAP dataset with the SVC model. The average number of channels is the average number of channels used in total for both arousal and valence classification.

The solutions with the highest accuracy and with the lowest number of utilized channels were also tested with the 5-fold cross-validation so that they can more easily be compared to the values in table 4.2. The results from the 5-fold cross-validation can be found in table 4.6.

The average accuracies of the solutions are again quite similar, with the difference between the lowest and the highest accuracy being less than 1.5 percentage points for both arousal and valence. The solution with the highest accuracy is the solution that utilizes the most electrodes. The more channels the solution utilizes. The highest achieved accuracy is 96.52% and 96.63% for valence and arousal classification. In this solution, 31.3 channels are used on average. The accuracy achieved with 16.2 channels on average is 95.26% for valence and 95.60% for

arousal. As seen in table 4.2, accuracies of 96.50% for the valence and 96.71% for arousal were achieved using all 32 channels, with 5-fold cross-validation. There are slight variations in the achieved average accuracies in the different solutions. However, the difference in the participant-wise achieved accuracy is not statistically significant between any of the proposed solutions listed in table 4.2, i.e. the p-value is above 0.05.

The solution with an average of 16.2 channels is statistically significant with regard to the participant-wise accuracy compared to the solution without any channel reduction. None of the other solutions listed in table 4.6 are statistically significant with regards to the participant-wise accuracy compared to the 32-channel model.

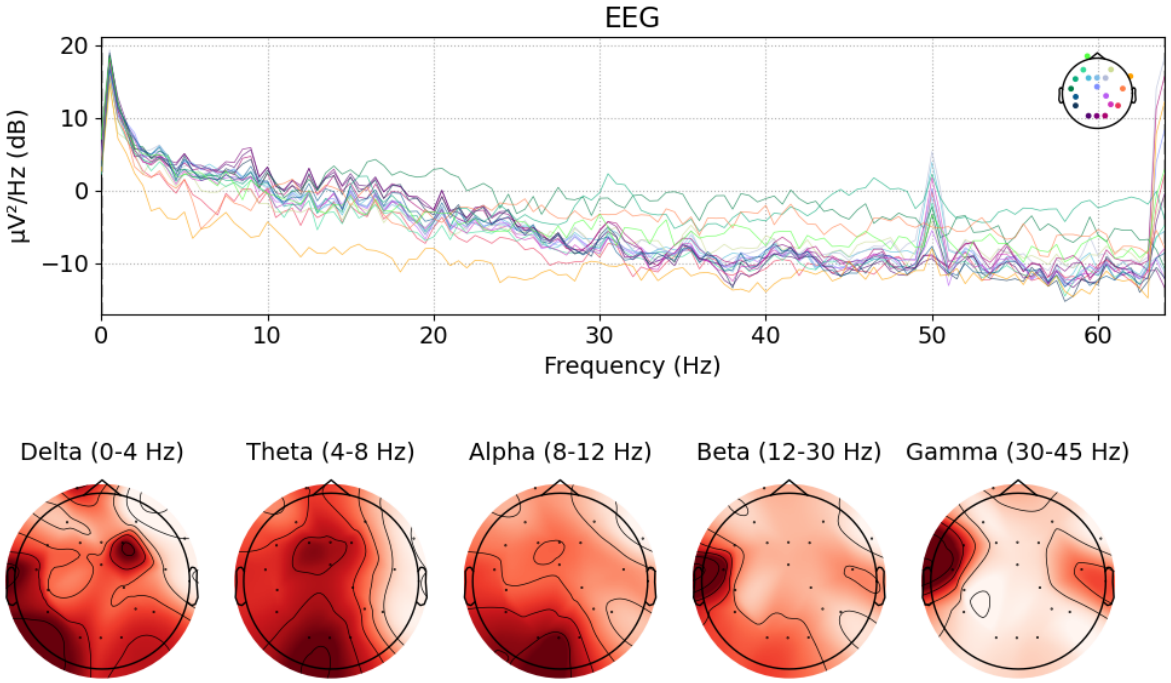
Specifications	Average total number of channels ± STD	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Population:70 Generations:35 Unconstrained	15.06±3.25	95.31%±2.35%	95.42%± 2.34%
Population:50 Generations:50 Unconstrained	17.53±3.71	95.45%± 2.04%	95.95%± 1.94%
Population:50 Generations:50 Channel constraint:25	30.25±1.03	96.42%±1.97%	96.57%±2.06%
Population:50 Generations:50 Channel constraint:29	31.25±0.71	**96.52%± 1.96%**	**96.63%± 2.13%**

Table 4.6: Average accuracies obtained when running 5-fold cross-validation over the entire dataset with the best-performing channel selections. This was calculated for the sake of comparing the results to those obtained in the specialization project.

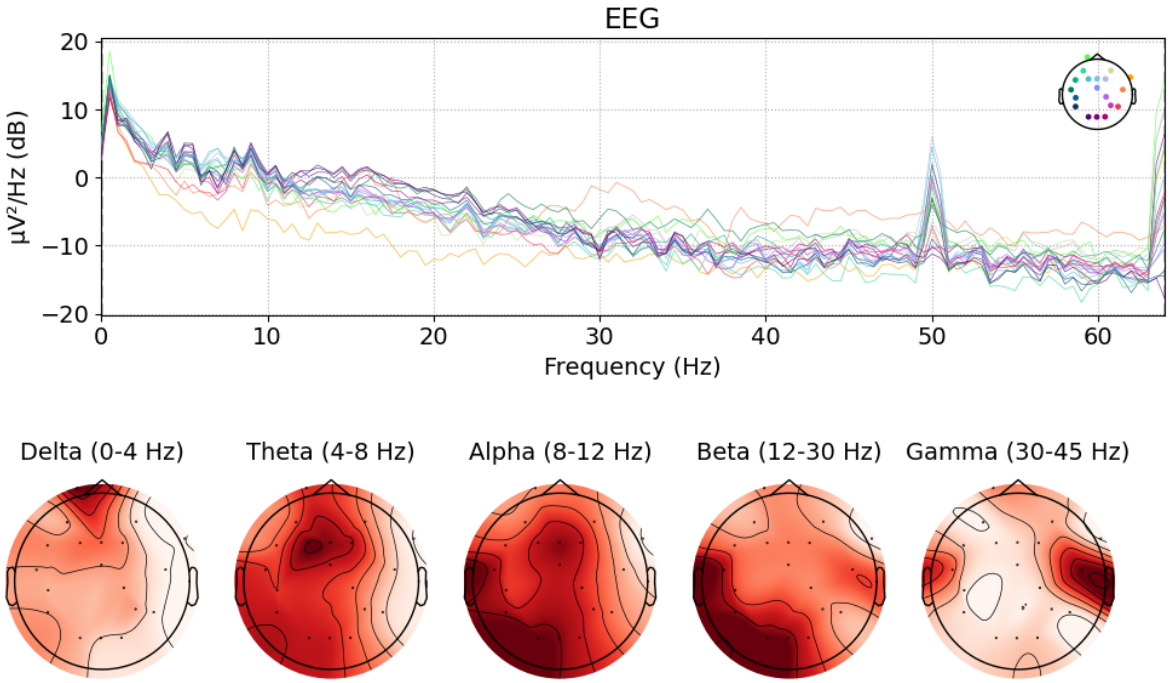
4.4 Power spectral density plots

Figure 4.1 shows the PSD of two trials (low valance/low arousal and low valance/high arousal) for participant no. 4. This participant had 19 good channels across the head. The topological plots of the PSD in the five frequency bands are also shown. ICA has been applied in order to remove noise from eye blinks, see more about this further down. Looking at the graphs, one can see differences in which channels stand out as the most active ones, but they are difficult to decipher. This is easier in the topographic plots. The plots do however not have the same scale, so the comparison is only about where the activity is located. One can clearly see differences in brain activity in the low and high arousal situations. The activity in the delta band in the high arousal situation is likely not from brain activity, and will be ignored for this purpose. The activity changes most in the alpha, beta, and gamma bands. For alpha and beta, the activity mostly spreads to a bigger part of the left side and the occipital part of the brain. In the gamma band, the activity is more or less mirrored in the high-arousal trial compared to the low-arousal trial.

The effect ICA has as a preprocessing method is shown in fig. 4.2. The plot shows the topological map of the PSD in the five frequency bands before and after using ICA to remove eye blinks. The data was downsampled in both plots. The eye blinks introduce a lot of noise to the data, this is apparent when looking at the topological plots. The eye blinks are so overpowering in the delta, theta, and alpha frequency bands that any other activity in the brain is not visible. After removing the eye blinks from the EEG data through ICA, the interesting activity is very clear.



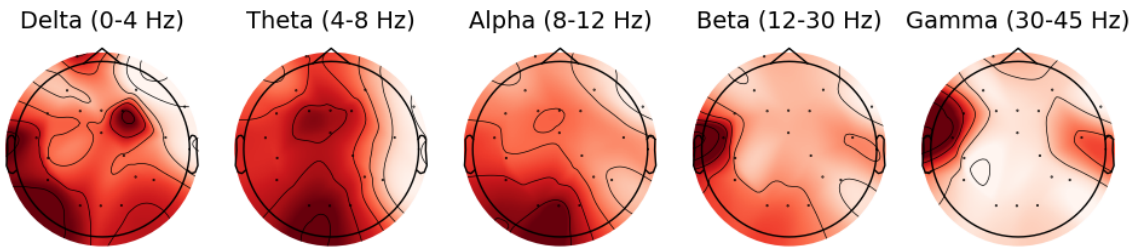
(a) Trial 21: valence=3 and arousal=4 in this trial, i.e. low valence and low arousal.



(b) Trial 23: valence=3 and arousal=8 in this trial, i.e. low valence and high arousal.

Figure 4.1: PSD plot and corresponding topological map for participant no. 4 with the Mentalab Explore+ 32 device. The data has been downsampled and ICA has been applied to remove eye blinks.

(a) Before ICA is applied there is a lot of activity at the forehead in the lower frequency bands.



(b) The overpowering activity at the forehead had been removed through ICA. Now the other activity in the brain is visible.

Figure 4.2: PSD topological map of trial 21 for participant no. 4 with the Mentalab Explore+ 32 device before and after removing eye blinks with ICA. The data had been downsampled in both instances. Please note that the topological plots are not on the same scale.

4.5 Preprocessing new data

The model performance on the new datasets will be presented in this section.

4.5.1 Mentalab Explore+ 32

Removing bad channels

Any channels that were bad during one trial were removed from the entire dataset for that participant. As seen in table 4.7, this resulted in a varying number of channels for each participant ranging from 8 to 19 usable channels. The good channels are plotted for each participant in fig. 4.3. In table 4.7, participant-wise average accuracy achieved from the 5-fold cross-validation is presented. The accuracy does decrease somewhat for most of the participants, where the worst case is with participant no. 7. This participant had only 8 good channels, but the model still manages to predict 85.03% of the instances correct for high/low arousal and 85.28% for valence. Using all channels, the model achieved accuracies of 93.92% and 93.09% for arousal and valence respectively for the same participant. The channel reduction due to bad channels resulted in a decrease in accuracy of 8-9 percentage points.

For participant 6 the channel reduction was actually beneficial as the prediction accuracy for arousal stayed the same, while the accuracy for valence predictions increased by 0.16%. This increase might just be due to change, but the channel reduction does not have seems to have harmed the model.

In table 4.8 the average accuracies from 5-fold cross-validation on the SVC and LR models before and after channel reduction to remove bad channels are presented. For both models, the average accuracies decrease after removing the bad channels, for both arousal and valence predictions. The biggest decrease is however just 2.6 percentage points, which is for the valence classification using the LR model.

	32 channels		Bad channels removed		
Participant	Valence accuracy	Arousal accuracy	Valence accuracy	Arousal accuracy	Channels remaining
1	97,17%	96,09%	93,65%	96,29%	14
2	95,50%	94,76%	94,76%	92,46%	12
3	95,80%	95,23%	95,07%	94,33%	15
4	96,72%	96,09%	95,31%	93,52%	19
5	96,38%	93,89%	92,26%	90,34%	14
6	94,31%	94,47%	94,47%	94,47%	18
7	93,09%	93,92%	85,28%	85,03%	8

Table 4.7: Participant-wise effect of removing the bad channels in the Mentalab Explore+ 32 dataset. The accuracies are obtained with the LR using 5-fold cross-validation.

	32 channels Average accuracy±STD		Bad channels removed Average accuracy±STD	
Classifier	Valence	Arousal	Valence	Arousal
SVC	93.10%±3.99%	93.10%±3.97%	91.96%±6.49%	91.70%±6.62%
LR	95.57%±1.33%	94.92%±8.57%	92.97%±3.28%	92.35%±3.44%

Table 4.8: Average accuracies before and after removing bad channels for the SVC and the LR models.

Preprocessing

The effects preprocessing has on the accuracy of the model in the Mentalab Explore + 32 dataset are reported in table 4.9. The table shows the average accuracy of the 5-fold cross-validation on the training data with the LR and SVC models using all six features HC, HM, DE, FBE, HMS, and HCS. For the three cases:

(1) Predicting high/low labels for arousal and valence

(2) Predicting the original (discrete 1-9) labels reported by participants for arousal and valence

(3) Converting the predicted 1-9 labels into high/low labels for arousal and valence

, table 4.9 covers the results of case (1). table 4.10 shows the results for cases (2) and table 4.11 for case (3).

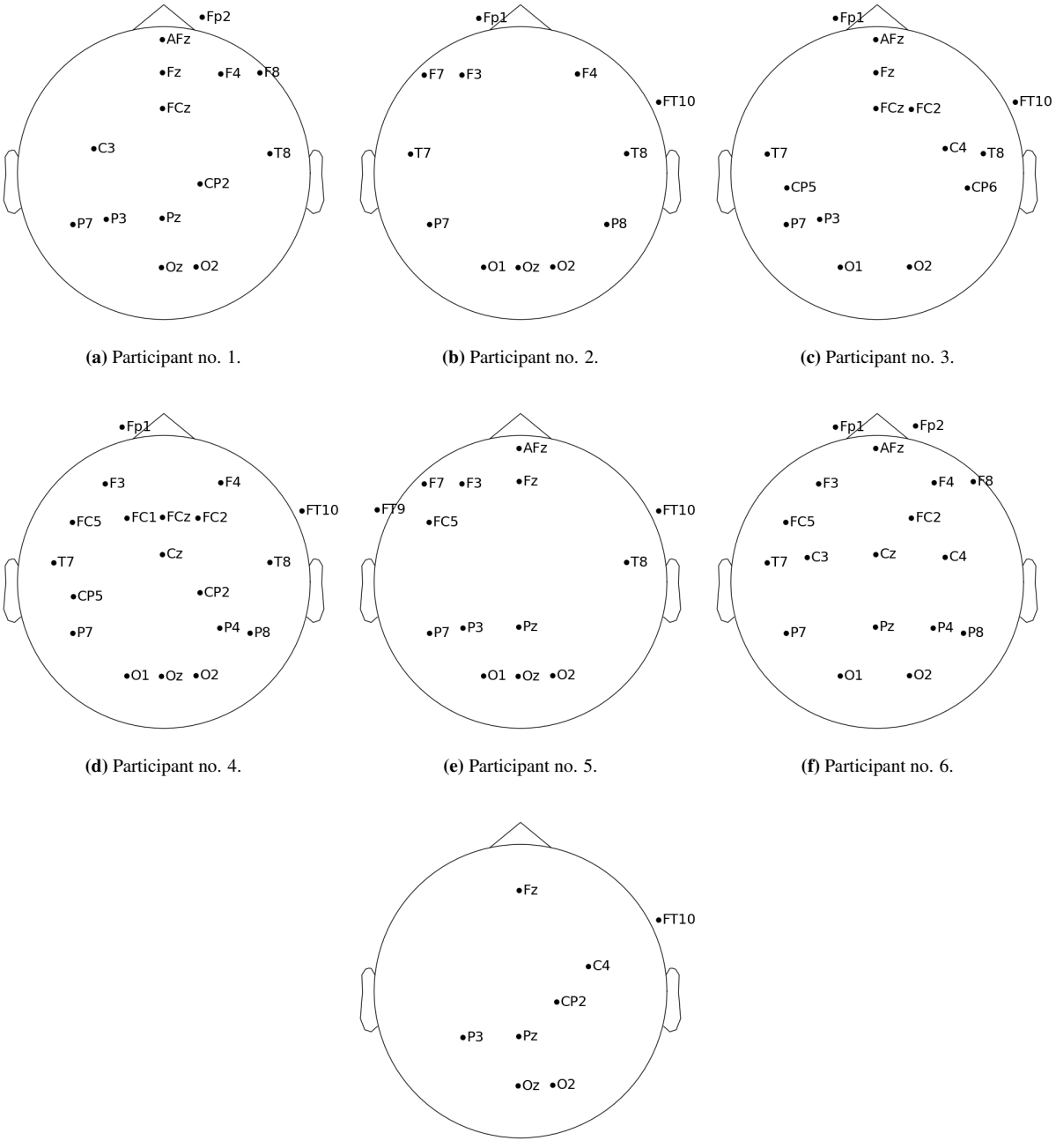
For case (1), the SVC model achieves an accuracy of 93.10% $\pm$ 3.99% for valence and 93.10% $\pm$ 3.97% for arousal without performing any preprocessing. The LR model achieves accuracies of 95.57% $\pm$ 1.33% and 94.92% $\pm$ 0.86% with no preprocessing for valence and arousal, respectively. The LR model performs better than the SVC model for all types of preprocessing.

The best-performing preprocessing method is downsampling to 128Hz. With this method, the accuracy of the LR model is increased to 93.33% $\pm$ 4.13% and 92.98% $\pm$ 3.47% for valence and arousal, respectively. The introduction of the notch filter, in addition to downsampling, reduced accuracy by less than 0.2 percentage points for both valence and arousal classification. The SVC model with downsampled and notch-filtered data obtains an accuracy of 93.08% $\pm$ 5.12% for arousal, but the standard deviation is much higher than for the LR model with downsampled data.

Removing eye blinks with ICA does not increase the accuracy of the LR model. In fact, it decreases it somewhat. For the SVC model, the accuracy decreases for valence classification but increases by one percentage point for arousal classification when using ICA.

The notch filter reduces the accuracy of the valence predictions by less than one percentage point for both models, while the arousal classification performance increases by around 0,5 percentage points.

For LR the standard deviation is about the same for all methods of preprocessing. For valence, it ranged from 3.28% with no preprocessing to 4.73% when notch filter and downsampling is applied. The standard deviation for arousal is kept within 3.40%-3.47% for all the preprocessing methods for the LR model. The standard deviation varies a bit more with the SVC model, but it follows the same trend for which preprocessing methods have the most and least standard deviation.



(a) Participant no. 1. **(b)** Participant no. 2. **(c)** Participant no. 3.

(d) Participant no. 4. **(e)** Participant no. 5. **(f)** Participant no. 6.

(g) Participant no. 7.

Figure 4.3: The good channels in the Mentalab Explore+ 32 dataset for each participant is plotted.

Predicting high/low arousal and valence.				
Features	Classifier	Preprocessing	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
HC, HM, DE, FBE, HMS, HCS	**SVC**	No preprocessing	91.96% ± 6.62%	91.70% ± 6.49%
		Downsampling	91.49% ± 8.50%	92.84% ± 5.83%
		ICA	91.22% ± 8.33%	92.03% ± 5.78%
		Notch filter	91.27% ± 8.60%	92.41% ± 5.00%
		Notch filter, downsampling	91.41% ± 8.83%	**93.08% ± 5.12%**
	LR	No preprocessing	92.97% ± 3.28%	92.35% ± 3.44%
		Downsampling	**93.33% ± 4.13%**	92.98% ± 3.47%
		ICA	92.57% ± 4.09%	92.30% ± 3.42%
		Notch filter	92.47% ± 4.46%	92.52% ± 3.40%
		Notch filter, downsampling	93.16% ± 4.73%	92.81% ± 3.40%

Table 4.9: The results for high/low classification from 5-fold cross-validation over the training set on the Mentalab Explore+ 32 dataset. The best mean accuracies obtained are marked with bold lettering.

As seen in table 4.10, the SVC model achieves an accuracy of 85.08% ± 13.71% and 84.69% ± 14.67% for valence and arousal respectively when predicting the original labels ranging from 1-9. The SVC model is again outperformed by the LR model in terms of average accuracy and standard deviation.

For the LR model, the average accuracy is 92.18% ± 5.41% for valence classification and 92.08% ± 6.08% for arousal. As opposed to the case with the high/low labels, removing blinks through ICA increases the prediction accuracy for the prediction of the original labels. For the other preprocessing methods, the methods increase and decrease the accuracy in the same manner as for the high/low prediction. The best accuracy was achieved with downsampling, resulting in 93.47% ± 4.73% for valence and 93.51% ± 5.71% for arousal. This accuracy is about the same as the accuracy obtained from high/low classification with downsampling, even though there are 9 possible levels. Applying the notch filter in addition to the downsampling reduces the accuracy to 92.76% ± 6.62% and 92.81 ± 7.18.

Predicting the original (discrete 1-9) labels for arousal and valence.				
Features	Classifier	Preprocessing	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	85.08% ± 13.71%	84.69% ± 14.67%
		Downsampling	85.42% ± 15.68%	85.51% ± 14.91%
		ICA	84.28% ± 15.51%	85.01% ± 14.96%
		Notch filter	84.37% ± 16.01%	84.59% ± 14.97%
		Notch filter, downsampling	85.27% ± 16.01%	85.18% ± 15.33%
	LR	No preprocessing	92.18% ± 5.41%	92.08% ± 6.08%
		Downsampling	**93.47% ± 4.73%**	**93.51% ± 5.71%**
		ICA	92.88% ± 4.61%	92.61% ± 5.49%
		Notch filter	92.52% ± 3.40%	91.89% ± 6.77%
		Notch filter, downsampling	92.76% ± 6.62%	92.81% ± 7.18%

Table 4.10: The results for classification of the original labels from 5-fold cross-validation over the training set on the Mentalab Explore+ 32 dataset. The best mean accuracies obtained are marked with bold lettering.

By converting the 1-9 label predictions to high/low labels, the accuracy increases to 96.43% ± 2.99% for valence and 96.24% ± 3.04% for arousal after downsampling, see table 4.11.

Predicting original labels and converting predicted labels to high/low arousal and valence.				
Features	Classifier	Preprocessing	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	91.92% ± 6.08%	92.24% ± 6.76%
		Downsampling	92.12% ± 8.54%	93.04% ± 5.66%
		ICA	91.66% ± 8.39%	92.70% ± 5.85%
		Notch filter	91.77% ± 8.76%	92.67% ± 5.80%
		Notch filter, downsampling	92.04% ± 8.85%	92.93% ± 5.78%
	LR	No preprocessing	95.77% ± 3.17%	95.99% ± 2.70%
		Downsampling	**96.43% ± 2.99%**	**96.24% ± 3.04%**
		ICA	96.03% ± 2.87%	96.05% ± 3.04%
		Notch filter	95.51% ± 4.18%	95.92% ± 3.46%
		Notch filter, downsampling	95.91% ± 4.21%	96.07% ± 3.69%

Table 4.11: The results for conversion into high/low labels from the prediction of the original labels from 5-fold cross-validation over the training set on the Mentalab Explore+ 32 dataset. The best mean accuracies obtained are marked with bold lettering.

4.5.2 Unicorn Hybrid Black

In table 4.12 the average accuracies achieved by the two models using different preprocessing techniques are reported. The performance is reported for the three cases (1), (2), and (3) mentioned in section 4.5.1.

The model's performance is very similar on the Mentalab 32+ Explore dataset and the Unicorn Hybrid Black dataset. The average accuracy is all over about one percentage point lower on the Unicorn Hybrid Black. The effects of applying the preprocessing to the data results in the same type of changes in accuracy for both datasets.

Using the raw data, both the SVC and LR models are able to predict at least 91% of the valence and arousal labels correctly for the high/low classification.

Downsampling is again the type of preprocessing that increases the accuracy the most. For case (1), the highest achieved average accuracy is 92.72% ± %2.50% for valence and 92.64% ± 3.11% for arousal with the SVC model. Using the approach of case (3) where you first classify the original (discrete 1-9) labels for valence and arousal and afterward convert them to high/low labels the average accuracy is increased to 95.49% ± 1.78% and 95.03% ± 2.04% for valence and arousal, respectively, with the LR model.

Features	Classifier	Preprocessing	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Predicting high/low arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	**SVC**	No preprocessing	91.54% ± 2.97%	91.00% ± 4.20%
		Notch Filter	90.45% ± 2.75%	90.45% ± 2.75%
		Downsampling	**92.72% ± 2.50%**	**92.64% ± 3.11%**
		Notch filter, downsampling	90.71% ± 2.68%	91.04% ± 3.49%
	LR	No preprocessing	91.27% ± 2.91%	91.32% ± 3.56%
		Notch Filter	90.69% ± 2.74%	90.90% ± 3.04%
		Downsampling	92.02% ±2.39%	92.11% ± 2.29%
		Notch filter, downsampling	90.96% ± 2.56%	91.16% ± 3.30%
Predicting the original (discrete 1-9) labels for arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	85.27% ± 5.38%	85.25% ± 5.87%
		Notch Filter	82.60% ± 3.49%	83.36% ± 3.77%
		Downsampling	87.16% ± 3.75%	87.91% ± 3.74%
		Notch filter, downsampling	91.05% ± 2.99%	**92.25% ± 2.90%**
	LR	No preprocessing	90.19% ± 2.88%	90.37% ± 3.32%
		Notch Filter	89.11% ± 2.57%	89.52% ± 2.71%
		Downsampling	**91.07% ±2.19%**	91.72% ±2.46%
		Notch filter, downsampling	89.57% ± 2.71%	90.17% ± 2.80%
Predicting original labels and converting predicted labels to high/low arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	92.19% ± 3.17%	92.50% ± 3.92%
		Notch Filter	90.56% ± 3.26%	91.94% ± 2.75%
		Downsampling	93.17% ± 2.31%	93.99% ± 2.49%
		Notch filter, downsampling	83.43% ± 4.14%	84.28% ± 4.06%
	LR	No preprocessing	94.99% ± 1.69%	95.21% ± 2.10%
		Notch Filter	94.23% ± 1.74%	94.93% ± 1.77%
		Downsampling	**95.49% ± 1.37%**	**95.80% ± 1.66%**
		Notch filter, downsampling	94.56% ± 1.78%	95.03% ± 2.04%

Table 4.12: The results from 5-fold cross-validation on the Unicorn Hybrid Black dataset. The best mean accuracies obtained for each method are marked with bold lettering.

4.6 Generalization properties

4.6.1 Mentalab Explore+ 32

Table 4.13 presents the average accuracy of the models on the Mentalab Explore 32+ validation dataset. The models are run on the raw data, the downsampled data, and the data that has been both notch filtered and downsampled.

For high/low classification, the average accuracy for valence classification is 85.08% ± 13.71% for the SVC model and 93.85% ± 2.93% for the LR model. For high/low valence classification the average accuracy is 84.69% ± 14.67% for the SVC model and 93.33% ± 3.12% for the LR model. The average accuracy is quite a bit higher for the LR model compared to the SVC, and the standard deviation is much lower.

The highest average accuracy in case (1) and (2) is achieved when applying only downsampling. For case (1) with high/low classification, the accuracy is 93.91% ± 2.97% and 94.18% ± 2.72% for valence and arousal

respectively. Interestingly, the accuracy of the model increases for the classification of the original (discrete 1-9) labels. The average accuracy increases to 94.65% $\pm$ 2.92% for valence and 94.42% $\pm$ 4.19% for arousal. By then converting the predicted 1-9 labels into high/low labels, an average prediction accuracy of 96.87% $\pm$ 2.11% and 96.95% $\pm$ 1.83% for valence and arousal is achieved with the downsampled data. For valence prediction in case (3), the average accuracy is 0.7 percentage points higher with the unprocessed data than with the downsampled data. The standard deviation is however 0.4 percentage points lower with the downsampled data.

The average accuracy of the model decreases to approximately 0.5-1 percentage point for all three cases when the notch filter is applied in addition to the downsampling. Additionally, the standard deviation increases somewhat.

The confusion matrices for classification with the downsampled LR model on the validation set for the participants with the most and least number of good channels are included in fig. 4.4 and fig. 4.5, respectively. In both figures, the subfigures (a) and (b) show the confusion matrices for arousal and valence from high/low classification, res (c) and (d) show confusion matrices for arousal and valence classification of the original (discrete 1-9) labels. Lastly, subfigures (e) and (f) show the confusion matrices when the predicted 1-9 labels are converted into high/low labels.

Participant no. 4 has 19 good channels. For this participant, the model works quite well. The high/low arousal labels are practically balanced, and the model does well at predicting both. The model has a very light preference toward predicting low arousal. The high/low valence labels are not balanced. There are a lot more instances of low valence than high valence in the data. Even though the dataset is quite unbalanced, the model does not have a strong preference for weather to predict high or low valence. It has 6 wrong high predictions and 8 wrong low predictions.

When looking at the confusion matrices for prediction of the original labels, not all values from 1 to 9 are present. For arousal, the labels are 3-9 and for valence, they range the interval 2-7. The datasets are not balanced but in both cases, there are a fair amount of instances for each label. For arousal, the model classifies all label classes wrong one or two times, except the label arousal=8. For valence, the model predicts arousal=2 correctly all 10 times and the label 6 correctly 29 out of 30 times. The model struggles most with classifying the instances with arousal=7. In this case 3 out of 35 epochs are classified incorrectly.

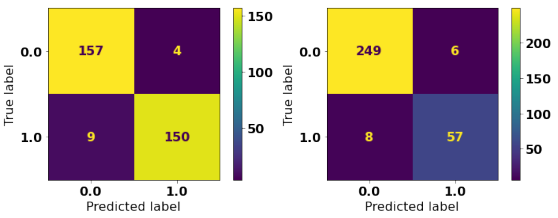
Features	Classifier	Preprocessing	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Predicting high/low arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	85.08% ± 13.71%	84.69% ± 14.67%
		Downsampling	92.67% ± 7.44%	93.72% ± 5.42%
		Notch filter, downsampling	92.62% ± 7.66%	93.93% ± 5.01%
	LR	No preprocessing	93.85% ± 2.93%	93.33% ± 3.12%
		Downsampling	**93.91% ± 2.97%**	**94.18% ± 2.72%**
		Notch filter, downsampling	93.76% ± 3.76%	94.12% ± 3.00%
Predicting the original (discrete 1-9) labels.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	87.27% ± 12.81%	15.41% ± 4.15%
		Downsampling	87.90% ± 14.22%	15.32% ± 4.80%
		Notch filter, downsampling	87.53% ± 15.28%	14.85% ± 4.86%
	LR	No preprocessing	94.33% ± 4.11%	94.25% ± 5.14%
		Downsampling	**94.65% ± 2.92%**	**94.42% ± 4.19%**
		Notch filter, downsampling	94.05% ± 4.31%	93.38% ± 6.10%
Predicting original labels and converting predicted labels to high/low arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	93.42% ± 6.34%	44.07% ± 9.09%
		Downsampling	93.83% ± 6.91%	44.62% ± 9.29%
		Notch filter, downsampling	93.75% ± 7.07%	44.69% ± 9.49%
	LR	**No preprocessing**	**96.94% ± 2.51%**	96.61% ± 2.43%
		Downsampling	96.87% ± 2.11%	**96.95% ± 1.83%**
		Notch filter, downsampling	96.56% ± 2.48%	96.31% ± 2.75%

Table 4.13: Accuracies obtained on the validation set with the Mentalab Explore+ 32 dataset. The best mean accuracies are marked with bold lettering.

Because the model usually predicts labels that are on or close to the true labels, converting the predicted 1-9 labels into high/low labels increases the accuracy further. For arousal, the predictions are very good. For valence, it is now apparent that the fact that there is more low valence than high valence epochs, makes the model somewhat prefer to predict low valence.

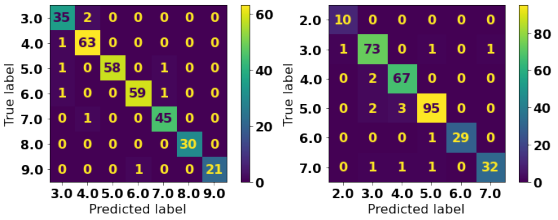
For participant no. 7, with only 8 good channels, the model predicts more false low instances than false high instances for both valence and arousal for the high/low predictions. The arousal dataset is somewhat unbalanced, whereas the valance dataset is very unbalanced. Both arousal and valence have more instances of low labels than high. It is worth noting that the model has a stronger preference for low arousal than it has for low valence, even though the arousal dataset is less balanced.

In the predictions of the original labels, the model falsely predicts label 4 for arousal especially many times. In the valence confusion matrix, there is not much that stands out, the incorrect predictions are spread across the board. When the predicted original labels are converted to high/low, half the false predictions are corrected compared to case (1) for arousal classification. For valence, 70% of the false low predictions are corrected, but only 12% of the false high predictions are corrected.



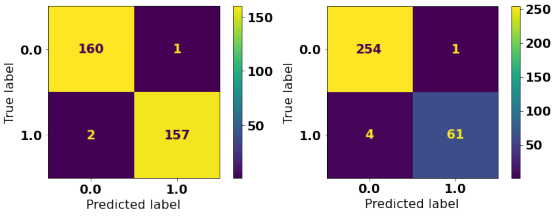
(a) Classification of high-/low arousal.

(b) Classification of high-/low valence.



(c) Classification of original (discrete 1-9) arousal labels.

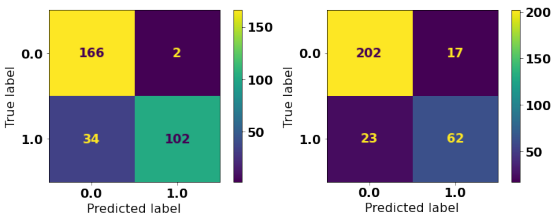
(d) Classification of original (discrete 1-9) valence labels.



(e) Converting the predicted 1-9 labels to high-/low arousal.

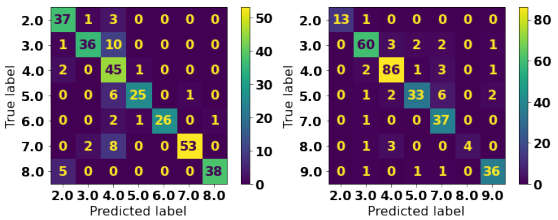
(f) Converting the predicted 1-9 labels to high-/low valence.

Figure 4.4: Confusion matrices for participant no. 4 with the downsampled Mentalab Explore+ 32 validation dataset. Participant no. 4 had 19 good channels.



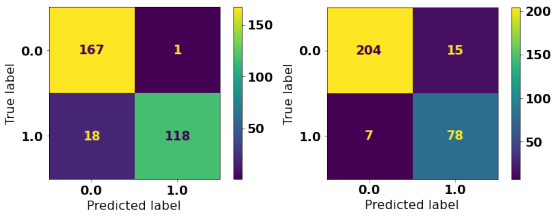
(a) Classification of high-/low arousal.

(b) Classification of high-/low valence.



(c) Classification of original (discrete 1-9) arousal labels.

(d) Classification of original (discrete 1-9) valence labels.



(e) Converting the predicted 1-9 labels to high-/low arousal.

(f) Converting the predicted 1-9 labels to high-/low valence.

Figure 4.5: Confusion matrices for participant no. 7 with the downsampled Mentalab Explore+ 32 validation dataset. Participant no. 7 had 8 good channels.

4.6.2 Unicorn Hybrid Black

The model performances on the Unicorn Hybrid Black validation dataset are presented in table 4.14. The average accuracy of the SVC and the LR models are presented for cases (1), (2), and (3).

As with the Mentalab Explore+ 32 dataset, the best results are obtained with the SVC model on the downsampled dataset. With high/low classification, average accuracies of 93.81% $\pm$ 2.54% and 93.80% $\pm$ 2.70% are obtained for valence and arousal, respectively. The performance can be increased by using the classification method of case (3) with the LR model. With that method, the accuracy increases to 96.60% $\pm$ 1.56% for valence and 96.63% $\pm$ 1.53 for arousal.

The LR model is able to achieve an accuracy as high as 93.03% $\pm$ 2.58% for valence and 92.89% $\pm$ 2.38% for arousal when predicting the original (discrete 1-9) labels from downsampled data.

The confusion matrices for the downsampled LR model on the validation set are included below. Figure 4.6 shows the confusion matrices for participant no. 11, which was one of the participants that achieved the lowest accuracy with the model. Figure 4.7 shows the same matrices for the best-performing participant, no. 15.

The high/low classification for participant no. 11 is better for low arousal and valance than for high arousal and valance epochs. The dataset is not balanced in a manner that there are more instances of low than high labels for both valance and arousal. There are in total 104 epochs that have arousal labels equal to 5. This can be seen in the confusion matrix for the classification of original labels. Labels 5 or smaller are defined as ”low”, and this has a big impact on the fact that the low arousal category is so much bigger than the high arousal, even though arousal equal 5 is very neutral. There are epochs with labels 1-8 for arousal and 1-9 for valance. Overall, the model does

about the same for categorizing all of them.

Converting the predicted original labels to high/low increases the accuracy, as always. For arousal, the improvement is about the same for both high and low instances. For valence, the improvement is biggest in predicting the high valence epochs correctly.

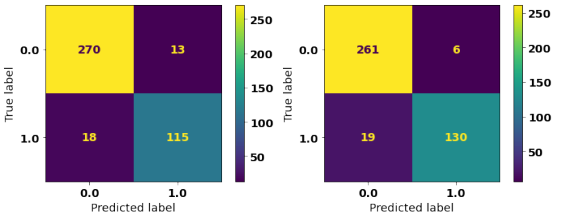
The dataset is very unbalanced for participant no. 17, which affects the accuracy. For this participant, the cut-off between high and low arousal had to be changed so that some of the epochs would be categorized as low arousal. In this case, arousal ≤ 6 is defined as low arousal. For high/low arousal classification, the model categorizes all instances correctly. But it is important to note that there are only 27 low arousal epochs. For valance, the model predicts three labels incorrectly. What is interesting is that these three labels are falsely categorized as high, even though the dataset is very unbalanced toward low valance.

Looking at the confusion matrices for classifying original labels, we see that there are only arousal labels from 6 to 9 and valence labels from 3 to 6. The participant mainly experienced that the video stimuli elicited an arousal of 8 and a valence of 5. Even though the model did perfectly on high/low classification for arousal, there are some mistakes when predicting the original labels.

For case (3) the mistakes made in the classification of the original arousal labels actually carry on into the conversion to high/low labels, resulting in one false low arousal prediction. For valance, this method corrected the errors in case (1), and now all epochs are predicted correctly.

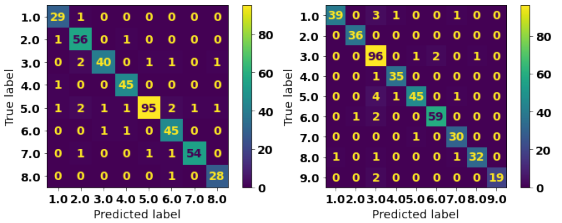
Features	Classifier	Preprocessing	Valence Average accuracy ± STD	Arousal Average accuracy ± STD
Predicting high/low arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	**SVC**	No preprocessing	92.15% ± 3.13%	92.13% ± 4.14%
		Downsampling	**93.81% ± 2.54%**	**93.80% ± 2.70%**
		Notch filter, downsampling	91.88% ± 2.90%	92.23% ± 3.48%
	LR	No preprocessing	92.13% ± 2.64%	92.23% ± 3.29%
		Downsampling	93.26% ± 2.74%	93.05% ± 2.44%
		Notch filter, downsampling	91.90% ± 2.83%	92.32% ± 2.80%
Predicting the original (discrete 1-9) labels.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	87.78% ± 5.01%	17.13% ± 10.09%
		Downsampling	89.46% ± 3.13%	16.83% ± 9.80%
		Notch filter, downsampling	86.64% ± 3.89%	17.02% ± 9.78%
	LR	No preprocessing	92.00% ± 3.23%	92.31% ± 2.65%
		Downsampling	**93.03% ± 2.58%**	**92.89% ± 2.38%**
		Notch filter, downsampling	92.08% ± 2.85%	91.47% ± 2.89%
Predicting original labels and converting predicted labels to high/low arousal and valence.				
HC, HM, DE, FBE, HMS, HCS	SVC	No preprocessing	93.45% ± 3.25%	45.24% ± 13.13%
		Downsampling	94.50% ± 2.42%	45.51% ± 13.41%
		Notch filter, downsampling	92.95% ± 3.04%	45.76% ± 13.52%
	LR	No preprocessing	95.82% ± 1.86%	96.09% ± 2.01%
		Downsampling	**96.60% ± 1.56%**	**96.63% ± 1.53%**
		Notch filter, downsampling	96.03% ± 1.94%	95.98% ± 2.06%

Table 4.14: Accuracies obtained on the validation set with the Unicorn Hybrid Black dataset. The best mean accuracies are marked with bold lettering.



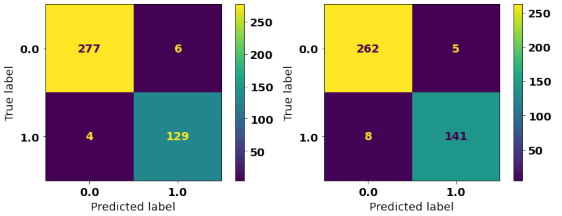
(a) Classification of high-/low arousal.

(b) Classification of high-/low valence.



(c) Classification of original (discrete 1-9) arousal labels.

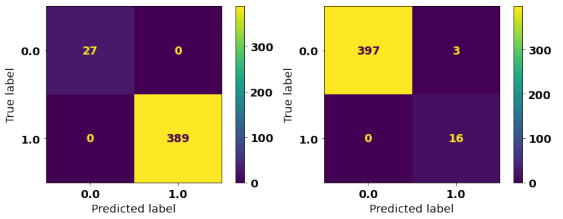
(d) Classification of original (discrete 1-9) valence labels.



(e) Converting the predicted 1-9 labels to high-/low arousal.

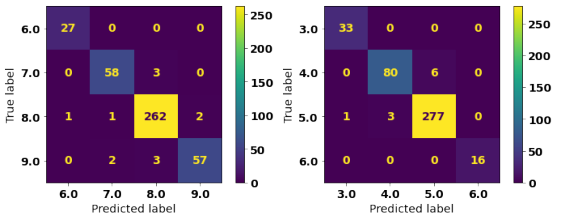
(f) Converting the predicted 1-9 labels to high-/low valence.

Figure 4.6: Confusion matrices for participant no. 11 with the downsampled Unicorn Hybrid Black validation dataset.



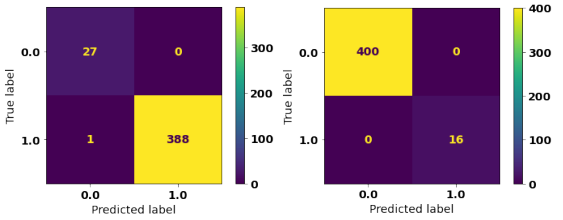
(a) Classification of high-/low arousal.

(b) Classification of high-/low valence.



(c) Classification of original (discrete 1-9) arousal labels.

(d) Classification of original (discrete 1-9) valence labels.



(e) Converting the predicted 1-9 labels to high-/low arousal.

(f) Converting the predicted 1-9 labels to high-/low valence.

Figure 4.7: Confusion matrices for participant no. 15 with the downsampled Unicorn Hybrid Black validation dataset.

5

Discussion

5.1 Channel selection

NSGA-II was able to improve the prediction accuracy of the SVC model marginally from 96.50%±2.10% to 96.52%±.96% for high/low valence classification by removing three channels. This improvement in accuracy is however not statistically significant. The accuracy of the high/low arousal classification decreased from 96.71%±.54% to 96.63%±.13% for the same solution.

It is however possible to reduce the number of channels drastically and still maintain a good emotion recognition performance. The model was able to predict on average over 95% of the labels correctly by using on average 15-16 electrodes. The original dataset has 32 electrodes, so this solution utilizes only half of that for a very small trade-off in accuracy.

In this project, the focus during the NSGA-II stage was mostly on trying to improve prediction accuracy by reducing the number of channels. When solutions were picked from the Pareto-front, the solutions with the highest accuracy instead of the fewest channels were chosen. More solutions using fewer channels should have been looked into but unfortunately, time did not allow for this. I hypothesize that it would have been possible to maintain almost the same prediction accuracy with even fewer electrodes.

The Mentalab Explore+ 32 dataset and the Unicorn Hybrid Black datasets both have less than 32 electrodes. The Mentalab Explore+ 32 dataset has from 8 to 19 channels, depending on the participant. Using 12 channels, the model is still able to achieve an accuracy well above 90% for binary arousal and valence classification from the raw data. It also has to be taken into account that the locations of these channels are not selected through any channel selection algorithm, but were a random selection due to bad data at malfunctioning channels.

The Unicorn Hybrid Black dataset utilizes 16 channels whose locations were chosen in order to maximize the emotion information in the recorded EEG data. In this dataset, an accuracy of almost 94% was obtained for binary arousal and valence classification on the downsampled data.

5.2 Data collection

Even though the EMDB is created to elicit a close-to-balanced dataset in terms of arousal, valence, and dominance, this is not necessarily the case with new participants. Some of the participants were very excited to be part of this project, which affected their emotional state. One participant reported their state of arousal as 6 or higher during all videos, even for the more calming videos like the landscape scenes.

The impedance of the channels was assessed before data collection with the Mentalab Explore+ 32 device. It was very participant-dependent what impedance we were able to achieve on the channels. One participant had impedance$\leq 37k\Omega$ on all channels, while another participant had an impedance equal to $104k\Omega$ on one channel. The difference in impedance did not seem to affect the emotion recognition ability of the models.

The DEAP dataset that was utilized in the specialization project used music videos as emotional stimuli. The emotional stimuli during data collection in this project were non-auditory movie clips. If the model only worked well for the DEAP dataset and not for the newly collected datasets, this could have been a sign that the models actually decoded the brain processing the music and not actually the emotions. The fact that the proposed models perform well for both types of stimuli suggests that the models decode emotions.

The data collected in this project til be used in further work at NTNU. Hopefully, it will also be published publicly online in the future. Collecting new data is quite time-consuming, and often takes more time than anticipated due to unforeseen problems such as the ones mentioned in section 3.5.2, but it is a very important step in improving the research. It is important to have different datasets in order to validate the robustness of the model.

5.3 Number of features

In the specialization project, it was determined that the five features HC, HM, DE, FBE, and HMS were preferable as input into the SVC. At that point, the statistical significance between the accuracy of that model and the SVC model with six features (HCS in addition to the other five) was not looked into. After working more with the SVC models during this project, it seems like HCS has a positive effect on the robustness and generalization properties of the model overall.

5.4 Data preprocessing

As seen in section 4.5, many of the preprocessing steps that were explored in this project did not necessarily increase the accuracy. As discussed, the frequency of the power line is very visible in the power spectrum distribution plot. This noise does not bring any useful information into the dataset, and it is, therefore, beneficial to remove it. In this project, it was removed with the notch filter algorithm that is implemented in the MNE-Python package. For valence predictions, applying the notch filter always lead to poorer classification accuracy. For arousal classification, the notch filter improved the accuracy somewhat in some of the situations with the LR model. The fact that the notch filter mostly had a negative effect on the prediction accuracies of the model is a sign that the notch filter removed some of the relevant information to emotion decoding.

Eye-blink removal proved itself to not be as straightforward as anticipated. ICA was only performed fully on the Mentalab Explore+ 32 dataset because this dataset had enough channels to locate the source more precisely. With the Unicorn Hybrid Black system, two 8-channel devices were used simultaneously to record the data. When trying to apply ICA, the channels from each device were looked at independently. Because of the few numbers of channels, ICA was not able to extract only the blinks in one source. I later realized the problem this caused and should have combined the data from the two devices before running ICA. That way the algorithm would find 16 instead of 8 independent components/sources, and would have had a better chance of identifying the blinking source.

For the Mentalab Explore+ 32 dataset, removing the sources that were determined to be eye blinks decreased the prediction accuracies of the LR model. Again, one has to assume that this might be because some useful information is removed from the model. The decrease is however so small that it might just be due to chance, and should be looked further into. For the SVC model, ICA improved the arousal classification but weakened the valence classification.

The models obtain good predictions with the eye blinks present. The electrooculography artifacts were removed from the preprocessed DEAP dataset, but from the results obtained in this project, the models seem to be able to look away from the eye blinks and not let that determine the predictions.

The preprocessing method that reliably improved the performance was downsampling. By downsampling the data, only the bigger more impactful structures of the data are kept. Downsampling is a way of removing less important data so that the model can focus on the important data structures. In this project, the data were downsampled from 250 Hz to 128 Hz. The choice of the downsampled frequency was based on the downsampling done in the DEAP dataset, as the model was trained on this data. Other downsampling rates should also be looked into.

It is, however, positive that the model works quite well with the raw data. In a real-life situation where the system is to be used in real-time, it is beneficial that not much processing of the data has to be done as this causes latency.

5.5 Generalization properties

The six-feature SVC model generalizes somewhat better to the new datasets than the five-feature model. In this project, the extra feature was included in order to look at the best possible accuracies. Increasing the complexity somewhat by adding this feature might not be necessary, as the increase in accuracy and robustness is not so big.

Both the SVC and LR model generalizes quite well to the new datasets. The average accuracy of the SVC model for binary classification on the downsampled Mentalab Explore+ 32 data is 92.67% $\pm$ 7.44% for valence and 93.72% $\pm$ 5.42% for arousal. This is a decrease of 3,76 percentage points for valence classification and 2,86 percentage points for arousal compared to the six-feature SVC model on the DEAP dataset. With LR the model achieved 93.91% $\pm$ 2.97% for binary valence classification and 94.18% $\pm$ 2.72% for arousal. Compared to the LR on the DEAP dataset, this is a decrease of 2.35 and 2,43 percentage points for the classification of valence and arousal, respectively. The decrease in accuracy is somewhat bigger with the SVC model than with the LR model, which might be a sign that the LR model generalizes a bit better to unseen data.

With binary classification on the downsampled Unicorn Hybrid Black dataset, the SVC achieves 93.81% $\pm$ 2.54% for valence and 93.80% $\pm$ 2.70% for arousal. This accuracy is better than the ones obtained on the Mentalab Explore+ 32 data. This is likely due to the fact that the latter dataset misses a lot of the recorded EEG data due to a malfunctioning device. There are a lot fewer epochs for the model to be trained on, and the dataset might be very unbalanced. It is however a very good sign that the model is able to achieve good predictions on both the Mentalab Explore+ 32 data and the Unicorn Hybrid Black data. The model handles both the small and bigger dataset well, which is a sign that it is quite robust and good at reading the underlying patterns of emotions. In addition to this, the EEG data is recorded using different devices and even different types of electrodes (wet and dry), but the model works well with both setups.

The LR model predicts 93.26% $\pm$ 2.74% of the valence labels correctly for the binary classification of the Unicorn Hybrid Black dataset. The average accuracy of the arousal predictions is 93.05% $\pm$ 2.44. In this case, the SVC model performs better than the LR model.

From the confusion matrices, we see that the LR model somewhat prefers to predict the labels that are most highly represented in the dataset, but even with very unbalanced datasets the model usually makes the correct predictions. This shows that the model is actually able to differentiate between the classes, even with few examples for one class.

Both the LR model and the SVC model that were proposed in the specialization project perform well on the two new datasets. Both the LR and the SVC model obtain accuracies around 93% on the downsampled datasets. This shows that the good performance on the DEAP data was not just due to overfitting, but the fact that the model is able to identify the important structures in the EEG data that correspond to emotions. It is not clear from the results obtained in this project whether the SVC or LR model is better for binary classification as long as the data is downsampled. The number of epochs, number of channels, and the channel locations are different in the two datasets, which may influence what model is better for the classification. It may also be due to individual differences in the participants.

The LR model also does quite a good job of predicting the original labels from the self-assessment. The LR model is able to differentiate between smaller nuances between the different levels of arousal and valence. This could be very useful for more practical uses of emotion recognition technology. During self-assessment in the experiments, the participants also rated their experienced level of dominance. This was not looked into during the project, but it would be interesting to see if also this could be predicted by the same model.

By letting the model train on the original labels, the model receives more information. If one still wants to classify only high and low levels of arousal and valence, the model performances can be increased by letting the model predict the original labels and converting these predicted labels into binary labels. This created an increase in the average accuracy of about 2 percentage points for both datasets.

6

Conclusion and Future work

6.1 Conclusion

This master's thesis has been a continuation of the work done in my specialization project [18], and has further explored the models proposed in that report. In this thesis, channel selection through NSGA-II has been explored for the DEAP dataset. The design of experiments and collection of data for two new datasets has been carried out. 52 40-second emotional movie clips were used as stimuli during experiments. The experiment was the same for both datasets, but different EEG devices have been used to record the EEG data. One device had 32 electrodes, where some of the channels had to be removed during data processing, and the other had 16 channels. Preprocessing as notch filtering (to remove noise from the power line), downsampling, and ICA (to remove eye blinks) was applied to the data. Lastly, the generalization properties of the models were tested on the new datasets.

From the work in this thesis, it can be concluded that good emotion decoding can be done using way fewer electrodes than 32. Using 16 channels, the average binary classification accuracy is above 93% for the Unicorn Hybrid Black dataset. Through further investigation of the channels, and possibly having participant-wise channel selection, the number of channels can probably be decreased even more without a big loss of accuracy.

Preprocessing the data can increase the prediction accuracy of the SVC model, but it is not necessary for the LR model. Downsampling is however a good method for removing unnecessary information from the data.

Both the proposed SVC and LR models are stable and reliable on the new datasets, so it can be concluded that the models are robust across recording devices, types of stimuli, and participants. The models are expected to generalize well to more new data. Although the accuracy is somewhat lower than the reported accuracies of the complex deep-learning state-of-the-art models, this shows that simple machine-learning models can be used for good emotion recognition.

The findings in this thesis suggest that the simple logistic regression model using six features can produce good emotion predictions with 16, maybe fewer, channels and without the need for preprocessing. The model is not only able to differentiate between high/low arousal and valence, but it can predict much more nuanced levels of arousal and valence with almost the same accuracy.

6.2 Future Work

Channel selection should be looked into even further. It would be beneficial to know more about the trade-off between the number of channels and prediction accuracy. What is actually most important for practical use? A system will not be used by patients if it is very big and uncomfortable, even if the accuracy is very good. Another important part of the channel selection problem is the location of the channels. In this project, the NSGA-II channel selection is done participant-wise but is there a universally good and stable channel selection for emotion recognition?

For this technology to someday be used by patients in a medical setting, it would have to be reliable, accurate, wearable for a long or semi-long period of time, and hopefully fast. The importance of each of these factors should be defined more clearly. Is it most important that the emotion decoding is correct every time, or that the system can read emotions close to instantly? If comfort level is the most important thing, how much accuracy are patients willing to trade for fewer channels? The type of recording device would also be important, as some electrodes are more comfortable than others, that way patients might accept more channels.

Bibliography

[1] Arnaud Delrome, . Infomax independent component analysis for dummies. URL: https://arnauddelorme.com/ica_for_dummies/. Accessed: 21.05.2023.

[2] bio-medical, . Electro-gel for electro-caps - 16 oz. URL: https://bio-medical.com/electro-gel-for-electro-caps.html. Accessed: 20.05.2023.

[3] Blank, J., Deb, K., 2020. pymoo: Multi-objective optimization in python. IEEE Access 8, 89497–89509. doi:10.1109/ACCESS.2020.2990567.

[4] Carvalho, S., Leite, J., Galdo-Alvarez, S., Gonçalves, , 2012. The emotional movie database (emdb): A self-report and psychophysiological study. Applied psychophysiology and biofeedback 37. doi:10.1007/s10484-012-9201-6.

[5] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Transactions on Evolutionary Computation 6, 182–197. doi:10.1109/4235.996017.

[6] g.tec medical engineering, 2023. g.gammasys active wet eeg electrodes. URL: https://www.gtec.at/product/ggammasys/. Accessed: 16.05.2023.

[7] Gannouni, S., Aledaily, A., Belwafi, K., Aboalsamh, H., 2021. Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. Scientific Reports 11. doi:10.1038/s41598-021-86345-5.

[8] g.tec medical engineering, 2022. Home — unicorn-bi.com. URL: https://www.unicorn-bi.com/. Accessed: 16.05.2023.

[9] Huang, D., Guan, C., Ang, K.K., Zhang, H., Pan, Y., 2012. Asymmetric spatial pattern for eeg-based emotion detection, in: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. doi:10.1109/IJCNN.2012.6252390.

[10] Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13, 411–430. URL: https://www.sciencedirect.com/science/article/pii/S0893608000000265, doi:https://doi.org/10.1016/S0893-6080(00)00026-5.

[11] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., 2012. DEAP: A database for emotion analysis ;using physiological signals. IEEE Transactions on Affective Computing 3, 18–31. doi:10.1109/T-AFFC.2011.15.

[12] Kothe, C., 2016. GitHub - chkothe/pylsl: Python interface to the Lab Streaming Layer. URL: https://github.com/chkothe/pylsl. Accessed: 01.03.2023.

[13] Kragel, P., Knodt, A., Hariri, A., LaBar, K., 2016. Decoding spontaneous emotional states in the human brain. PLoS biology 14, e2000106. doi:10.1371/journal.pbio.2000106.

[14] Lindquist, K., Wager, T., Kober, H., Bliss-Moreau, E., Barrett, L., 2012. The brain basis of emotion: A meta-analytic review. The Behavioral and brain sciences 35, 121–43. doi:10.1017/S0140525X11000446.

[15] Mentalab, 2022. Mentalab explore+ quick start guide. URL: https://wiki.mentalab.com/pdfs/Mentalab_Explore_Plus_Quick_Start_Guide.pdf. Accessed: 16.05.2023.

[16] Mentalab, 2023. Meet mentalab explore. URL: https://mentalab.com/mobile-eeg/. Accessed: 16.05.2023.

[17] Moctezuma, L., Abe, T., Molinas, M., 2022. Two-dimensional CNN-based distinction of human emotions from EEG channels selected by multi-objective evolutionary algorithm. Scientific Reports 12. doi:10.1038/s41598-022-07517-5.

[18] Neverlien, E.C.S., 2022. Automatic emotion decoding from EEG signals using the DEAP dataset. doi:10.13140/RG.2.2.36600.11522.

[19] Pane, E.S., Wibawa, A.D., 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), pp. 14–19. doi:10.1109/CENIM.2018.8711196.

[20] R. Krol, L., 2020. File:EEG 10-10 system.svg - Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:EEG_10-10_system.svg.

[21] Srinivas, N., Deb, K., 1994. Muiltiobjective optimization using nondominated sorting in genetic algorithms. Evol. Comput. 2, 221–248. URL: https://doi.org/10.1162/evco.1994.2.3.221, doi:10.1162/evco.1994.2.3.221.

[22] TerniMed, . Signa gel electrode gel. URL: https://www.ternimed.de/Signa-Gel-Electrode-Gel. Accessed: 20.05.2023.

[23] Unho Choi, . Pca ica ? URL: https://wooono.tistory.com/389. Accessed: 21.05.2023.

[24] Zheng, W.L., 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, 162–175. doi:10.1109/TAMD.2015.2431497.

[25] Zilio, F., 2017. Review: Georg northoff, minding the brain. a guide to philosophy neuroscience. Universa. Recensioni di Filosofia 6, 65–70. doi:10.14658/URF-6-2.

Appendix

A Information 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.

B Abstract accepted by 10th International BCI Meeting

EEG-based Automatic Emotion Recognition Using Machine Learning

Rose Lu[1⋆], Embla C. S. Neverlien[1], Mohit Kumar[1], and Marta Molinas[1]

[1]*Norwegian University of Science and Technology, Trondheim, Norway*
⋆O. S. Bragstads Plass 7034, Trondheim, Norway. E-mail: roselu@stud.ntnu.no

Introduction: The field of EEG-based emotion recognition has been widely explored in the last decade. Methods proposed by recent literature mostly use complex deep learning methods to achieve good predictions. To obtain comparable results using simpler machine learning techniques would be preferable as these models are much more intuitive to implement and understand. This work explores support vector machine (SVM) classifier on two datasets (DEAP and SEED) and compares the performance with more complex models.

Material, Methods and Results: This work uses the publicly available datasets, DEAP [1] and SEED [2], to evaluate SVM. The preprocessed EEG data from DEAP is segmented into one-second epochs without overlapping, resulting in total 2400 epochs. Further, it is filtered into four frequency sub-bands using the Butterworth bandpass filter. The data from SEED is also segmented into epochs of one second without overlapping, resulting in a total of 3394 epochs. Several features and their different combinations are extracted from each epoch of the EEG signal. For DEAP, the same features are derived for each one-second epoch in a three-second baseline recording. The computed features for each trial are corrected by subtracting the average value of the baseline features. To decipher the emotions, the extracted features are used as an input to SVM with RBF kernel and K-nearest neighbor (KNN) classifiers. The performance of SVM is found to be superior as compared to KNN. The best average accuracies achieved for each dataset are presented in Table 1. For DEAP, the most suitable feature set is the combination of Hjorth mobility (HM), Hjorth complexity (HC), differential entropy (DE), frequency bands energy (FBE), and Hjorth mobility spectrum (HMS). For SEED, the feature HM is found to be more useful than the combination of several features in detecting human emotion. The performance of the SVM on both datasets is verified using a 5-fold cross-validation approach.

Dataset	Feature	Accuracy
DEAP arousal	HC, HM, DE, FBE, HMS	96.71 %
DEAP valence	HC, HM, DE, FBE, HMS	96.50 %
SEED	HM	83.87 %
SEED	HC, HM, DE, FBE, HMS	63.76 %

Table 1: *Mean accuracies*

Discussion: It can be observed from the obtained results that the simple machine-learning techniques can produce satisfactory predictions for EEG-based emotion recognition. The obtained results for the model proposed for DEAP are comparable to the ones obtained with deep learning [3]. For SEED, deep learning methods outperform the best-performing proposed model by almost 11 percentage points [4]. Nevertheless, both the proposed models show promise for utilizing simpler models in EEG-based emotion recognition. In future work, a new dataset will be developed to test the generalization of the proposed machine-learning method. And, a channel selection approach will be adopted for complexity reduction of the models. Related studies [2, 5] have shown that the number of electrodes can be reduced significantly without decreasing the prediction accuracy of the model.

Significance: The potential for emotion recognition devices will be huge both medically and commercially given that only a few electrodes in combination with simple and fast models perform well.

References:

[1] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, pages 18–31, 2012.

[2] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, pages 162–175, 2015.

[3] Guowen Xiao, Meng Shi, Mengwen Ye, Bowen Xu, Zhendi Chen, and Quansheng Ren. 4D attention-based neural network for EEG emotion recognition. *Cognitive Neurodynamics*, 16:1–14, 08 2022.

[4] Fangyao Shen, Guojun Dai, Guang Lin, Jianhai Zhang, Wanzeng Kong, and Hong Zeng. EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cognitive Neurodynamics*, 14:1–14, 12 2020.

[5] Luis Moctezuma, Takashi Abe, and Marta Molinas. Two-dimensional CNN-based distinction of human emotions from EEG channels selected by multi-objective evolutionary algorithm. *Scientific Reports*, 12, 03 2022.

First and second author contributed equally to the abstract.

C Paper accepted by 45th Annual Conference of the IEEE engineering in the Medicine and Biology society

Decoding Emotions From EEG Responses Elicited by Videos Using Machine Learning Techniques on Two Datasets

Embla C. S. Neverlien*, Rose Lu, Mohit Kumar, and Marta Molinas

***Abstract*— In recent times, we have seen extensive research in the field of EEG-based emotion identification. The majority of solutions suggested by current literature use sophisticated deep learning techniques for the identification of human emotions. These models are very complex and need huge resources to implement. Hence, in this work, a method for human emotion recognition is proposed which is based on much simpler architecture. For that, two publicly available datasets SEED and DEAP are used to perform experiments. First, the EEG signals of the two datasets are segmented into epochs of 1-second duration. The epochs are also decomposed into different brain rhythms. The features computation is performed in two different ways, one is directly from the epochs and the other way is from the brain rhythms obtained after the decomposition of the epochs. Several features and their combination are examined with different classifiers. For the DEAP dataset baseline features are also utilised. It is observed that the support vector machine (SVM) has shown the best performance for the DEAP dataset when baseline feature correction and epoch decomposition are implemented together. The best achieved average accuracy is 96.50% and 96.71% for high versus low valence classes and high versus low arousal classes, respectively. For the SEED dataset, the best average accuracy of 86.89% is achieved using the multilayer perceptron (MLP) with 2 hidden layers.**

***Clinical relevance*— This work can be further explored to develop an automated mental health monitor which can assist doctors in their primary screening.**

I. INTRODUCTION

Human emotions are fundamental to non-verbal communication and play a significant role in our daily lives. Several aspects of life are affected by the emotional state such as decision-making, communication skills, behaviour, mental and physical health. Automated emotion recognition can improve the quality of service in these areas. The EEG-based methods are the most reliable and common techniques as these approaches are less vulnerable to imposters [1].

In literature, several researchers have focused on identifying human emotions using EEG signals [2]–[4]. In [2], various emotions such as, joy, pleasure, anger, and sadness are separated using support vector machine (SVM) based method. In [3], time-domain based features are used with SVM classifier for the detection of different human emotions. Fourier transform, wavelet transform, and power spectral density, based method are found to be useful to distinguish among the different human emotions [4]. A multiresolution analysis based feature extraction method is proposed in [5]. Differential entropy (DE) is widely used for human emotions classification and has shown its effectiveness for emotional states representation in EEG signals [1], [6]–[9].

In recent years, deep learning based methodologies are widely explored for emotion detection using EEG signals [7], [8], [10], [11]. DE is used with a deep belief network to discriminate among positive, neutral, and negative emotional states of humans [7], [10]. A long short-term memory (LSTM) based method is explored to identify the temporal information present in the EEG signals corresponding to different emotions [11]. To identify the spatial relationship among various channels in EEG signals related to different emotional states, a convolution neural network (CNN) based approach is suggested in [8]. In [1], a parallel convolution recurrent neural network is proposed which utilized both CNN and LSTM for discriminating the different human emotions. The EEGNet is also explored for the classification of different human emotions [12].

In this work, we aim to explore the less complex machine learning techniques to design an automated method for the identification of human emotions using EEG signals as the current state-of-the-art methods [1], [9] are based on very complex network architectures which require huge computational resources. For that, we have explored various machine learning methods on the two publicly available datasets namely, SEED and DEAP.

All the authors are with the Department of Engineering Cybernetics, Norwegian University of Science and Technology, Trondheim, Norway ecneverl@stud.ntnu.no

II. MATERIALS AND METHODS

A. Description of dataset

The two publicly available datasets which are used to evaluate the performance of the models are described below:

The DEAP dataset [13] describes emotions on a two-dimensional spectrum spanning arousal and valence, i.e. intensity and pleasantness of an emotional state. The dataset contains 32-channel EEG recordings from 40 trials on each of the 32 participants with a sampling frequency of 512 Hz. These signals are downsampled to 128 Hz and preprocessed using a bandpass filter (BPF) having a cutoff frequency of 4-45 Hz. After preprocessing, the final data has 3 sec. baseline and 60 sec. trial recordings. Further information regarding the data is available in [13]

The SEED dataset [7] consists of 62-channel EEG signals from 45 trials on 15 participants in three different sessions with a sampling frequency of 1000 Hz. These signals are downsampled to 200 Hz and preprocessed using a bandpass filter of 0-75 Hz [14]. The emotions are categorized as positive, negative and neutral. Further information regarding the data is available in [7].

B. Framework of the proposed method

The framework of the proposed method includes mainly three parts: segmentation and decomposition of EEG signals into epochs and sub-bands, feature extraction, and subject-wise classification. The open-source MNE-Features module is used for feature extraction [15].

1) DEAP dataset: The data is segmented into 2400 1-second trial epochs and 120 1-second baseline epochs for each participant. The epochs have no overlap. In the work presented here, the effects on model performance are explored when introducing frequency sub-bands and baseline removal through four different cases: 1. extraction of features from the data without any decomposition and baseline removal, 2. introduce frequency sub-bands before feature extraction, 3. feature extraction followed by baseline removal, and 4. introduce frequency sub-bands before both feature extraction and baseline removal. For case 1, the features are calculated for each channel and every epoch, leading to 32 features. In case 2, the data is filtered into frequency sub-bands theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz), and gamma (31-45 Hz) using the Butterworth band-pass filter. These sub-bands are known to be strongly associated with emotions [7], [16]. Feature extraction is then performed, resulting in 128 features. For case 3, feature extraction is performed on both the trial and baseline epochs. Next, the average value (average over 3 epochs of 1-second) of the baseline features is calculated for each trial. The average baseline features are then subtracted from the corresponding trial features. In case 4, the data is decomposed into sub-bands before performing the feature extraction and baseline removal. The features to be extracted are Hjorth mobility in the time domain (HM) and Hjorth mobility derived from the power spectrum (HMS), Hjorth complexity in the time domain (HC), and Hjorth complexity derived from the power spectrum (HCS), DE and frequency band energy (FBE). These features are explored individually and in combination. When features are combined, normalization is applied to the feature matrix. Two independent classifiers are trained to classify the two different sets of classification problems, one is high versus low arousal (HLA) and the other is high versus low valence (HLV) of human emotions for each participant.

2) SEED dataset: For each trial from the SEED dataset, the provided data is segmented into 3394 1-second epochs without overlapping. Unlike the DEAP dataset, the SEED dataset has no baseline signals available. Three approaches are investigated to evaluate the model performances. In the first approach, individual features are calculated channel-wise on each epoch. This leads to 62 features in total. The second approach includes combining different features. Features from the time domain, frequency domain, and also nonlinear features are combined and explored. For the third approach, decomposition into five sub-bands before feature extraction is examined. The five sub-bands are delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-50 Hz) [14]. Features of interest in this work are HM, HC, Katz fractal dimension (KFD), Higuchi fractal dimensions (HFD), and power spectral density (PSD). In addition, several combinations of them are also explored.

C. Classifiers

To classify the EEG signals, several classifiers have been used in this work. Table I shows an overview of the classifiers and the details of the used parameters. For KNN, K = {3,5} is tested. For SVM, the radial basis function (RBF) kernel is used.

For multilayer perceptron (MLP), three simple architectures were tested, i.e. MLP-v1, MLP-v2, and MLP-v3. MLP-v1 has two hidden layers with 100 and 50 nodes, respectively, and a dropout rate of 0.1. MLP-v2 also consists of two hidden layers, although the number of nodes is 500 and 300. A dropout rate of 0.2 is applied as well. MLP-v3 has the same dropout rate as MLP-v1 but consists of three hidden layers of 2000, 1000 and 500 nodes instead. For all MLP, the activation function used is ReLU and the output is fed into a softmax classifier. The loss function is set to the categorical cross-entropy loss and the Adam optimizer is used. The number of nodes and dropout rate is selected through the trial and experimentation method.

TABLE I: Details about the used classifiers

Classifier	Parameter details
KNN	K = {3,5}
SVM	Kernel: RBF Decision function: One-vs-One
MLP-v1	2 hidden layers with 100 and 50 nodes dropout rate: 0.1
MLP-v2	2 hidden layers with 500 and 300 nodes dropout rate: 0.2
MLP-v3	3 hidden layers with 2000, 1000 and 500 nodes dropout rate: 0.1

III. Results and Discussion

The obtained results in several experiments performed on the DEAP and SEED datasets with different features and classifiers are summarised below.

A. DEAP dataset

When assessing the individual features the data is split into 67% training set and 33% test set. The performance of each model was determined by the classification accuracy on the test set. Table II contains the average accuracies of the four cases mentioned in section II-B.1 for the two best-performing features using the SVM classifier for HLV and HLA category. The average accuracy of a model is the mean value of the prediction accuracies of all 32 participants, using that model. HC and HM were the two best-performing individual features, and are therefore presented here. Baseline removal has improved the performance of the model.

The best-performing models are, however, obtained by employing both sub-bands and baseline removal. This was true for all six features. For both HM and HC, the performance has increased by over 20 percentage points as compared to directly computing the features from the data. Using HC, the

average accuracy is 88.63% for HLV and 89.28% for HLA. The model performed best on participant nr. 23 achieving prediction accuracies of 94.70% for HLV and 95.71% for HLA. Participant nr. 22 got the worst prediction accuracies with 76.52% for HLV and 74.87% for HLA.

The 5-fold cross-validation method was used to test the best-performing combination of features for both HLV and HLA categories. The average accuracies of the proposed models are compared to state-of-the-art models [9], [17] in Table III.

The inclusion of different types of features in the model has improved the performance significantly. By combining the five best-performing individual features (HC, HM, DE, FBE and HMS), the average accuracy increased to 96.50% for HLV and 96.71% for HLA with the SVM classifier. The standard deviation simultaneously decreased.

TABLE II: Average accuracies for HLV and HLA prediction for the two best-performing features using SVM classifier with RBF kernel.

Feature	Feature extraction method	Average accuracy±STD
HLV		
Hjorth complexity	Original data	63.63% ± 6.26%
	Sub-bands	60.85% ± 6.10%
	Baseline removal	77.91% ± 5.83%
	Sub-bands + baseline removal	88.63% ± 3.92%
Hjorth mobility	Original data	65.60% ± 5.68%
	Sub-bands	64.01% ± 5.53%
	Baseline removal	73.85% ± 6.67%
	Sub-bands + baseline removal	88.38% ± 4.70%
HLA		
Hjorth complexity	Original data	67.25% ± 7.99%
	Sub-bands	65.60% ± 8.88%
	Baseline removal	79.50% ± 6.26%
	Sub-bands + baseline removal	89.28% ± 4.36%
Hjorth mobility	Original data	68.15% ± 7.78%
	Sub-bands	66.71%± 8.36%
	Baseline removal	76.68% ± 6.08%
	Sub-bands + baseline removal	88.54% ± 5.88%

Using the same features combined with the KNN classifier also provided good predictions, but the classifier is outperformed by SVM both in terms of average accuracy and standard deviation. For the SVM model, participant nr. 1 achieved the best prediction accuracies with 99.21% for HLV and 99.04% for HLA. The model had the worst performance on participant nr. 11 with accuracies of 90.38% for HLV and 92.17% for HLA. Performing the t-test on achieved accuracies of the SVM and KNN models, $p < 0.05$ for both HLA and HLV is obtained. This implies that the difference in the model performances is statistically significant. The proposed SVM model is outperformed with less than a percentage point by the 4D-aNN model proposed in [17] but outperforms the 4D-CRNN model presented in [9].

B. SEED dataset

The data was split into a training set and a testing set, in addition to a validation set for multilayer perceptron (MLP). For SVM and KNN, the split is kept at 80% for training and 20% for testing, while for MLP the training, validation and testing split percentage is kept at 60%, 20%, and 20%, respectively.

TABLE III: Comparison of results for HLV and HLA prediction with deep learning models and the proposed machine learning model using 5-fold cross-validation approach on DEAP dataset.

Features	Methods	Average accuracy±STD	
		HLV	HLA
DE	4D-aNN [17]	96.90% ± 1.65%	97.39% ± 1.75%
DE	4D-CRNN [9]	94.22% ± 2.61%	94.58% ± 3.69%
HC,HM,DE, FBE,HMS	SVM	96.50% ± 2.10%	96.71% ± 2.06%
	KNN	94.63% ± 3.56%	94.70 % ± 3.66%

TABLE IV: Average accuracies for the best-performing features, i.e. Hjorth mobility and complexity, and a combination of the two, on SEED dataset

Feature	Method used	Average accuracy
Hjorth complexity	SVM	79.05% ± 7.38%
	KNN	67.44% ± 8.88%
	MLP-v1	84.24% ± 4.88%
	MLP-v2	86.45% ± 3.45%
	MLP-v3	83.95% ± 3.29%
Hjorth mobility	SVM	83.87% ± 8.75%
	KNN	75.56% ± 10.87%
	MLP-v1	83.65% ± 5.32%
	MLP-v2	86.89% ± 4.20%
	MLP-v3	83.95% ± 3.21%
Hjorth mobility, Hjorth complexity	SVM	78.85% ± 7.48%
	KNN	68.80% ± 8.68%
	MLP-v1	85.57% ± 4.44%
	MLP-v2	86.01% ± 3.29%
	MLP-v3	85.71% ± 3.18%

Table IV presents the average accuracies for the three best-performing features, namely HM, HC and a combination of HM and HC. This is a result of channel-wise feature extraction for each epoch. The average accuracies are the average of the performance accuracies across all 45 trials on 15 participants. The results obtained with KFD and HFD features are poor as compared to the Hjorth parameters and are therefore not included. The individual feature HM has shown the best average accuracy of 86.89% when MLP-v2 is used as a classifier. This is just three percentage points better than the classification with SVM. However, SVM has slightly outperformed MLP-v3 for the HM feature. For HC and the combinatorial features involving HM and HC, the performance is improved by 7.40% and 7.16%, respectively, when applying MLP-v2 instead of SVM. The improvement in accuracy can be explained by the increased complexity of the architecture. Compared to two even more complex architectures, 4D-aNN model [17] and 4D-CRNN [9], the results of MLP-v2 on HM are 9.36% and 7.85% lower, respectively. The t-test was performed on the accuracies

obtained with the SMV, KNN, and MLP-v2 models. The difference in achieved accuracy of the KNN and SVM models is statistically significant, and so is the difference in accuracy between the MLP-v2 and KNN. However, the difference in the performance of the SVM and MLP-v2 models is not significant with $p = 0.355 > 0.05$.

TABLE V: Comparison of the average accuracies for the best-performing feature HM and deep learning models on SEED dataset

Feature	Method used	Average accuracy
DE	4D-aNN [17]	96.25% ± 1.86%
DE	4D-CRNN [9]	94.74% ± 2.32%
Hjorth mobility	MLP-v2	86.89% ± 3.45%

Finally, Table VI shows prediction accuracies of the best-performing model for the DEAP dataset when applied to the SEED dataset and the other way around. The five-feature SVM model on the SEED dataset has a performance comparable to that of the MLP-v2 model using only HM. The performance of the one-feature MLP-v2 model on the DEAP dataset is somewhat worse than the five-feature SVM model, but would probably be improved by including more features.

The results presented above for SEED and DEAP datasets suggest that it is possible to create models using less complex machine-learning techniques that can achieve comparable performance to the more complex deep learning architecture reported in the state-of-the-art work.

TABLE VI: Average accuracies when the best-performing models on DEAP and SEED dataset are applied to each other.

Dataset	Feature	Method	Average accuracy
SEED	HC, HM, DE, FBE, HMS (sub-bands)	SVM	85.90% ± 6.19%
DEAP HLV	HM (sub-bands, baseline removal)	MLP-v2	91.19% ± 2.55%
DEAP HLA	HM (sub-bands, baseline removal)	MLP-v2	91.78% ± 3.33%

IV. Conclusion

In this work, an automated emotion recognition approach is proposed and tested over two different publically available datasets SEED and DEAP. Several features such as HM, HC, DE, FBE, and HMS are computed from the epochs of 1-second duration which are extracted from the EEG signals of SEED and DEAP datasets. The HC and HM features have achieved the best accuracy for DEAP and SEED datasets. Further, we also computed the features from the different brain rhythms and also tested the different combinations of features. It is seen that combining features has improved the classification performance for the DEAP dataset. However, for the SEED dataset, the classification performance with the combination of features is not improved compared to that of the individual feature. With the incorporation of baseline feature correction and brain rhythms, the classification performance on the DEAP dataset is further improved. The SVM classifier yielded the best accuracy of 96.5% and 96.71% for HLV and HLA classes. The best accuracy with the MLP classifier is found to be 86.89% with the HM feature on the SEED dataset. It is apparent that the implementation of sub-bands and baseline removal are two important steps in order to improve the performance of EEG-based emotion recognition. Even though deep learning models are a small step ahead of machine learning models, simpler machine learning models might be a viable option to decode emotions from EEG. In future work, channel selection methods will be explored to reduce the complexity of the model even further.

References

[1] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-Channel EEG through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.

[2] Y.-P. Lin, C.-H. Wang, T.-L. Wu, S.-K. Jeng, and J.-H. Chen, "EEG-based emotion recognition in music listening: a comparison of schemes for multiclass support vector machine," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 489–492.

[3] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 589–597, 2010.

[4] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: a survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.

[5] M. Murugappan, M. Rizon, R. Nagarajan, and S. Yaacob, "Inferring of human emotional states using multichannel EEG," *European Journal of Scientific Research*, vol. 48, pp. 281–299, 12 2010.

[6] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 81–84.

[7] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[8] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," in *Neural Information Processing*, L. Cheng, A. C. S. Leung, and S. Ozawa, Eds., 2018, pp. 433–443.

[9] F. Shen *et al.*, "EEG-based emotion recognition using 4D convolutional recurrent neural network," *Cognitive Neurodynamics*, vol. 14, no. 1, p. 815–828, 2020.

[10] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 2014, pp. 1–6.

[11] L.-Y. Tao and B.-L. Lu, "Emotion recognition under sleep deprivation using a multimodal residual LSTM network," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[12] L. Moctezuma, T. Abe, and M. Molinas, "Two-dimensional CNN-based distinction of human emotions from EEG channels selected by multi-objective evolutionary algorithm." *Scientific Reports*, vol. 12, no. 3522, 2022.

[13] S. Koelstra *et al.*, "DEAP: a database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[14] "SEED Dataset," https://bcmi.sjtu.edu.cn/home/seed/seed.html, ., accessed: 2022-08-30.

[15] "MNE-Feature module," https://mne.tools/mne-features/api.html, ., accessed: 2022-09-28.

[16] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," 10 2018.

[17] G. Xiao, M. Shi, M. Ye, B. Xu, Z. Chen, and Q. Ren, "4D attention-based neural network for EEG emotion recognition," *Cognitive Neurodynamics*, vol. 16, pp. 1–14, 2022.

NTNU

Norwegian University of
Science and Technology