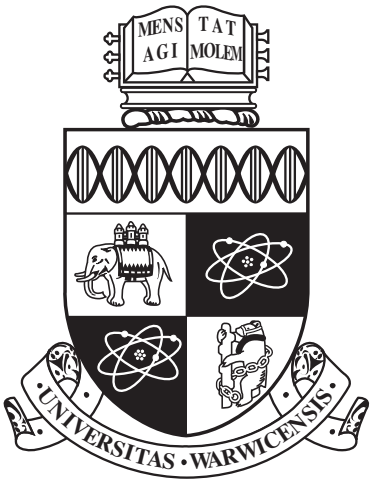


Signal processing and machine learning methods with applications in EEG-based emotion recognition

by

Laura Piho

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

School of Engineering

June 2019

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

First I would like to thank my supervisor Dr. Tardi Tjahjadi for his guidance and support. I would like to thank the Warwick School of Engineering for their three year scholarship that would allow me to undertake the Engineering PhD course.

Declarations

The author hereby declares, that the work presented in this thesis is entirely her own unless explicitly acknowledged, including citations of published and unpublished sources.

Parts of this thesis have been previously published by the author in the following:

[80] Laura Piho and Tardi Tjahjadi. A mutual information based adaptive windowing of informative eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 2018

The author confirms that the thesis has not been submitted for a degree at any another university.

List of Publications

- L. Piho and T. Tjahjadi, "A mutual information based adaptive windowing of Informative EEG for Emotion Recognition," IEEE Trans. Affective Comput., in press, 13 pages, DOI: 10.1109/TAFFC.2018.2840973.
- L. Piho and T. Tjahjadi, "Subject-dependent and subject-independent EEG-based emotion recognition using Gaussian process classification" *In progress*

Abstract

Automatic emotion recognition has become increasingly popular, with applications in marketing, advertising, e-learning, entertainment, and more. Currently, the majority of automated emotion recognition is performed using facial expressions, body language, and speech intonation patterns. In recent years, using brain signals has become increasing popular. Being able to understand and analyse brain signals is beneficial in many applications. The goal of this thesis is to develop an effective method for extracting and representing EEG signals associated with human emotions, and to develop a robust classifier using machine learning tools for emotion recognition.

The thesis aims to address the common problems related to the EEG-based emotion recognition datasets, including dealing with small sample sizes, low signal-to-noise-ratio and high dimensional data. The contributions of this thesis lie in the proposed subject-dependent and subject-independent EEG-based emotion recognition frameworks. These frameworks are shown to accurately perform two-class classification as well as multi-class classification. In addition, a novel mutual information based signal reduction algorithm is introduced, aiming to increase the accuracy of EEG-based emotion recognition when the duration of the recording due to chosen stimuli is long. Furthermore, Gaussian Process classification is introduced for the purpose of EEG-based emotion recognition. This classifier is combined with the subject-dependent and subject-independent emotion recognition schemes and is shown to increase the accuracy when compaerd to the previous commonly used clasifiers.

By using publicly available EEG datasets, the proposed novel frameworks are evaluated and shown to improve the EEG-based emotion recognition when compared against state-of-the-art methods. In addition, different signal processing methods suitable for EEG-based emotion recognition are introduced, explored, and analysed. An in-depth comparison of different feature extraction, feature selection, and classification methods is given using the proposed subject-dependent and subject-independent emotion recognition schemes.

Sponsorships and Grants

This work was supported by the Warwick University, School of Engineering scholarship.

Acronyms

ACF	Auto Correlation Function
BCI	Brain-Computer Interface
BSS	Blind Source Separation
CSP	Common Spatial Patterns
CWT	Continuous Wavelet Transform
DTFT	Discrete Time Fourier Transform
DWT	Discrete Wavelet Transform
ECG	Electrocardiography
EEG	Electroencephalography
EMD	Empirical Mode Decomposition
EMDB	Emotional Movie Data Base
EMG	Electromyography
EOG	Electrooculography
fMRI	Functional Magnetic Resonance Imaging
GA	Genetic Algorithm
GP	Gaussian Process
HCI	Human Computer Interaction
HHS	Hilbert-Huang Spectrum
HOC	Higher Order Crossing
HOS	Higher Order Spectral

IAPS . . . . . International Affective Picture System

ICA . . . . . Independent Component Analysis

IMF . . . . . Intrinsic Mode Functions

kNN . . . . . k-Nearest Neighbours

LDA . . . . . Linear Discriminant Analysis

MEG . . . . . Magnetoencephalography

mRMR . . . . . Minimum Redundancy Maximum Relevance

NB . . . . . Naive Bayes

NN . . . . . Neural Networks

PCA . . . . . Principal Component Analysis

PSD . . . . . Power Spectral Density

PSE . . . . . Power Spectral Entropy

ROC . . . . . Receiver Operating Characteristic

SF . . . . . Statistical Features

SL . . . . . Surface Laplacian

SVD . . . . . Singular Value Decomposition

SVM . . . . . Support Vector Machine

VA . . . . . Valence-Arousal

VDA . . . . . Valence-Dominance-Arousal

WT . . . . . Wavelet Transform

Chapter 1

Introduction

Emotions play an important role in everyday life, with an important part of the communication being the expression and recognition of emotions. The goal of the automatic emotion recognition is to accurately classify temporal emotion states given some input. Being able to extract and understand emotions has a significant benefit in human-computer interactions (HCI), where these benefits can be incorporated into telecommunications, video games, automobile safety, and educational software [106].

The majority of emotion recognition is achieved through facial expressions (often through still images or videos), body language, and speech patterns. In recent years, emotion recognition through brain signals has become more popular. Growing advancements in technology and development of human-centric (and human-driven) interactions with digital media have increased the need and significance of automated emotion recognition [59].

The human brain has fascinated physicians and scientists for hundreds of years. In the last century extensive discoveries and innovations have been made in the areas of science and technology that have now reached a point where direct interaction with the human brain is possible. This has lead to brain-computer interfaces (BCIs) which allow communication between users (human) and computers (machines) that do not depend on the brain's normal output pathways of peripheral nerves and muscles [99].

In the past few decades, brain signals have been utilised in a wider area of applications including communication, entertainment and gaming (e.g., virtual reality), lie detection, trust assessment, brain fingerprinting, neurorehabilitation, and medicine (e.g., sleep-stage or mood monitoring, diagnosing conditions like autism, narcolepsy, epilepsy, and alzhimers). Furthermore brain signals can be used to assess cognitive-state (including workload, fatigue, and alertness) of pilots and others

in high-risk employments [29, 60, 99]. Therefore, with the rising importance of applications which use brain signals, there is an increasing need for robust and accurate methods for signal processing and analysis.

Emotion recognition data can be collected using for example electroencephalogram (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) equipment. Measuring brain activity with EEG (or MEG or fMRI) is non-invasive and traditionally used in laboratory environment for medical purposes. The growing research in BCIs has expanded the possible applications of EEG, MEG and fMRI signals (images).

The EEG-based emotion recognition can be split into two parts: data collection, and data processing. In emotion recognition using brain signals, the first task deals with recording of the data (signals). In addition to recording the signals, it also covers responsibilities like finding a test group, choosing appropriate stimuli, and setting up recording equipment. The second task, data processing, concentrates entirely on handling the recorded information. The data processing involves preprocessing, extracting information, and creating a recognition framework that is able to classify the signals.

Emotions are thought to be subject-dependent. In most emotion classification based on image and speech, this matters less as there are a number of similarities between how people display emotions. At the brain signal level, these similarities are not as clear and often emotions are modelled subject-dependently. This means that the datasets used to model subject-dependent emotions are usually small (or very small). Despite this, high accuracy can still be reached for subject-dependent emotion recognition using brain signals.

Even though emotions are thought to be subject-dependent, there exists a need for subject-independent emotion recognition. Being able to find underlying similarities between how people process emotions on brain signal level would be a great step forward in the area of emotion recognition.

The signals extracted using EEG, MEG, or other methods are called raw signals, which often includes some artefacts. These artefacts can be movement related potentials, eye blinks, and facial movement. Furthermore, biomedical signals (e.g., electrocardiogram (ECG), electromyography (EMG), and electrooculography (EOG)) can be mixed in the raw brain signals. The biomedical signals are difficult to separate from brain signals as they resemble the actual brain signals [40].

The concentration of this thesis is on EEG signal processing, with the aim of achieving EEG-based emotion recognition. To create a reliable and robust emotion classification using EEG signals, different mathematical and statistical signal processing and machine learning methods can be used for pre- and post-processing

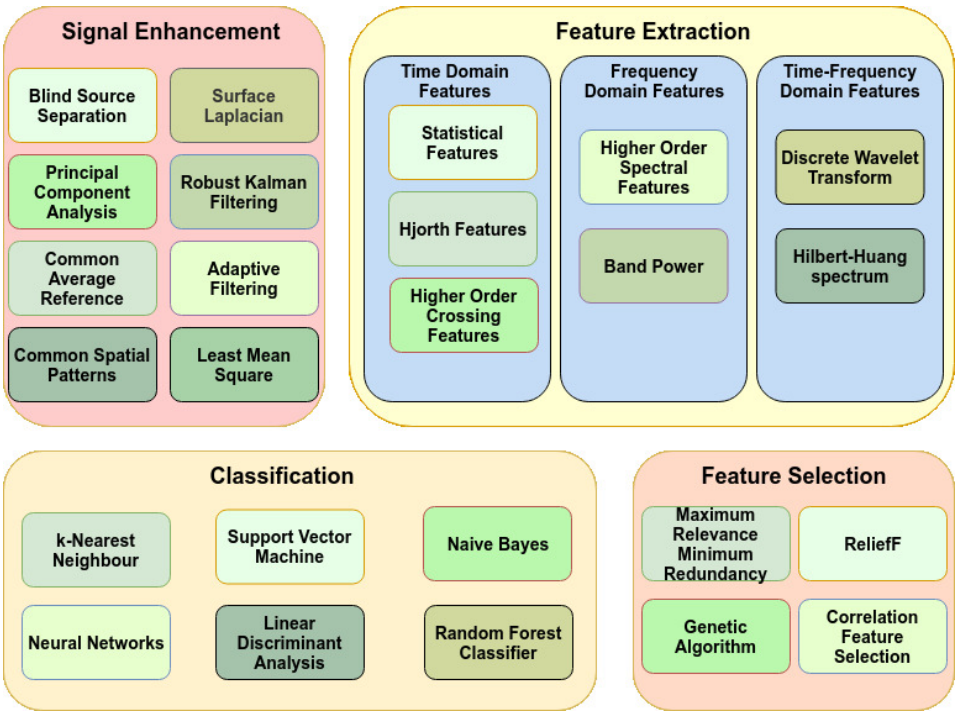


Figure 1.1: Different methods that can be used for EEG-based emotion recognition.

of the signals. In general, the EEG-based emotion recognition framework has four steps: signal pre-processing, feature extraction, feature selection, and classification. There exists a wide variety of methods that can be used for each of these steps. A selection of different methods is given on Figure 1.1.

Due to high levels of noise in the EEG data, the first step is to enhance the signal. The methods that can be used for EEG signal pre-processing include independent component analysis (ICA), common average reference (CAR), principal component analysis (PCA), common spatial patterns (CSP), surface Laplacian (SL), robust Kalman filtering, adaptive filtering, etc.

After pre-processing the EEG signals, features can be extracted and selected from the signals for classification. The feature extraction methods include time-domain, frequency-domain, and time-frequency domain features. The examples of time-domain features include statistical features (e.g., mean, variance, first and second order difference), Hjorton features, and higher order crossing features (HOC). Frequency-domain features include band power and higher order spectrum (HOS), and time-frequency domain features including Hilbert-Huang spectrum (HHS) features. However, most of the feature extraction methods give rise to a large number of features and consequently a high-dimensional problem. Therefore feature selection techniques are used to reduce the dimensionality. These methods include ReliefF,

maximum relevance minimum redundancy (mRMR), and genetic algorithm (GA).

Finally, the reduced set of features will be used for classification. Common methods of emotion classification using brain signals include support vector machines (SVM), k-nearest neighbours (kNN), naive Bayes (NB), and neural networks (NN). Furthermore, recently transfer learning has been used for EEG-based emotion recognition.

The EEG-based emotion recognition datasets are often small, multidimensional, and noisy. This creates major challenges in the signal processing. With most noise removal techniques there is no good way to identify which part of the signal is related to emotions. One has to be careful not to remove emotion related information from the signals. Unlike emotion recognition using images or voice, there is no clear way to identify the emotions from EEG signals. In this thesis, the goal is to address the problems that arise from these datasets. Using different signal processing methods, it is shown that accurate recognition can be made using the available small datasets.

1.1 Contributions

The aim of this thesis is to develop an effective method for extracting and representing EEG signals associated with human emotions, and to develop a robust classifier using machine learning tools for emotion recognition. The contributions made in this thesis are:

- A novel mutual information based signal reduction algorithm is introduced.
- The Gaussian Process classification is introduced for the purpose of EEG emotion recognition, and a novel subject-independent emotion recognition scheme using EEG signals is proposed using Gaussian Process classification.
- Using publicly available EEG datasets the proposed novel methods are evaluated and shown to improve the EEG-based emotion recognition when compared to state of the art models for both two-class classification and multi-class classification.

1.2 Associated publications

The content of this thesis is based on work done during the doctorate studies, where the following articles are the outcome of this work:

1. L. Piho and T. Tjahjadi, "A mutual information based adaptive windowing of Informative EEG for Emotion Recognition," IEEE Trans. Affective Comput., in press, 13 pages, DOI: 10.1109/TAFFC.2018.2840973.

2. L. Piho and T. Tjahjadi, "Subject-dependent and subject-independent EEG-based emotion recognition using Gaussian process classification" *In progress*

1.3 Outline of the thesis

This thesis is presented as follows. The pre-requisites of affective computing are discussed in Chapter 2. The chapter begins with giving an overview of recording brain signals, creating datasets, and ethical concerns related to EEG-based emotion recognition. Next, different pre-processing, feature extraction, feature selection, and classification methods used for EEG-based emotion recognition are given. Finally, the EEG datasets used in this thesis are introduced, and state-of-the-art results presented.

Chapter 3 introduces a mutual information based algorithm for signal reduction. It is shown that the signal reduction improves the classification accuracy. Examples of subject-dependent emotion recognition are given using three different EEG datasets. Furthermore, the chapter gives an in-depth comparison between different feature extraction and classification methods.

In Chapter 4, the primary focus is on Gaussian Process classification. The method is introduced, applied, and analysed for the purpose of EEG-based emotion recognition. The examples for both subject-dependent and subject-independent emotion recognition are given.

Finally, in Chapter 5, the discussion and analysis of the work done, future work, and the conclusion is given.

Chapter 2

Affective computing

In [79], affective computing is defined as computing that relates to, arises from, or deliberately influences emotion or other affective phenomena. It is an interdisciplinary field of research combining engineering with mathematics, and computer science with cognitive and neuroscience. In addition it is strongly linked to psychology and sociology, especially when dealing with ethical questions that affective computing gives rise to.

The aim of affective computing is to develop advanced methods to recognise affective-cognitive states. This can lead to possible assessment and communication of affective-cognitive states, more emotionally intelligent technology, and understanding how emotional responses impact health and vice versa. This chapter aims to give an in-depth overview of EEG-based emotion recognition, with some mention of other methods used to recognise emotions (e.g., facial image and speech recognition).

The first part of this chapter will discuss the relations between emotions and the human brain. The focus is on different methods for measuring brain activity, especially for the purpose of affective computing. It is discussed how to create different datasets for emotion recognition, and how to choose stimuli. In addition, different models that can be used to label the emotions are discussed.

Following this, the overview of the existing methods used in this thesis is given. Different pre-processing topics, including signal downsampling, blind source separation, wavelet analysis and empirical mode decomposition are discussed, making use of both time and space domains. Multiple feature extraction and selection methods are introduced. Likewise, classification methods that are commonly used for EEG emotion recognition are discussed.

Next, the in-depth details of the three publicly available emotion recognition datasets used in this thesis to illustrate the proposed work are given. Furthermore, the results achieved using state-of-the-art methods are discussed, giving an overview

of different results, which are later used for comparison. Finally, a short summary of the chapter is given.

2.1 Human brain and emotions

The electrical signals emitted from muscle nerves were first registered in the 19th century [89]. The research into measuring brain signal on humans and following a more detailed study into application and analysis of these signals started in 1920s. Research into connections between emotions and human brain was one of the first aspects that sparked interest.

Together with increasing knowledge of brain activity, and advanced techniques in signal processing and computation, the number of studies published relating emotions to brain signals has increased in the recent decade. Furthermore, compared to other emotion recognition models (e.g., based on images and speech), using brain activity is advantageous as it is available all the time, and it is difficult to be manipulated via voluntary control.

Emotions have been linked to the area in the brain called the limbic system, where different parts of the limbic system have different tasks in processing emotion information. The most important part of the limbic system (regarding emotion recognition) is the amygdala, whose purpose is to learn to connect stimuli to emotional reactions and to evaluate new stimuli by comparing it to past experiences [9].

As the limbic system is located inside the brain, the activity cannot be directly measured from the scalp. The amygdala is connected to the temporal and prefrontal cortices, indicating that the brain activity recorded from there could be linked to emotions. Both [47] and [74] reported a positive correlation of valence linking to temporal sources. Furthermore, in [61] and [14] the signals recorded from the frontal cortex have also been shown to describe the changes in valence. However, this does not mean that one should only use temporal and frontal lobe electrodes to record brain signals for emotion recognition. On the contrary, in [61], it has been shown that using more channels increases the emotion classification rate.

2.1.1 Measuring brain activity

In general, there are three common methods used to measure brain activity: electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG). All these methods are non-invasive to the participants, however they measure brain activity using different methods. The EEG records the electrical activity generated by neural firing, MEG captures the magnetic fields generated by neural activity, and fMRI measures changes in blood flow related

to brain activity [92].

For the purpose of emotion recognition, EEG systems are most commonly used for data collection. Although fMRI has a good spatial resolution, the temporal resolution is low. In addition, the action potential of the EEG signal takes approximately 0.5-130 milliseconds to propagate across single neutron, while for fMRI it takes seconds [88]. Alternatively, MEG systems have good spatial and temporal resolution. However, both MEG and fMRI are very costly. Furthermore, both can only be used in laboratory setting. They are sensitive to head and body movement and not portable. In addition, to use MEG and fMRI one needs to go through training to get specialised license to operate the equipment.

EEG recording devices are cheap and can be used with minimal training. Furthermore, in addition to the laboratory EEG equipment, there exist portable EEG devices. In addition, the EEG is not limited to movement, making it possible to use outside laboratory with flexible stimulus.

There are some challenges using EEG signals for emotion recognition. First, the EEG signals are very noisy and there is no definitive way to separate the signals related to the stimulus (emotion states) to other activities taking part at the same time. Therefore, even though EEG devices are not limited to movement, the experiments tend to take place in controlled laboratory environment to minimise the signals related to other activities. Furthermore, due to variations in emotional response to the stimuli, it can be difficulty to differentiate between the signals corresponding to different emotional states.

Second, gathering data for EEG-based emotion recognition takes time, especially when the stimuli chosen is audio-visual (e.g. videos). Therefore, all the existing datasets are relatively small, resulting in limited number of emotions being recognised.

The examples given in this thesis use signals recorded on EEG systems. Two of the publicly available example datasets were created using traditional non-portable laboratory EEG recording device. The third dataset was created using commercially available Emotiv EPOC wireless EEG headset[1] to record the signals.

2.1.2 Creating datasets and choosing stimulus

Due to the nature of emotions, most research on emotion recognition is done using experimental data. This data can include facial expressions from static images or video sequences [4, 63, 82], speech (audio) [31, 67], text [76, 90], brain signals [1, 6], and sometimes a combination of these modalities [38, 41]. In general, to

[1]The information on the wireless EEG headsets can be found on www.emotiv.com and was used for recording one of the datasets used in this thesis DREAMER [39]

train good robust classifiers, the number of training samples has to be large and the sample set diverse. Therefore, in some cases, emotion recognition is performed using information not specifically created for emotion recognition, e.g., movie clips and audio recordings from movies. However, a more common approach is to use datasets created specifically for emotion recognition. These datastes are constructed by creating an emotional response using some kind of stimuli.

There exists a large number of datasets for facial emotion recognition. These datasets have still facial images and (or) facial videos sequences of participants displaying certain emotions. For facial emotion recognition, the stimuli used to evoke emotions are often audio-visual (i.e., movie clips) [13, 47, 94], however still images [52], laboratory based emotion inducing tasks [93] or posed emotions [30] have also been used.

The datasets for speech-based emotion recognition can be divided into three classes: imitated, induced, and natural. The simulated datasets (e.g., [51]) are most commonly used and often have recordings of people with some acting experience. The recordings are of linguistically neutral content but spoken to express different emotions [50]. There exists a wide variety of simulated datasets (in multiple languages) for multiple different emotions, and these datasets can usually be easily standardised, however they do not always give an accurate representation of real world emotions.

Induced speech recognition datasets represent real world emotions more accurately than simulated datasets. These datasets are created by recording a conversation between a participant and an anchor, where the latter is leading the conversation to evoke an emotional response [44]. However, the number of emotions present could be limited.

In real life, the emotion may not be as dominant in the speech as simulated and induced emotion datasets and can be harder to recognise. There exist some datasets that use recording from natural settings (e.g., call centers, cockpit) which can be used for emotion recognition. Similarly to induced emotion datasets the natural data may not include all the emotions. Furthermore, natural databases do not include emotion labels (or true emotion labels) to compare the classification results against.

The emotion recognition using brain waves (signals) differs considerably from the previous methods. Recognising emotions by looking and hearing someone comes naturally for most people. However, in a natural situation, the brain signals cannot be interpereted by others. Most of the brain signal datasets use EEG recordings [47, 94]. This is because EEG systems are cheaper and easier to operate than other brain signal recording devices. However, MEG recordings [2] and fMRI [42] have been used to recognise emotions and locate areas in the brain that are related to

emotions.

The stimuli used for emotion recognition from brain signals are usually audio (e.g., music) [12], visual (e.g., still images) [55], or audio-visual (e.g., movie clips, music videos) [13, 47, 94]. For most brain signal recordings participants need to be still. In theory, with portable EEG devices, brain signal recordings for emotion recognition could be performed in a natural setting.

There has also been some work done in classifying emotions using multiple different modalities, for example [7] use both EEG signals and facial expressions, and [16, 18] use both speech and facial videos. In addition, emotion recognition can be achieved from text. These datasets tend to consist of sentences (and/or short paragraphs) which have been annotated with labels (sometimes by third party).

Choosing the appropriate stimuli to evoke emotions can be complicated as the reaction to stimuli is very subject dependent. For example, one might find images of cats pleasing (relate to happy emotion) where as another might find it indifferent (relate to neutral or base line) or even dislike it. For brain signal based emotion recognition the choice of stimuli is also limited to audio, visual, and audio-visual stimuli due to the nature of the recordings.

In general, video clips tend to be better in evoking emotions. In [13] an emotional movie database (EMDB) includes a set of tested non-auditory clips which can be used for emotion recognition together with labels (using VDA scale) for the videos. These labels have been given by investigating 131 participants' self-assessment [13]. Using video clips is useful when one wishes to have emotions recorded over a longer time frame.

A longer time frame often means less samples in an EEG-based emotion recognition dataset. When shorter time frame is preferable, often images are used as stimuli. An example of image-based emotion dataset with VDA labels included is the International Affective Picture System (IAPS) [55].

2.1.3 Labeling emotions

Most of the emotion recognition using EEG signals is achieved using supervised learning algorithms. For supervised learning, the datasets have to include labels, or one has to be able to annotate the data to include labels. Depending on the data used, a label can be either the root of stimulus (or stimulus itself) or response to the stimulus. An example of emotion being a root of the label in facial emotion recognition would be letting a participant to create a facial expression according to the label. Showing a participant different images and recording their facial expressions and annotating the response (by third party or by participant themselves) would be an example of the label being the response to the stimulus.

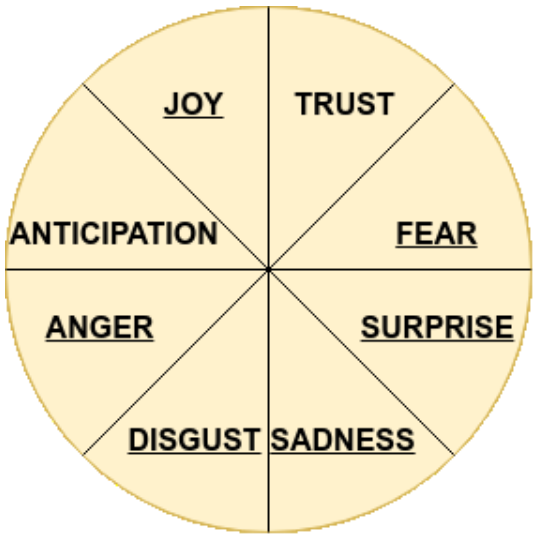


Figure 2.1: The eight primary emotions proposed by Robert Plutchik. The six underlined emotions also correspond to the basic emotions proposed by Paul Ekman.

There does not exist a general standard for labelling emotions for EEG emotion recognition. Two more common ways to describe emotions for BCI applications are either using Paul Ekman's basic emotions [22] or Valence-Arousal (VA) scale [87].

In his research, Ekman highlighted six basic emotions: anger, happiness, sadness, disgust, fear, and surprise. A lot of the emotion recognition databases are designed around these so called basic emotions. However, when using Ekman's approach to emotions, one is limited to six emotions. In his later works, Ekman himself theorised more universal emotions existing.

Agreeing with Ekman's approach, Robert Plutchik created a scale of emotions called Plutchick's wheel of emotions. It starts out with eight primary emotions, i.e., Ekman's six emotions plus anticipation and trust, to create a positive-negative view, see Figure 2.1. The idea behind the wheel is to enable the creation of a whole range of emotions by combining the primary emotions, e.g., combining anger and disgust results in contempt.

An alternative to labeling emotions is the valence-arousal model (or circumplex model) proposed by Russell in 1980, Figure 2.2. This model is a more common way to label the emotions for EEG-based emotion recognition. It is a two-dimensional model that on the x-axes highlights if the emotion is positive or negative, whereas the y-axis shows the activeness or passiveness of the emotions. This can give more flexibility when labelling emotions.

There is no right or wrong way to label the emotions. There are two ways to label data: prior stimulus affect labelling and post stimulus affect labelling. With prior stimulus affect labelling, the subjects (participants) are often aware of the

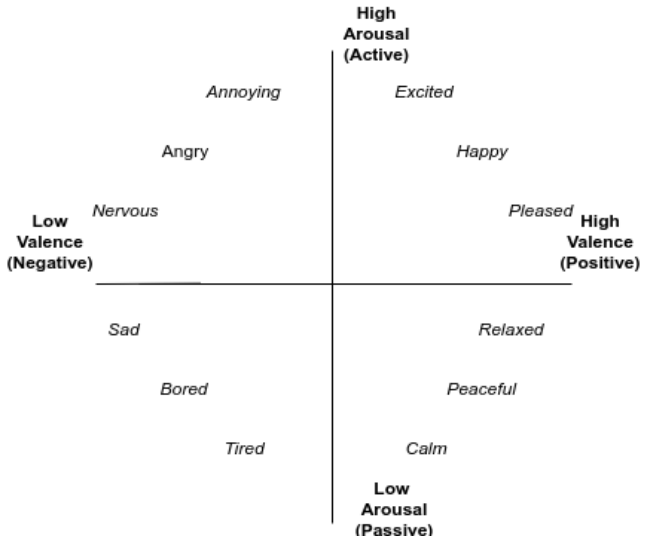


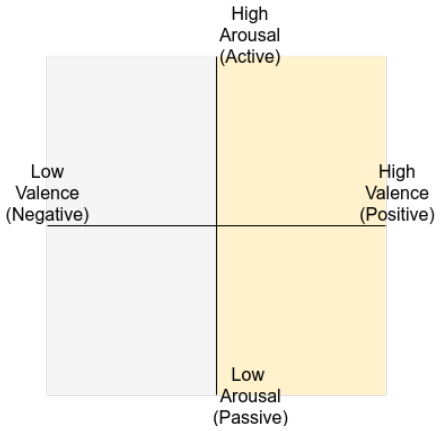
Figure 2.2: The Valence-Arousal scale proposed by James Russell.

emotions they need to present. This is often the case when classifying emotions using images, where the dataset can be created by asking participants to for example look happy or sad depending what emotions are needed. In this case, either Ekman's and/or Russell's models are often chosen.

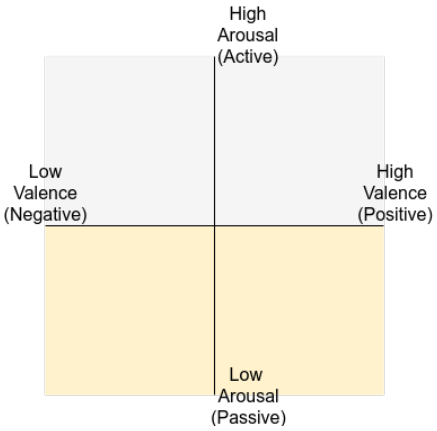
The post stimulus affect labelling works by first evoking an emotion, and then labelling it. This can be achieved by either self-assessment or by third-party. In facial emotion recognition, labelling by third party is not uncommon. However, in EEG-based emotion recognition most all labeling is done via self-assessment. Furthermore, in EEG-based emotion recognition, post stimulus affect labelling is more often used. Due to the nature of EEG-based emotion recognition, the self-assessment of the signals is the most practical way to create datasets.

In addition, as the EEG datasets are generally small, the Valence-arousal using Valence-Arousal scale, the same set of trials (EEG recordings) can be used for two-class classification and for multiclass classification. All examples in this thesis use datasets that label the trials using valence and arousal model. Furthermore, the classification is done for valence and arousal separately in all examples. The labelling of the data is achieved by splitting the trials according to the valence and arousal scores given in self assessment.

To get the labels from the self assessment, the valence and arousal scores are normalised between -1 and 1. For two-class classification, the positive valence (arousal) has scores > 0 and the negative valence (arousal) has scores < 0. Similarly, the three class classification labels are created by letting the positive class have scores between 1 and $\frac{1}{3}$, neutral class have scores between $\frac{1}{3}$ and $-\frac{1}{3}$, and negative

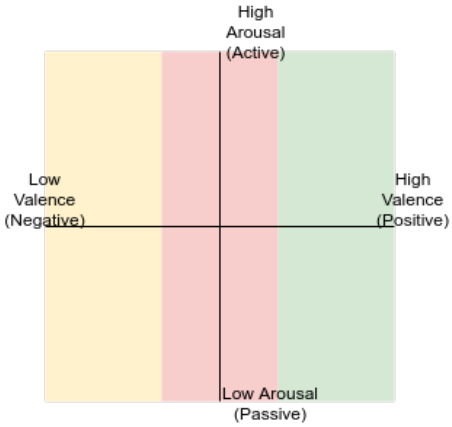


(a) The two-class valence labels.

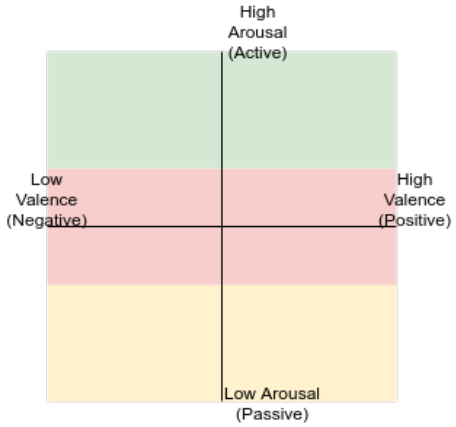


(b) The two-class arousal labels.

Figure 2.3: Valence-Arousal labels for two-class classification. For both two-class valence and arousal, the positive class is highlighted in yellow and the negative class in blue.



(a) The three-class valence labels.



(b) The three-class arousal labels.

Figure 2.4: Valence-Arousal labels for three-class classification. For both three-class valence and arousal, the positive class is highlighted in yellow, the neutral class is highlighted in pink, and negative class in blue.

class have scores between $-\frac{1}{3}$ and -1. Finally, the five class classification labels are created by letting the positive class have scores between 1 and 0.6, semi-positive class to have scores between 0.6 and 0.2, neutral class have scores between 0.2 and -0.2, semi-negative class between -0.2 and -0.6, and negative class have scores between -0.6 and -1. Figures 2.3 and 2.4 illustrate the two and three-class classification labels using Valence-Arousal model.

2.1.4 Ethical concerns

Even though the focus of the thesis is on signal processing and machine learning methods, it is essential to discuss some of the ethical concerns related to affect recognition. The ethics of different brain computer interfaces (BCIs) have been discussed in [32, 65, 107]. In this Section, a brief overview of ethical concerns concentrating on affective BCI[2] is given.

Emotions play a central role in human everyday life, and therefore it is essential to accompany the development of affective BCIs with ethical considerations as early as possible. First, the main concerns overlapping with other BCI systems ethics will be looked at. Things that needed to be addressed include:

- Risks to the human health;
- Security, privacy, and consent;
- Liability;
- Impact on identity, and personhood;
- Biases embedded in the device.

The main concern about risk to the human health is when using invasive forms of affective BCIs which are embedded in the brain. The benefits and risks have to be carefully considered due to the chance of serious harm (e.g., infection, brain tissue injury). However, for affect recognition, most often non-invasive methods are used (e.g., EEG). There are no risks associated with an EEG recording, and the test is painless and safe.

The affect recognition data collected is very sensitive, and the security and privacy of this data needs to be addressed. In [84], the main concerns about privacy asks questions like: Who will collect emotional data? What type of emotions are recognised? What task the recognised emotional data aims to fulfill? It was concluded that participants strongly favoured ethical contracts, when it came to affect recognition [84]. In general, the EEG-based emotion recognition datasets are anonymous and only available for research purposes.

With a functioning affective BCI systems, it is important for the users to understand the aim of the system, what it does and why? In addition, it is important to give a clear overview of what kind of data is collected and processed. Problems of shared control, criminal guilt and liability are important to consider, by considering the independence of the BCI device. One of the more importatnt questions in the

[2]Affective BCI is the sub-field of BCI, that aims to extract information related to affective states (e.g. emotions and moods).

literature is to consider whether an action done by a device (either solely or mostly) can truly be attributed to a human [26, 62, 96].

Finally, problems related to the biases in neural devices have been pointed out in [23, 96, 112]. These biases arise when scientific or technological decisions are based on a narrow set of systemic, structural or social concepts and norms. In the case of affective BCI (and emotion recognition), it is necessary to consider the potential biases regarding affective states [96].

In addition to the concerns discussed above that are universal to all BCI applications, there are some additional considerations directly related to affect recognition. It is worth mentioning that the ability to monitor affective states introduces a serious concern of mental privacy. The subject may not wish to share all mental states that the system is able to recognise (monitor). In addition, the ability to recognise (monitor) different emotions (states) can lead to the introduction to new stereotypes linked to the emotions or even the social pressure to self-regulate the emotions [96].

2.2 Methods

In this Chapter, the methods for EEG signal processing will be discussed. The methods will be considered in four sections, namely pre-processing, feature extraction, feature selection, and classification.

2.2.1 Pre-processing

The EEG signal data is often extremely noisy. The noise sources can be external, environmental, and physiological. A sample of raw EEG data from ten channels can be seen on Figure 2.5. The external and environmental sources, e.g., noise from equipment and electro-magnetic (EM) noise can often be dealt with by mitigation [83]. This can simply be done by insuring that the equipment is in good order and by removing EM sources from the recording room [83].

The main noise source that needs to be accounted for is physiological noise, i.e., cardiac signals (electrocardiogram, ECG), movements caused by muscle contraction (electromyogram, EMG), and ocular signal caused by eyeball movement (electrooculogram, EOG) [83]. Some of the physiological signals can be reduced by asking participants to find a comfortable position to sit which reduces noise caused by EMG [83]. The EOG signals are generated by eye saccades or movement of the eye as well as blinking [83]. The noise caused by eye movement can be reduced by using stimuli that does not require much eye movement. However, the noise due to blinking is harder to deal with, as blinking is considered as an involuntary

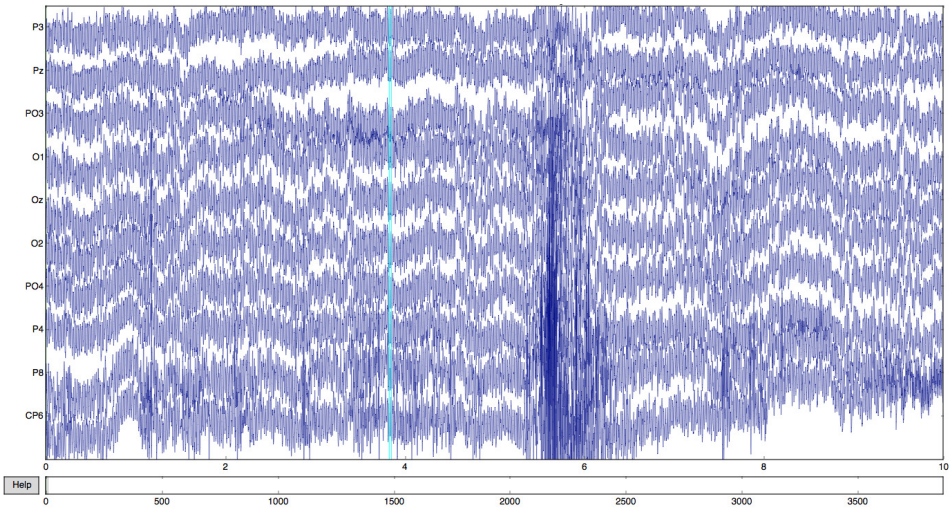


Figure 2.5: The raw EEG data from parietal (P3, PZ, P4, P8, PO3, PO4), occipital (O1, OZ, O2), and central (CP6) channels. A "Z" (zero) refers to an electrode placed on the midline sagittal plane of the skull, and "PO" and "CP" denote the intermediate electrode placements between parietal and occipital, and central and parietal respectively.

action. There are some options to minimise the noise due to blinking, e.g., by asking participants not to blink during the critical periods of the experiment using cues to inform them when they can blink freely. But withholding involuntary muscle movement can become a noise source.

Furthermore, to create datasets for EEG-based emotion recognition, often audio-visual (e.g., movie clips, music video clips) stimuli are used, as this tends to trigger a stronger emotional response than just audio stimuli. Therefore a compromise has to be found in trying to avoid ocular/eye movement noise, and having effective emotional stimuli.

The aim of pre-processing (signal enhancement) is to deal with noisy data by increasing the signal-to-noise ratio. Preprocessing comprises artefact and noise removal using digital signal processing techniques. There is a need to remove noise that has not been produced neurologically, e.g., blinking, vascular effects, muscular effects etc. In addition, there is noise from neurological sources that has to be removed.

There are a wide variety of methods used for signal pre-processing. Most often, spatial filtering methods are used. Furthermore, spatial filters like surface Laplacian and Wiener filtering can be used. In addition to spatial filters, regression methods can be used for artefact removal. The following sections present the different useful pre-processing methods.

Signal downsampling

EEG signals are often recorded using higher sampling rate than needed, hence for practical applications, the first step is to downsample the signal. The reason for downsampling is to reduce the memory requirements. When doing this, the main concern is in preserving the original information given by the recorded data. EEG signals contain a wide frequency spectrum. The ultraslow and ultrafast frequency components play no significant role in clinical EEG. In [73], it has been shown that the emotion-related signals lie between $3.5Hz$ and $40Hz$.

When choosing the sampling rate for downsampling, the Nyguist rate is often used. The Nyguist theorem states that the sampling rate should be at least twice the highest-frequency signal. Therefore, if the cut-off frequency of the low-pass filter is around $40Hz$, the sampling rate has to be higher than $80Hz$. Often, the emotion-related EEG is downsampled to $128Hz$, which satisfies the Nyquist theorem.

Blind source separation (BSS)

There exist multiple methods for removing ECG, EOG, and EMG artefacts. These methods are based on BSS and aim to separate the source signals from the mixed signal without any (or very little) information about the mixing process or the original source signals.

Consider a set of observed signals $x_i(t)$, $i = 1, ..., N$ and $t = 1, ..., M$, where M is the number of samples, and N is the number of sources (i.e., in the scope of this thesis EEG channels). These observations can be written in a matrix form, $\mathbf{X} \in \mathrm{R}^{M \times N}$.

Furthermore, consider a set of source signals $z_i(t)$, $i = 1, ..., N$ and $t = 1, ..., M$, where M is the number of samples, and N is the number of sources. Similarly to observed signals, the source signals can be written in a matrix form $\mathbf{Z} \in \mathrm{R}^{M \times N}$. The matrix $\mathbf{Z}$ is an unknown. In BSS, the observed signals and source signals are related to each other by a mixing matrix $\mathbf{A} \in \mathrm{R}^{N \times N}$, such that $\mathbf{X}^T = \mathbf{A}\mathbf{Z}^T$.

The assumption of mutual independence between signals is the key idea behind BSS. Based on this assumption, the *demixing* matrix can be found using principal component analysis (PCA) and independent component analysis (ICA). Both of these techniques aim to project the data onto a new basis that fulfill some statistical criterion, which removes correlation. For PCA, second order moment (variance) is used to find the new basis, whereas for ICA other measures of independence are used.

The aim of the BSS is to find a *demixing* matrix $\mathbf{B}$, such that $\hat{\mathbf{Z}}^T = \mathbf{B}\mathbf{X}^T$, and $\hat{\mathbf{Z}} \approx \mathbf{Z}$. PCA uses the second order moments to find the recovered approximations

called component vectors $\hat{\mathbf{z}}_i(t)$. This means, the component vectors are found such that they explain the maximum amount of variance possible by N linear transformed components. This leads to a set of orthogonal axes, the dot product of the axes and cross-correlation of the projections are both close to zero, resulting in orthogonal axes.

The principal components of the multidimentional signal can be found by using singular value decomposition (SVD). The matrix of observations can be decomposed as:

$$\mathbf{X} = \mathbf{USV}^T \tag{2.1}$$

where the matrix $\mathbf{S}$ of size $M \times N$ is a non-square matrix with zero entries everywhere except on the lead diagonal, i.e.,

$$\mathbf{S} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_N} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}. \tag{2.2}$$

The other two matrices $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal square matrices of size $M \times M$ and $N \times N$ respectively[3]. The columns of matrix $\mathbf{U}$ are orthonormal eigenvectors of $\mathbf{XX}^T$, and $\mathbf{V}$ is a matrix whose columns are the orthonormal eigenvectors of $\mathbf{X}^T\mathbf{X}$. Furthermore, the non-zero elements of the matrix S are the square roots of the eigenvalues from U or V in descending order.

In order to find $\mathbf{U}$, first one has to calculate the eigenvalues and eigenvectors of $\mathbf{XX}^T$. These eigenvectors become column vectors in a matrix, such that the eigenvector with the largest eigenvalue is in column one, following the eigenvector of the next largest eigenvalue.The eigenvector with the smallest eigenvalue is positioned in the last column. Following, to get the matrix $\mathbf{U}$, this matrix has to be converted into an orthonormal matrix. This can be done by applying the Gram-Schmidt (orthonormalization) process[4] to the column vectors. The matrix $\mathbf{V}$ is found similarly to $\mathbf{U}$, by calculating the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$, and converting the resulting matrix into an orthonormal matrix by applying the Gram-Schmidt process.

[3]This means $\mathbf{UU}^T = I$ and $\mathbf{VV}^T = I$

[4]The Gram-Schmidt process is defined as: Let $B = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ be a basis for an inner product space V. Let $B' = \{\mathbf{u}_1, ..., \mathbf{u}_n\}$, where $\mathbf{u}_1 = \frac{\mathbf{v}_1}{||\mathbf{v}_1||}$, $\mathbf{u}_{k+1} = \frac{\mathbf{v}_k - \sum_{j=1}^{K} <\mathbf{v}_{k+1}, \mathbf{u}_j> \mathbf{u}_j}{\sum_{j=1}^{K} <\mathbf{v}_{k+1}, \mathbf{u}_j> \mathbf{u}_j}$. Then B' is an orthonormal basis for V. The $< \cdot, \cdot >$ is the inner product and $|| \cdot ||$ is the norm in Euclidean space.

Another application of PCA is dimensionality reduction, PCA can be used to reduce the dimension of the data from D to p. This is done by assuming that the data contained in the last $D - p$ components are mostly noise.

However, for the purpose of EEG signal pre-processing, ICA is more often used. As the orthogonality implies independence, but independence does not necessarily imply orthogonality, the form of independence imposed by PCA is weaker than one imposed by ICA. ICA refers to a variety of techniques which aim to uncover the independent source signals from a set of observations that are composed of linear mixtures of the underlying sources [35]. The fundamental idea of ICA is to apply operations to the observed data or to the mixing matrix and measure the independence of the sources. In PCA, the measure for independence is variance. However, for ICA, the choice of measures is greater. More common choices for the cost function are mutual information, entropy, and kurtosis[5]. Iterative methods are used to maximise (or minimise depending on the measure) the cost function.

There exists multiple methods to implement BSS using ICA. These methods include *FastICA* [28], *EFICA* [48], *MULTICOMBI* [101], *FCOMBI*, and *iWASOBI* [102]. The *FastICA* and *EFICA* aim to find maximally non-Gaussian elements. However, the *EFICA* algorithm is asymptotically efficient version of *FastICA*[6]. The *MULTICOMBI* and *FCOMBI* are similar in the sense that the latter is an efficient version of the first. The advantage of using these lies in being able to simultaneously separate non-Gaussian and time-correlated sources [27]. Although the *FCOMBI* is computationally more efficient, it is not as stable and reliable as *MULTICOMBI*.

Finally, the algorithm used in this thesis to find the mixing matrix is the *iWASOBI* [102] algorithm. The *iWASOBI* algorithm is based on the *SOBI* [8] and *WASOBI* [111] algorithm which use second order statistics to find spatial components. The *SOBI* algorithm uses approximate joint diagonalization to find the mixing matrix $\mathbf{A}$. The *WASOBI* algorithm is an adaptation of *SOBI*, where in *WASOBI* algorithm joint diagonalization is transformed into a properly weighted nonlinear least squares problem. The *iWASOBI* has two main advantages: enhanced running speed, important especially in high-dimensional problems; and capability to use the specially structured weight-matrices with approximate joint diagonalisation (AJD) criterion.

Common average reference (CAR)

The EEG signals (measured in voltages) recorded at each (specific) electrode are relative to the signals recorded at all other electrodes. Hence, the reference could lie

[5]Kurtosis is the fourth order moment, which is a measure of non-Gaussianity.

[6]The *EFICA* estimator can asymptotically reach the CrameèrRao lower bound.

anywhere, however for signal pre-processing the reference has to be carefully chosen as any activity in the reference electrode will be reflected in the activity at other electrodes.

The common practice is to choose the position of a reference electrode to be away from the expected main effects. Therefore depending on the activity related to the EEG signals the reference point can differ. A commonly used reference electrode is 'Cz', located on the top of the head. However, this node should not be used when the brain activity related to the task is expected to have activity near its location. In addition it is not suggested to select a reference the data around an electrode of one hemisphere, as this can add a laterality bias into the signal data. Often, instead of referencing the data around a specific electrode, a common average reference (CAR) is used. This method is based on the assumption that the mean of all recording channels is approximately neutral.

To reference the data to a CAR, consider a signal data matrix $\mathbf{Z} \in \mathrm{R}^{N \times M}$, where N is the number of signals (electrodes) and M is the number of samples. Denote the re-referenced signal $\mathbf{Z}_r \in \mathrm{R}^{N \times M}$, i.e.,

$$\mathbf{Z}_r = \mathbf{A}\mathbf{Z}, \tag{2.3}$$

$$\mathbf{A} = \mathbf{I} - \begin{bmatrix} 1/N & \dots & 1/N \\ \vdots & \ddots & \vdots \\ 1/N & \dots & 1/N \end{bmatrix} \tag{2.4}$$

where $\mathbf{I} \in \mathrm{R}^{N \times N}$ is the identity matrix, and therefore $\mathbf{A} \in \mathrm{R}^{N \times N}$.

When using CAR for EEG signals, the assumption of mean of the recording channels being approximately neutral is only valid with full coverage of the head surface and accurate spatial sampling. This however requires a substantial number of electrodes. The consequence of this condition not holding is that CAR effect will exist and bias the recorded signals [19]. For EEG emotion recognition, CAR is often used even with 32 electrodes, as the recordings do cover all of the heads surface. In addition, the useful signals related to emotions can occur in all electrode locations and therefore using a specific electrode as a reference can have a drastic biasing effect.

Wavelet analysis

In addition to manipulating the EEG data in the spatial domain, i.e., projecting the data to a new set bases (the process of BSS), one can process the data in the frequency domain. The wavelet transform is a notably effective tool for time-frequency analysis.

Table 2.1: Different EEG frequency bands and their biological meanings in healthy adults.

Frequency	**Frequency Band**	**Normally Occurring**
Gamma	32 Hz & up	Concentration, Problem solving
Beta	16-32 Hz	Busy, Active, Anxiety dominant, external attention, relaxed
Alpha	8-16 Hz	Relaxed, passive attention, restful, reflective
Theta	4-8 Hz	Deeply relaxed, inward focused, tiredness, Drowsiness
Delta	0-4 Hz	Sleep, Dreaming

It can represent elements of non-stationary signals such as trends, discontinuities, and patterns, even when other signal processing tools are unsuccessful or less effective.

The recorded EEG signals include oscillations at a various frequencies. The recordings for clinical and physiological interests lie in the range between $3.5Hz$ and $40Hz$. For EEG-based emotion recognition the frequency range of interest is usually chosen to be between $4Hz$ and $45Hz$. The frequency range can be broken down even further into frequency bands, which are used later to extract features. The frequency bands are delta (δ, $0-4$ Hz), theta (θ, $4-8$ Hz), alpha (α, $8-16$ Hz), beta (β, $16-32$ Hz), and gamma (γ, 32 Hz and above) [70]. The exact definition of these bands can differ a little in some texts, for example in [64] the alpha and beta frequency bands have been defined between $8-14$ Hz and $14-31$ Hz respectively. In this work, the first (more common) definition is used.

Different frequency bands can indicate different effects. The delta rhythms are described as slow brain activities that are dominant only in deep sleep stages of normal adults [100]. As the EEG recordings for emotion recognition are made during the awake states of healthy participants, these signals are most often discarded.

The theta frequency band is more dominant in normal infants and children. It has also been noted during drowsiness and sleep in adults. Since a small amount of theta rhythms appear in healthy adults who are awake, this frequency range cannot be discarded.

The alpha rhythms have been noted most often in normal adults during relaxed and mentally inactive state. The amplitude of these signal frequencies are often less than $50mV$ and are prominent in the occipital area. However, these rhythms are blocked by opening the eyes (visual attention) and other mental efforts such as thinking [64].

The beta activity is often seen in the frontocentral region and usually have less

amplitude than alpha rhythms. The beta rhythms have been related to expectancy states and tension. The beta band is the largest frequency band covering the frequencies from $16Hz$ to $32Hz$. Due to the large range of frequency it covers, sometimes the beta band is split into two, the *lower beta* band and the *higher beta* band.

The highest frequency is in the gamma rhythms and are often discarded as they are not clinically and physiologically of interest. However, in some applications, it is useful to include some of the higher frequencies. This can be solved by defining the gamma with a maximum band, e.g., in [69] the gamma band is defined as $32 - 64Hz$ and frequencies above $64Hz$ are considered noise.

The most common way to extract these frequency bands is by applying the wavelet transform (WT) [17] to the signals. WT is a spectral estimation technique that expresses all functions as an infinite series of wavelets. It maps a one dimensional signal to a two-dimensional function by decomposing a signal as a superposition of simple units from which the original signal can be reconstructed. The decomposition of the EEG signal via WT leads to a set of wavelet coefficients which represents its energy distribution in time and frequency.

The two types of wavelet analysis are continuous wavelet transform (CWT) and discrete wavelet transform (DWT). CWT considers data in to be continuous in time and frequency space. Therefore, the original signal can be expressed as a weighted integral of the continuous basis wavelet function. DWT considers data at discrete points. This results in the inner product of the original signal with the basis wavelet function to be taken at discrete points, hence the original signal is expressed as a weighted sum of a series of basis functions.

In DWT, filters of different cutoff frequencies are used to analyse the signal at different scales. The main difference between CWT and DWT lies in the way the two methods handle scale parameters, i.e., the CWT discretises scale more finely than the DWT.

Consider a DWT, and let $a = a_0^m$ and $b = nb_0$. Therefore, the analysing wavelets are discretised as

$$\psi_{n,m}(t) = a_0^{\frac{-m}{2}} \psi \left(\frac{t - nb_0}{a_0^m} \right), \tag{2.5}$$

where n, m are integers. Hence,

$$W_{n,m} = \int_{-\infty}^{\infty} \psi^*_{n,m}(t) x(t) \, \mathrm{d}t, \tag{2.6}$$

$$x(t) = k_\psi \sum_m \sum_n W_{n,m} \psi_{n,m}(t), \tag{2.7}$$

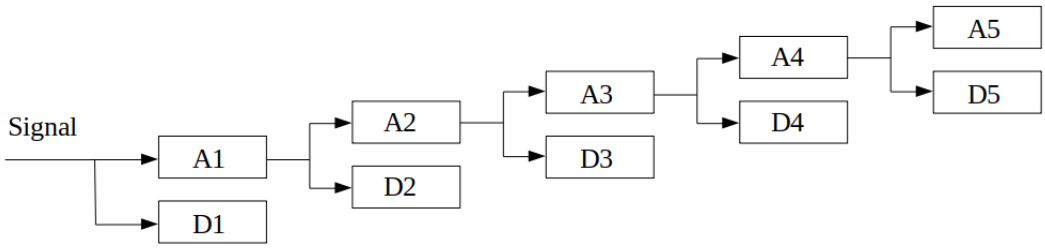


Figure 2.6: Wavelet decomposition.

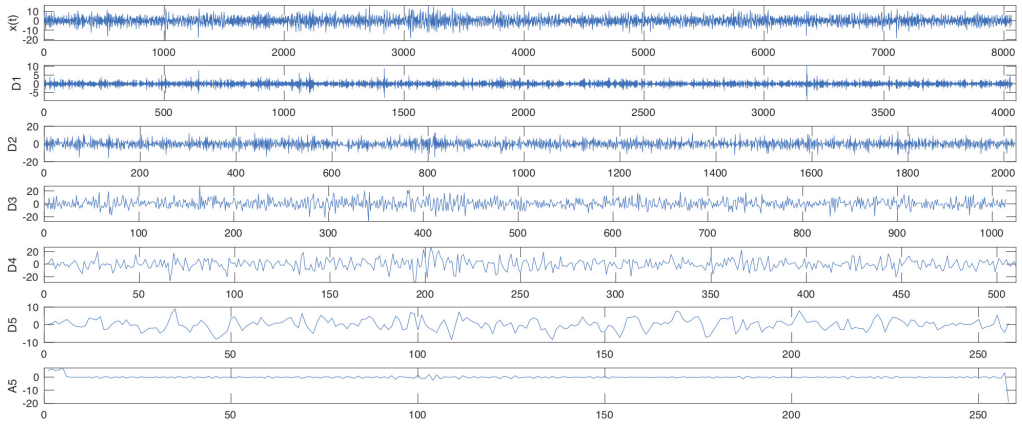


Figure 2.7: Wavelet decomposition of an EEG signal using "db4" wavelet.

where k_ψ is a constant value of normalisation. The equations (2.6) and (2.7) are respectively the DWT and its inverse.

An important task when using wavelets is to choose the wavelet decomposition level. In EEG signal processing, the the level selection is closely linked with the sampling rate of the EEG signal and the choice of the frequency bands used for further analysis. In this thesis, the datasets used are downsampled to 128 Hz and five levels of wavelet decomposition have been used. When using five levels of decomposition, shown on Figure 2.6, the first level coefficients correspond to the Gamma frequency band, second level coefficients correspond to the Beta frequency band , and so on. An example of wavelet decomposition of an EEG signal can be seen in Figure 2.7, where the $x(t)$ shows the full noise free signal with sampling rate 128 Hz, $D1$ shows the corresponding gamma frequency band, $D2$ shows the corresponding beta frequency band, $D3$ the corresponding alpha frequency band, the $D4$ corresponding theta frequency band, $D5$ the delta frequency band, and finally $A5$ is the corresponding noise.

In this thesis the main wavelets used are Symlets and Daubechies wavelets,

namely "db4", and "sym8". They are chosen due to their good fit for EEG signals [70].

For CWT, consider a real signal $x(t)$ at time t. The CWT of the signal is defined as

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \psi^* \left(\frac{t-b}{a} \right) x(t) dt \ , \tag{2.8}$$

where a is the scale of the analysed wavelet, b is the time shift factor, a, $b \in \mathrm{R}$, and ψ is a wavelet function with complex conjugate ψ^*. Define $\psi_{a,b}(t)$ as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right). \tag{2.9}$$

Hence, the (2.8) can be rewritten as

$$W_x(a,b) = \int_{-\infty}^{\infty} \psi^*_{a,b}(t) x(t) dt \ , \tag{2.10}$$

i.e., as the scalar or inner product of the signal $x(t)$ and function $\psi_{a,b}(t)$.

The most important properties of the wavelet transform are the admissibility and the regularity conditions. The admissibility condition is

$$\int \frac{|\Psi(\omega)|^2}{|\omega|} \, \mathrm{d}\omega < +\infty, \tag{2.11}$$

where $\Psi(\omega)$ is the Fourier transform[7] of $\psi(t)$. From the admissibility condition it follows that the Fourier transform of $\psi(t)$ vanishes at the zero frequency, i.e.,

$$|\Psi(\omega)|^2 \Big|_{\omega=0} = 0 \tag{2.12}$$

This implies that wavelets must have a bandbass-like spectrum, which turns out to be an important observation that will be used to build efficient wavelet transform.

Furthermore, the Fourier transform of the wavelet basis function vanishing at the zero frequency also means that the average value of the wavelet in the time domain must be zero, i.e.,

$$\int \psi(\omega) \, \mathrm{d}t = 0, \tag{2.13}$$

hence, the wavelet basis function must be oscillatory, i.e., it must be a *wave*.

The second important property of the wavelet transform is the regularity condition. This is of importance as the time-bandwidth product of the wavelet transform is the square of the input signal. For most practical applications this is not a desirable property. Hence, additional conditions on the wavelet functions to make

[7]The Fourier transform is not covered in detail in this thesis, for more information see[11]

the wavelet transform decrease quickly with decreasing scale a are needed. These conditions, the regularity conditions, state that the wavelet function should have some smoothness and concentration in both time and frequency domains. Regularity is a complex concept which can be explained using the concept of vanishing moments.

First, expand the equation (2.10) into the Taylor series at $t = 0$ of order n. For simplicity, let $b = 0$. Thus,

$$W_x(a,0) = \frac{1}{\sqrt{a}} \left[\sum_{p=0}^{n} x^p(0) \int \frac{t^p}{p!} \psi\left(\frac{t}{a}\right) \mathrm{d}t + O(n+1) \right], \tag{2.14}$$

where x^p denotes the pth derivative of function x, and $O(n+1)$ denotes the rest of the Taylor series expansion. Consider the moments M_p of the wavelet, defined as

$$M_p = \int t^p \psi(t)\, \mathrm{d}t. \tag{2.15}$$

Using (2.15), the equation (2.14) can be rewritten as:

$$W_x(a,0) = \frac{1}{\sqrt{a}} \left[x(0) M_0 a + \frac{x^{(1)}(0)}{1!} M_1 a^2 + ... + \frac{x^{(n)}(0)}{n!} M_n a^{n+1} + O(a^{n+2}) \right] \tag{2.16}$$

As at the 0th moment $M_0 = 0$, the first term of the right hand side of (2.16) is also zero. If the other moments up to M_n are zero as well, then the wavelet transform coefficients $W_x(a,b)$ will decay as fast as a^{n+2} for a smooth signal $x(t)$. This is known as vanishing moments or approximation order, meaning that if a wavelet has N vanishing moments, then the approximation order of the wavelet is also N. The number of vanishing moments depends on the application[8].

The aim of CWT is to compare a signal to shifted and scaled copies of a basic wavelet. However, CWT is not always practical, as the obtained wavelet coefficients will be highly redundant and for practical applications these redundancies should be removed. Furthermore, there exists an infinite number of wavelets in the wavelet transform which should be reduced to a finite number. Finally, for most functions, the CWT has no analytical solutions, i.e., is intractable, and therefore can only be obtained via numerical approximate methods.

Empirical mode decomposition (EMD)

Similarly to wavelet analysis, empirical mode decomposition (EMD) is a non-stationary data processing method and is often used for signal processing. EMD is

[8]The moments do not have to be exactly zero. Often a small value is sufficient, depending on the application.

based on the assumption that any non-stationary and non-linear time series consists of different simple intrinsic modes of oscillation [113]. These intrinsic oscillatory modes can be identified empirically by their characteristic time scales.

The goal of EMD is to decompose the signal using these intrinsic oscillatory modes. The majority of the riding waves[9] can be eliminated using a process called sifting. Hence, the EMD algorithm breaks the signal down into its intrinsic mode functions (IMF) by considering the signal oscillations at a very local level and separating the data into locally non-overlapping time scale components [113].

Algorithm 1 EMD algorithm: The shifting process to decompose the data set (signal) $x(t)$ into IMFs $x_n(t)$ and a residuum $r(t)$ such that the signal can be represented as $x(t) = \sum_n x_n(t) + r(t)$

$n := 1$, $k := 1$
$r_0(t) = x(t)$
$h_0 := r_{n-1}(t)$,
while do $r_{n-1}(t) \neq 0$ or $r_{n-1}(t) = r_{n-1}(t)$
  while do I_i has non-negligible local mean
    $U(t)$ = spline through local maxima of I_i
    $L(t)$ = spline through local minima of I_i
    $Av(t) = \frac{1}{2}(U(t) + L(T))$
    $I_i(t) = I_i(t) - Av(t)$
    $i = i + 1$
  end while
  $IMF_n(t) = I_i(t)$
  $r_n = r_{n-1}(t) - IMF_n$
end while

The two main characteristics of IMFs are: only one extremum between two subsequent zero crossings, and all IMFs have a mean value of zero. The first condition means that the number of local minima and maxima differs by a maximum of one, whereas the second condition implies that the IMF is stationary and simplifying its analysis[10]. The detailed algorithm for EMD showing how IMFs are obtained is given in Algorithm 1.

Figure 2.8 illustrates an example of how EMD works in practice, using an EEG signal. It shows 3 out of 8 IMFs of an EEG signal. It also shows the original signal and its residual.

[9]The riding waves are oscillations with no zero crossing between extrema
[10]IMFs may have amplitude modulation and changing frequency.

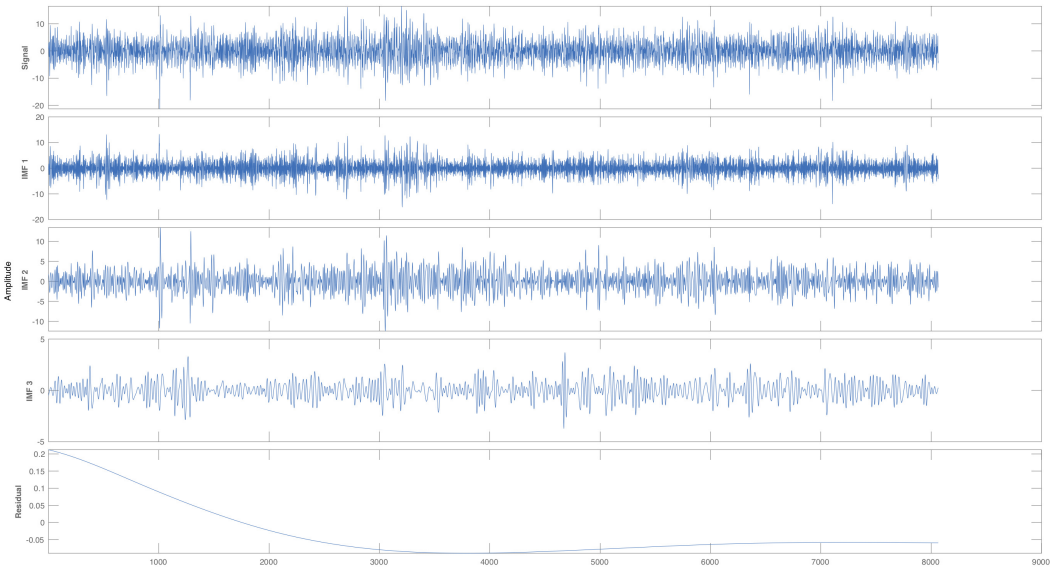


Figure 2.8: EMD of an EEG signal, showing 3 out of 8 IMFs.

2.2.2 Feature extraction

Feature extraction is used to extract the underlying characteristics of the data as uniquely and explicitly as possible. Following some of the commonly used feature extraction methods that have been shown to give good results in studies that used EEG signals, the methods in [34, 69, 77, 114] will be explained.

The useful signal features can be split into three sets depending on the source of this information: spatial, spectral, and temporal information. The features from spectral information (i.e., spectral features) describe how the power varies in the relevant frequency bands. Temporal features are useful to describe how the relevant signal varies within time. Finally, the spatial features describe the information relevant to the location of that signal. In terms of EEG signal these features concentrate on the EEG channels.

There exists a number of features that can be extracted from (EEG) signals. In this thesis, temporal features, namely statistical features (SF) and higher order crossing (HOC) features will be considered. In addition, the spectral features, namely power spectral density (PSD) and higher order spectral (HOS) features will be discussed.

Statistical features (SF)

One of the most common feature extraction techniques is to use statistical measures to characterise the data. Statistical features are time domain features (temporal).

Most commonly, these measures involve calculating mean (μ), standard deviation (σ), first and second order differences (Δ and Γ, respectively), and normalised first and second order differences (i.e., $\bar{\Delta}$ and $\bar{\Gamma}$, respectively). These measures are computed as follows:

Consider a signal $x_i(t)$, $i = 1, .., N$ and $t = 1, ..., M$, where N is the number of signals, and M is the number of samples. The mean of a signal $x_i(t)$ is:

$$\mu_{x_i} = \frac{1}{M} \sum_{t=1}^{M} x_i(t). \tag{2.17}$$

The standard deviation of a signal $x_i(t)$ is:

$$\sigma_{x_i} = \left(\frac{1}{M-1} \sum_{t=1}^{M} (x_i(t) - \mu_{x_i})^2 \right)^{\frac{1}{2}}. \tag{2.18}$$

The first and second order differences of the signal $x_i(t)$ are respectively:

$$\Delta_{x_i} = \frac{1}{M-1} \sum_{t=1}^{M-1} |x_i(t+1) - x_i(t)|, \tag{2.19}$$

and

$$\Gamma_{x_i} = \frac{1}{M-2} \sum_{t=1}^{M-2} |x_i(t+2) - x_i(t)| \tag{2.20}$$

respectively. Finally, the normalised first and second order differences are respectively:

$$\bar{\Delta}_{x_i} = \frac{\bar{\Delta}_{x_i}}{\sigma_{x_i}}, \tag{2.21}$$

and

$$\bar{\Gamma}_{x_i} = \frac{\bar{\Gamma}_{x_i}}{\sigma_{x_i}} \tag{2.22}$$

respectively.

These six features are the most commonly used in EEG-based emotion recognition. However, there exist multiple additional functions that allow for measuring common distinctive features of signals, that are required for additional features for robust in-depth analysis. For example, in addition to the regular mean and variance, robust estimates of mean and variance can be considered. The robust estimates are useful when dealing with noisy data, which is the case with EEG data even after pre-processing. In addition, when addressing noisy data, the mean excluding outliers (or trimmed mean) can be calculated where a number of highest and lowest data vales are excluded.

The additional statistical features that can be useful are as follows. Minimum and maximum amplitudes of the signal, respectively defined as:

$$x_{i,\min} = \min(x_i(t)), \tag{2.23}$$

and

$$x_{i,\max} = \max(x_i(t)). \tag{2.24}$$

Peak magnitude to root-mean-squared (RMS) ratio of the signal, i.e.,

$$r_{p2rms} = \frac{||x_i(t)||_\infty}{\sqrt{\frac{1}{M}\sum_{m=1}^{M}|x_i(t=m)|^2}}, \tag{2.25}$$

where $||\cdot||_\infty$ is the L^∞ norm. The peak to RMS ratio is the ratio of the largest absolute value in $x_i(t)$ to the RMS value of $x_i(t)$. The mean absolute deviation is a measure of variability, i.e.,

$$MAD = \frac{1}{M}\sum_{t=1}^{M}|x_i(t) - m(x_i)|, \tag{2.26}$$

where the function $m(\cdot)$ is the measure of central tendency. Furthermore, skewness, mobility, complexity, occupied bandwidth of the signal, and mean normalised frequency can be calculated and are used as features in this thesis.

The skewness of a signal measures the asymmetry of that signal (data) around the sample mean. Consider a signal $x_i(t)$, as before. The skewness is given by

$$s = \frac{E(x_i(t) - \mu)^3}{\sigma^3}, \tag{2.27}$$

where μ is the sample mean, σ the variance, and $E(\cdot)$ is the expected value. Negative skewness indicates that the data is spread more to the left of the mean than to the right, and positive skewness indicates that the data is spread more to the right. In the case of any perfectly symmetric distribution of the data, the skewness is equal to zero.

Mobility and complexity are both Hjorton's parameters [33]. The mobility parameter represents the mean frequency, and the complexity parameter represents the change in frequency. The mobility of a signal $x_i(t)$ is defined as:

$$Mob = \sqrt{\mathrm{var}\left(\frac{dx_i(t)}{dt}\right) \Big/ \mathrm{var}(x_i(t))}. \tag{2.28}$$

The complexity parameter compares similarity of a signal to a pure sine wave, where

its value converges to 1 if the signal is more similar. The complexity parameter is defined in terms of mobility,

$$Comp = \frac{Mob(\frac{dx_i(t)}{dt})}{Mob(x_i(t))}. \tag{2.29}$$

The occupied bandwidth is the difference in frequency between the points where the integrated power crosses 0.5% and 99.5% of the total power in the spectrum. Finally, the mean normalised frequency is the estimate of the mean normalized frequency of the power spectrum of a time-domain signal x_i.

Power spectral density features(PSD)

The second set of features involve the power spectral density (PSD). PSD is a measure of a signal's power intensity in the frequency domain (and therefore classify under frequency domain features), and is computed using the discrete Fourier transform.

Consider a signal $x_i(t)$, $i = 1, .., N$ and $t = 1, ..., M$, where N is the number of signals, and M is the number of samples. The PSD $\Phi_{xx}e^{j\Omega}$ is given as the discrete time Fourier transform (DTFT), denoted $\mathcal{F}_*\{\}$, of the autocorrelation function (ACF), denoted $\varphi_{xx}(t)$, i.e.,

$$\Phi_{xx}e^{j\Omega} = \mathcal{F}_*\{\varphi_{xx}(t)\}. \tag{2.30}$$

Multiple methods can be used to compute the PSD, e.g., the periodogram estimator[11], Bartlett's method, and Welch's method. To estimate the PSD using either Bartlett's or Welch's method, the signal is split into segments. The PSD is estimated for each segment separately, and finally to reduce the variance of the PSD estimate, an average over these local estimates is computed. The main difference between these two methods is that while the Bartlett's method uses non-overlapping segments, the Welch's method is a generalisation using overlapping windowed segments.

As before, consider a signal $x_i(t)$, $i = 1, .., N$ and $t = 1, ..., M$, where N is the number of signals, and M is the number of samples. First, the signal $x_i(t)$ is split into L segments. Let $x_{i,l}(t)$, $l = 1, ..., L$, be the segment l of length P, $P < M$, starting at multiples of step size $h \in 1, ..., P$. In addition, let R denote the overlapping points[12]. Next, all segments $x_{i,l}(t)$ are windowed by the window $w(t)$ of

[11]The periodogram estimator of the PSD is not consistent. This is the result of variance not converging towards zero even when the signal length is increased towards infinity.

[12]For Bartlett's method, $R = 0$, and there is no overlap.

length P[13]. The DTFT of the windowed segment l is given by:

$$X_l(e^{j\Omega}) = \sum_{t=0}^{P-1} x(t + l \cdot h)w(t)e^{-j\Omega t}, \tag{2.31}$$

where the window $w(t)$, $0 \leq t \leq P-1$, is normalised as

$$\frac{1}{P}\sum_{t=0}^{P-1} |w(t)|^2 = 1. \tag{2.32}$$

The step size h determines the overlap between the segments. In general the number of overlapping samples in adjacent segments is $P - h$.

As the Welch's method estimates the PSD, for clarity, the resulting estimate of the PSD of the segment is denoted as $\hat{\Psi}_{xx,l}(e^{j\Omega})$. Hence, the PSD is achieved by introducing $X_l(e^{j\Omega})$ into the definition ot the periodogram[14], and the use of the l-th segment, i.e.,

$$\hat{\Psi}_{xx,l}(e^{j\Omega}) = \frac{1}{P}|X_l(e^{j\Omega})|^2, \tag{2.33}$$

and hence the estimated PSD is given by averaging over all the segments periodograms, i.e.,

$$\hat{\Psi}_{xx}(e^{j\Omega}) = \frac{1}{L}\sum_{l=0}^{L-1} \hat{\Psi}_{xx,l}(e^{j\Omega}). \tag{2.34}$$

The main PSD feature used in this thesis is the PSE [114]. PSE is able to quantify the spectral complexity and the amount of potential information conveyed in the power spectrum of a given signal. The PSE is computed using the entropy function, i.e.,

$$H = -\sum_{m=1}^{M} P(m)\log P(m), \tag{2.35}$$

where $P(m)$, $m = 1, ..., M$, is the probability distribution given by $P(m) = \hat{\Psi}_{xx}(e^{j\Omega})/\sum \hat{\Psi}_{xx}(e^{j\Omega})$. However, one can extract many features from PSD, including statistical features explained in section 2.2.2.

Higher order spectral (HOS)

The third method to extract features is based on higher order spectral (HOS). The HOS features are also frequency domain features. This set utilizes the spectral representations of higher order moments, i.e., cumulants of the signal. The HOS analysis is useful when dealing with non-Gaussian signals which have Gaussian noise

[13]Often Hamming or Hann windows are used.
[14]The detailed description of periodogram can be found in [108].

and mixed-phase signals, and when dealing with non-linear signal (or signals).

To compute the third order correlation, i.e., bispectrum, consider a signal $x_i(t)$ with a discrete Fourier transform [11] evaluated on N data points, i.e.,

$$X\left(e^{j\Omega}\right) = \sum_{t=0}^{N-1} x_i(t) \exp^{j\Omega} , \tag{2.36}$$

where f is the frequency variable. The bispectrum is the Fourier transform of the third order correlation of the signal and is given by [34], i.e.,

$$Bis(f_1, f_2) = E[X(f_1)X(f_2)X^*(f_1 + f_2)] , \tag{2.37}$$

where $X^*(f)$ denotes the complex conjugate of $X(f)$, and $E[\cdot]$ is the statistical expectation operator. The normalised bispectrum, i.e., bicoherence, is [34]

$$Bic(f_1, f_2) = \frac{Bis(f_1, f_2)}{\sqrt{P(f_1)P(f_2)P(f_1 + f_2)}} , \tag{2.38}$$

where the power [34, 97]

$$P(f) = E[X(f)X^*(f)]. \tag{2.39}$$

Five features are computed, namely, sum of the bispectrum magnitudes $\left(f^1\right)$, sum of the squares of the bispectrum magnitudes $\left(f^2\right)$, sum of the bicoherence magnitudes $\left(f^3\right)$, sum of the squares of the bicoherence magnitudes $\left(f^4\right)$, and test of Gaussianity $\left(f^5\right)$.

Higher order crossing (HOC)

The higher order crossing (HOC) is based on the local and global movement (up and down) of the time series and is classified under time domain feature methods. This behaviour can be described by applying a sequence of high-pass filters to the zero-mean time series $x_i(t)$ [77], i.e.,

$$\mathfrak{T}_k\{X(t)\} = \nabla^{k-1}X(t) , \tag{2.40}$$

where ∇ is the iterative difference operator. We use $\nabla \equiv X(t) - X(t-1)$, and $k = 1, ..., L$, where L is the number of filters. The HOC sequence D_k, i.e., the resulting k features, comprises the number of zero-crossings of the filtered time series

by counting its sign changes, i.e.,

$$T_k\{X(t)\} = \sum_{j=1}^{k} \begin{pmatrix} k-1 \\ j-1 \end{pmatrix} (-1)^{j-1} X(t-j+1). \tag{2.41}$$

We construct a binary time series

$$Y_t(k) = \begin{cases} 1 & \text{if } \mathfrak{T}_k\{X(t)\} \geq 0 \\ 0 & \text{if } \mathfrak{T}_k\{X(t)\} < 0 \end{cases}, \quad k = 1, 2, ...;\ t = 1, ..., N. \tag{2.42}$$

Hence, the simple HOC is estimated by counting the symbol changes in binary time series $Y_t(k)$, giving the feature vector

$$V_{HOC} = [D_1, ..., D_L], \tag{2.43}$$

where $D_k = \sum_{t=2}^{N} [Y_t(k) - Y_{t-1}(k)]^2$. The different HOC features are computed to represent the oscillatory patterns present in the EEG data.

2.2.3 Feature selection

The extraction of features from EEG signals using 5 wavelet bands results in a large number of features. Due to the limited number of data points in the datasets available for EEG-based emotion recognition, the number of features is significantly higher than the number of data points, resulting in model over-fitting. To overcome over-fitting, feature selection can be used to reduce the number of features used to train the model, where the aim of the feature selection algorithm is to find a new set of the most informative features. The rule of thumb in feature selection is to make the number of features fewer than the number of observations to obtain a well-specified model.

Performing feature selection on small datasets can turn out to be problematic, as the features selected in the test set may not be the best features to use for the training set. For small datasets, leave-one-out cross validation is often used. However, for small datasets, even one training sample can make a big difference. To overcome this, the feature selection was performed on the training set in a leave-one-out fashion. That is, in the feature selection step, an additional sample was left out from the training set, and the features selected in a loop. The final set of features used to train the classifiers were the most favored ones over the whole loop.

In addition to feature selection methods, feature extraction methods (e.g.

PCA, kernel PCA) can be used for dimensionality reduction. The difference between feature extraction and feature selection is that the first generates a new set of features from functions of the original features, whereas the latter returns a subset of the original features leaving out the redundant features. There exists a lot of work using both methods, however in general the feature selection methods have been shown to work better for EEG-based emotion recognition. This can be because the used datasets tend to have a considerably more features than samples, and uncorrelated features reduce the performance of classifiers. In [20] the PCA feature extraction has been compared to mRMR feature selection. It has been shown that the mRMR feature selection resulted in higher accuracy for all cases. In [86], the kernel PCA has been used reaching accuracy of 76.9% for valence, and 69.1% for arousal. As they also use DEAP dataset, [86] has been taken as one of the comparison works in this thesis.

ReliefF

One of the feature selection methods considered in this work is the ReliefF algorithm [49], an extension of Relief algorithm [45]. The Relief algorithm is a favourable method as it is not dependent on heuristics, it uses low-order polynomial time, and is tolerant to noise and robust to feature interactions. It is a simple method with low computational time. However, Relief does not behave well when the datasets are small and cannot be extended to multi-class problem. To address this, [49] proposed an extension called ReliefF, which behaves better with small training sets and can be extended to the multi-class problem.

In the examples given later on, the feature selection method ReliefF, is applied to select features by first selecting an instance and finding k near misses and hits. That means the instances corresponding to the same class (called hits), and instances from different classes (called misses) are counted. These are used to calculate the weight vector which is used to describe the quality of features. Finally, the features with the highest quality (i.e., in accordance with the weight vector) are chosen. In general, the number of selected features is chosen to be smaller than the number of samples.

Maximum relevance minimum redundancy (mRMR)

In general, the purpose of feature selection is to find a reduced set of features S from the whole set of features S_{ALL} using a certain criterion. Maximum relevance

criterion is to find features that satisfy

$$\frac{1}{|S|}\sum_{x_i \in S} I(x_i; c). \tag{2.44}$$

where c are the target labels, x_i are the features in S, i.e., $x_i \in S$, and $I(x_i; c)$ is the mutual information between the feature x_i and the labels c.

While the maximum relevance criterion approximates the maximum dependency $D(S, c)$ with mean of the mutual information of all individual features and class c, there could be a large dependency between the features. It is noted that if two features are highly dependent on each other, removing one would generally not make a large difference, as the respective class-discriminative power would not change significantly. In other words, maximum relevance criterion can have high redundancy.

Consider the minimum redundancy condition

$$\frac{1}{|S|^2}\sum_{x_i, x_j} I(x_i, x_j). \tag{2.45}$$

Combining the conditions (2.45) and (2.44) gives rise to the mRMR feature selection, which can check for the superfluous features and give a set of features without any redundancy. The first feature is chosen to be the one with highest mutual information between the feature set and the output. All the additional features are chosen inductively using:

$$\underset{x_j \in S_{ALL} - S_m}{\operatorname{argmax}} I(x_j; c) - \frac{1}{m}\sum_{x_i \in S_m} I(x_i, x_j), \tag{2.46}$$

where the set S_m denotes m already chosen features.

2.2.4 Classification

The most common classifiers used in an EEG-based emotion recognition framework include the support vector machine (SVM), k-nearest neighbours (kNN), and Naive Bayes (NB). An overview of these methods will be given in this section.

Support vector machine (SVM)

SVM [105] is a binary classifier which can be extended into a multiclass classifier. It is chosen due its flexibility, and it has been shown to work well for classification. Consider a training set $(\mathbf{x}_j, y_j),\ j, ..., N$, where $\mathbf{x}_j$ denote the feature vectors extracted from EEG signals, y_j denote the corresponding emotion labels, and N is the number

of data.

The SVM decision function can be written as

$$f(\mathbf{x}) = \sum_{j}^{N} \alpha_j y_j k(s_j, \mathbf{x}) + b, \tag{2.47}$$

where $\mathbf{x}$ is the input vector (in this case feature vector extracted from EEG signals), k is the kernel function, s_j denote support vectors, α_i are the weights and b is the bias. In the scope of this paper, we used the Gaussian kernel, i.e.,

$$k(s_j, \mathbf{x}) = \exp\left(-\gamma ||s_j - \mathbf{x}||^2\right), \tag{2.48}$$

where γ is the kernel scale parameter. To train the SVM, weights α_j are found for existing data such that

$$f(\mathbf{x}_j) = \begin{cases} \geq 0 & y_j = +1 \\ < 0 & y_j = -1 \end{cases}, \tag{2.49}$$

where $+1$ and -1 denote positive and negative emotion classes, respectively.

K-Nearest neighbours (KNN)

KNN has been shown to work well with EEG signals in [69]. The classification is based on user-defined constant integer k, where a new case is assigned to the class most common amongst its k nearest neighbours measured by a distance metric. Most commonly, the Euclidean distance is used as the distance metric, but Manhattan, Minkowski, and Hamming distances can also be used. The problem with KNN is when the training set is imbalanced, the classes with more examples tend to dominate the classification.

For the purpose of this thesis, MATLAB inbuilt function *fitcknn* is used to fit KNN model to the data. This function attempts to minimize the cross-validation loss for the *fitcknn* by varying the parameters, including the number of neighbours and distance metric depending on the dataset.

Naive bayes (NB)

A NB classifier [21, 24] assumes all input variables are independent. It aims to find the conditional probability that data samples belong to a specific class given the input features, and chooses the class with the highest probability. Thus, the goal of NB classifier is to find the probability $p(C|F_1, ..., F_n)$, where C is the class indicator

variable and $F_1, ..., F_n$ are the features. This probability is difficult to compute, and thus the Bayes theorem,

$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)} \tag{2.50}$$

is used instead. For all classes, the marginal of (2.50) is the same and thus only the numerator must be computed. Assuming that all input variables F_i are independent,

$$p(C|F_1, ..., F_n) \propto p(C)p(F_1, ..., F_n|C) = p(F_1, ..., F_n, C)$$

where

$$\begin{aligned} p(F_1, ..., F_n, C) &= p(F_1|F_2, ..., F_n, C)p(F_2, ..., F_n, C) \\ &= p(F_1|F_2, ..., F_n, C)p(F_2|F_3, ..., F_n, C)p(F_3, ..., F_n, C) \\ &= p(F_1|F_2, ..., F_n, C)p(F_2|F_3, ..., F_n, C)\times \\ &\quad \times ...p(F_{n-1}|F_n, C)p(F_n|C)p(C) \end{aligned}$$

and so forth. Assuming all input variables (features) F_i are independent, $p(F_i|F_{i+1}, ..., F_n, C) = P(F_i|C)$, and (2.50),

$$p(C|F_1, ..., F_n) = \frac{p(C)\prod_{i=1}^{n} p(F_i|C)}{p(F_1, ..., F_n)} \tag{2.51}$$

$$= \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C), \tag{2.52}$$

where Z is a scaling factor dependent on $F_1, ..., F_n$. Despite the assumption, a NB classifier still performs surprisingly well even when the assumption is not entirely accurate.

Other classifiers

Neural networks are often proposed as a good method for EEG based emotion recognition. Neural networks work best when dealing with very large number of samples. The publicly available datasets contain small number of training samples, and therefore, in the scope of this thesis, neural networks have not been used as a classification method.

In some cases Linear Discriminant Analysis (LDA) is used for emotion recognition. LDA is used as it is easy to implement, and the classification is fast. However, as LDA assumes the multivariate normal distribution for all classes, and in practice

this is very rare, the LDA classification tends to under perform when compared to other classifiers.

2.2.5 Complexity analysis

One of the major problems with EEG-based emotion recognition system is that it is computationally very complex, and therefore the required training time is long. The computational complexity varies for different methods. The computational complexity of WT using discrete wavelet transform is $O(N)$, where N is the size of the signal. The EMD however has a the computational complexity of $O(N \log N)$, which is same as the complexity of fast Fourier transform. In signal pre-processing, the computationally most costly part is the BSS, where even when the complexity of estimation of correlation matrix is excluded, the post-processing cost of *iWASOBI* is $O(d^2M)$, where d is the number of signal components and M is the number of covariance matrices [102].

The feature extraction methods considered have varying computational cost. The HOC and statistical feature extraction have relatively low computational complexity, namely $O(N)$. However, HOS features need computation of bispectrum and bicoherence. Bispectrum invokes a computational cost of $O(N^2)$. For our emotion recognition framework, there is a need to compute bispectrum multiple times, which results in large computation complexity. In addition, to compute bicoherence, there is a need to compute PSD, which also has the computational complexity of $O(N)$ for a window size of N.

The computational complexity of the feature selection algorithms is $O(nFm)$ *reliefF* and $O(nFd)$ for mRMR, where n is the number of samples, F the total number of features, d number of selected features, and m number of training instances (user chosen). However, additional complexity is added when performing feature selection in leave-one-out fashion.

Finally, the complexity of classification step is discussed. Let the number of samples be n, each with d dimensions (i.e., features selected). The complexity of kNN classifier is $O(nd + nk)$, and the complexity of NB classifier is $O(nd)$. The linear SVM has the computational complexity of $O(d)$, however the kernel SVM using either polynomial kernel or Gaussian kernel would result in computational complexity of $O(n_{SV}d)$, where n_{SV} is the number of support vectors.

2.3 Datasets

Publicly available datasets are used in this thesis to illustrate the EEG-based emotion recognition methods proposed. Three EEG datasets have been chosen,

namely DEAP [47], MAHNOB [94], and DREAMER [39]. All these datasets used valence-arousal (dominance) scale to label the datasets. All three datasets are created to be used for EEG-based emotion recognition, however they differ in size and choice of stimuli. Using all three datasets aims to validate the robustness of the model, by demonstrating the models proposed generalise well. DEAP is the largest publicly available EEG dataset for emotion recognition. MAHNOB-HCI and DREAMER are smaller, making the classification task more complicated.

The advantage of the publicly available datasets lies in the possible comparisons between other methods using the same datasets. By using the same dataset, one-on-one comparisons can be made with state-of-the-art methods. In this thesis, comparisons between methods proposed in literature using DEAP, MAHNOB, or DREAMER datasets and methods proposed in this thesis are presented.

2.3.1 DEAP

The DEAP dataset [47] is the largest publicly available dataset. It includes recordings from 32 participants, of which 16 are men and 16 are female. Each participant was asked to watch 40 music videos. The EEG signals were recorded using Biosemi technology, with 32 electrodes set according to the international 10-20 system [36]. In addition, the dataset includes recordings from 12 peripheral channels, 3 unused channels and 1 status channel.

The data was labelled through participant self-assessment, where at the end of each trial (video) each participant gave the rating using the valence-arousal-dominance scale. The scales range from unhappy/sad to happy/joyful for valence, calm/bored to stimulated/excited for arousal, and submissive to dominant for dominance scale. For each video, the participant rated the valence, arousal, dominance and liking on a continuous scale (up to one decimal point) between 1 and 9. In addition, the dataset includes the familiarity ratings (between 1-5).

The dataset includes both raw signals and pre-processed signal. In this thesis, the raw signals are used and processed to keep the pre-processing step consistent over all datasets. Furthermore, the dataset includes the EMG and EOG signals that can be used to remove artifacts from the EEG signal. However, as not all these signal are available for other datasets, the artifact removal is performed using automated methods, that do not require additional information.

2.3.2 MAHNOB-HCI

The MAHNOB-HCI [94] database consists of two experiments, emotion recognition and implicit tagging. For the purpose of this thesis, the interest lies in the experiment

considering emotional responses to videos. The emotion response experiment has recordings from 27 participants, from which 11 are males and 16 are females. However, only the complete recordings from 25 of the participants were used. Similarly to DEAP dataset, the recordings were made using Biosemi system, with 32 electrodes set according to the international 10-20 system. Since the dataset included the raw signals, these signals are pre-processed similarly to DEAP dataset.

The MAHNOB dataset is significantly smaller than DEAP dataset. Each of the participant watched 20 video clips, during which the EEG signals were recorded. The DEAP dataset used music videos as stimuli, whereas the MAHNOB dataset used video clips from movies.

The self-assessment of the videos was also performed. The participants were asked to rate their valence, arousal, and dominance on a nine-point scale (discrete scale). Furthermore, participants were asked to give emotional labels to the trials. The labels included neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear. For the purpose of consistency, all experiments were performed using valence and arousal results.

2.3.3 DREAMER

The third multimodal affect dataset used in this thesis to test the proposed models is DREAMER [39]. The database consists of EEG signals and ECG signals from 25 participants, of which 11 are females and 14 are males. However, due to incomplete data recordings, only 23 participants have full recordings.

The DREAMER dataset includes 18 trials per subject. Similarly to MAHNOB dataset, the stimuli used to evoke emotional responses was movie clips, and each participant was asked to label their emotions for valence, arousal, and dominance. The DREAMER dataset used the scale from 1 to 5 to label the trials.

The main difference between DREAMER dataset and previously introduced datasets (i.e., DEAP, MAHNOB-HCI) is the use of equipment. The Biosemi system used to record the DEAP and MAHNOB dataset signals is meant for research, which provides enhanced capturing capabilities and increased signal quality. Furthermore, the recording can be made from large number of electrodes (up to 128 channels). However, the system is costly, non-portable and, non-wearable, and therefore the set-up is not suitable for everyday usage. The DREAMER dataset uses Emotiv EPOC wireless EEG headset, which is portable and low cost. The number of channels the Emotiv EPOC can record is only 14. The raw signals are pre-processed similarly to DEAP and MAHNOB.

The inclusion of this dataset is to show that accurate classification of emotions can be achieved using portable equipment. However, as the dataset is small, only

two-class classification is performed.

2.3.4 EEG emotion recognition

The EEG-based emotion recognition can be either subject-dependent or subject-independent. The difference between the two approaches is how the classifiers are trained. In subject-dependent approach, a new classifier is trained for each subject (participant) separately. In general this approach gives more accurate results, as the emotions have been shown to be dependent on the subject. Nevertheless, the subject-independent approach to EEG-based emotion recognition is possible. The subject-independent approach trains only one classifier over all subjects and is therefore more robust. The results, however, tend to be less accurate than in the subject-dependent approach.

An overview of subject-dependent emotion recognition models, for up to eight emotions are given in [59]. Most of the work in EEG-based emotion recognition is subject-dependent due to different participants experiencing emotions differently. The method used by [59] combines fractal dimensions and HOC, and used SVM classifier. Similarly to the work presented in this thesis, [59] uses DEAP dataset. However, they only used data from six participants out of 32.

Emotion recognition has been shown to work well for two-class classification. In [56], two-class classification gives a recognition accuracy of 93.5% where images of facial expressions of smile and cry were used as stimuli using common spatial pattern feature extraction and SVM. In [78] the two-class classification reaches accuracy up to 94.4%. One of the main aims of emotion recognition is to increase the number of emotions recognised.

In [6], the two, three, and five-class classification were attempted using subject-dependent approach. They perform feature selection using combined genetic algorithm and SVM, and mRMR, where the latter was shown to give better results. The classification was done using SVM. The accuracy of two-class classification was 73.14% for valence and 73.06% for arousal. As expected, the accuracy of the multi-class classification is lower. The three and five-class valence classification reached the accuracy of 62.33% and 45.32% respectively. Similarly, the arousal accuracy was 60.70% for three-classes and 46.69% for five-classes. The work presented in [6] is one of the main texts used to compare the work proposed in this thesis.

Comparing the EEG-based emotion recognition results is somewhat complicated, as the datasets are small and differ from one another by their size, the recording equipment, and stimuli used. Furthermore, not all datasets are publicly available to make comparisons. Therefore, the comparisons are made using only the methods that use the same datasets. To highlight this phenomenon, comparing the

two-class classification results in [56] and [78] to [6], is unfavourable to the latter. However, this does not necessarily mean that the methods presented in [6] are less good, but can be an indication of the quality and size of the different datasets used.

This is why comparing the results presented in [6] to [86] and [57] give a better comparison. All three papers use the same dataset (DEAP dataset) making the comparison more accurate. In [86], the accuracy of valence is 76.9% and arousal is 69.1%, and in [57] the accuracy is 58.49%, and 64.3% for valence and arousal respectively. Hence, the results from [6] are far superior when compared to the models using the same datasets. It is important to note the effects the experimental data has to the results.

The highest accuracy for DEAP dataset is reached is [25], using Gaussian process latent variable model, where the accuracy of valence was 88.33% and the accuracy of arousal is 90.56%. As the DEAP is one of the main datasets used in this thesis, these results are used to compare the results achieved in this thesis.

There has been a significant amount of work published using the DEAP dataset, however,the amount of work published using MAHNOB and DREAMER is considerably smaller. Classifying valence and arousal has been done in [117] and [46] using MAHNOB dataset. In [117], canonical correlation analysis is used to find relations between EEG signals and the video content. The accuracy reaced in [117] for two-class valence classification is 55.72% and the accuracy reached for arousal is 60.23%. In [46], EEG-based emotion classification is done using PSD feature, recursive feature elimination, SVM classification. The accuracy for two-class EEG-based emotion classification is 67.5% and 70.0% respectively.

The final dataset used in this thesis for subject-dependent emotion recognition is DREAMER dataset. This dataset is only a few years old, and contains the smallest number of samples. Nevertheless, some work using this dataset has been published. In [95], a model using dynamical graph convolutional neural networks is introduced. The accuracy reached for two-class valence classification is 86.23% and the accuracy reached for arousal is 84.54%. This is the main paper that is used to compare the DREAMER dataset results presented in this thesis.

In most subject-dependent emotion recognition the feature selection is also done for every subject separately. This means that even though the same features are extracted from the signals, the features selected might be different for every subject. This, however, is not always the case. In [72], the accuracy reached for two-class subject-dependent classification using subject-independent features was 89.22%. The dataset they used contained of six subjects and the EEG signals were recorded using 64 electrode cap. The method used in [72] combined log band energy and SVM. In addition, the features were smoothed using linear dynamic system approach.

The subject-independent emotion recognition is a lot more challenging task. In addition, the physiological expressions of emotion also depend on subjects age, gender, culture and other social factors [3], as well as the environment a subject lives. Hence, for a successful subject-independent classification, often the models have to be more complex. However, the amount of training data is larger as one can make use of all the available data over all subjects.

Subject-independent classification model for EEG-based emotion recognition has been attempted in [5, 15, 98]. The best classification results are reached in [5], with an accuracy of 94.27% achieved using 256 electrodes. In the aforementioned paper, the spectral power features were used for feature extraction with multilayer perceptron based classifier. However, the dataset used in the paper only contained EEG recordings from five subjects.

When using datasets with smaller number of electrodes and larger number of subjects, the accuracy reached is useally smaller. For example, three-class classification in [15] reached 56% accuracy. The classifier used to obtain these results was LDA with ANOVA feature selection. The dataset used in this work contained EEG recordings from 20 participants.

In [98] the author used SVM classifier and statistical features for five-class emotion recognition. The accuracy reached in [98] is 41.7%, with dataset containing EEG recordings from 12 subjects.

Sometimes, a way to validate the subject-independent EEG-based emotion recognition is by using 'leave-one-subject-out' corss validation. Validating results this way gives a good idea how the model behaves on previously unknown subjects. In [58], two datasets are used for EEG-based emotion recognition, namely DEAP and SEED datasets. The highest average accuracy for two-class classification using 'leave-one-subject-out' cross validation and DEAP dataset is 59.06%. The two-class classification using SEED dataset gave the accuracy of 83.33%. This highlights well the difference in accuracy using various datasets. In SEED dataset, the EEG recordings are given for 15 subjects, where as in DEAP dataset, the EEG recordings are given for 32 subjects resulting in a lot smaller accuracy. The SEED dataset was also used in [116], where the three-class classification reached 76.31% using transfer learning.

Other papers that used subject-independent emotion recognition include [14, 37, 75, 115]. In [14] the three class SVM classification reached accuracy of 63% when using only EEG signals. The dataset used consisted of EEG recording from 11 participants.

In [37, 75, 115] the same dataset, DEAP dataset, is used for subject-independent classification. In [75] and [37] two-class classification is performed for both valence

and arousal. In [75] deep neural networks have been used for classification and wavelet coefficients have been used as features. The accuracy achieved is 62.5% for valence and 64.25% for arousal. In [37] deep learning network has been used for classification, with accuracy of 53% and 52% for valence and arousal respectively. The feature extraction technique used in [37] is power spectral density and PCA is used to select the most important features.

In [115], four class classification using 16 subjects from DEAP dataset is performed using SVM classifier. In addition, ReliefF-based channel selection is used to reduce the input size. The accuracy reached is 57.67%.

It is possible to see from the different subject-independent EEG-based emotion recognition results, that multiple different factors can have an effect on the emotion recognition results. Most of the state-of-the-art methods have only been tested on one dataset. However, as all the data is experimental, the results can be highly dependent on the datasets. In this thesis, multiple publicly available datasets are used to compare different methods, and to show that the models proposed in this theses are robust.

2.4 Summary

In this Chapter, an in-depth overview of affective computing is given. The Chapter starts out by outlining different ways of recording the brain activity, how the datasets are created, and discussing different ways to label the emotions. The foundations of different computational steps are then given. Methods used to pre-process EEG signals, extract features from the signals, select more informative features, and finally the methods used to classify the signals are presented. These methods have been used in different stages of this thesis.

Finally, the overview of all three datasets (DEAP, MAHNOB, and DREAMER) used in this thesis is given. The size, stimuli used, and differences between the different datasets are highlighted. In addition, a review of the state-of-the-art work done in EEG-based emotion recognition is presented for comparison in the later chapters.

Chapter 3

Signal reduction using adaptive windowing

In this chapter, a novel mutual information based signal windowing method is introduced for the purpose of signal reduction. For the purpose of this thesis, signal reduction is defined as the process of identifying a subsignal from the original signal that is used in the later analysis. In addition, an EEG emotion recognition framework that includes the proposed method is given. It is shown that using reduced signals will result in higher accuracy when compared to the full signal. Furthermore, an indepth comparison of multiple feature extraction and classification methods is given. This chapter is based on the paper published by the author in [80], with additional two-class classification results using DREAMER dataset, and an in-depth overview of multiclass classification using DEAP and MAHNOB datasets.

Using all available EEG data can be computationally expensive, and often will not give a viable emotion recognition. It has been shown that mutual information is a good criterion for measuring the importance of EEG based information, and has often been used for feature extraction in [6, 103].

In addition, the EEG signals corresponding to the emotions are often noisy. Some of the noise can be removed using different pre-processing techniques. However, when removing the noise some of the information related to emotions can be lost. Furthermore, audio-visual (videos) stimuli are commonly used to evoke emotions in EEG-based emotion recognition. Often these stimuli are long which result a large amounts of data (e.g., the DEAP dataset uses stimuli each of one minute duration using sampling rate of 512 Hz). If the recording period is long, a person can experience emotions with different intensity or even different emotions during the period. This can be overcome by tailoring the stimuli to evoke specific emotions. However, even the tailored stimuli can evoke the emotions at different times during

the stimuli or in rare occasion different emotions.

As EEG signals are high dimensional, having these signals recorded over long period results in large data samples. To make the signal processing faster and emotion classification more accurate, the author suggests that finding and using the part of the signal where the concentration of the useful information (i.e. information related to the emotion) is highest. The proposed mutual information based signal reduction aims to achieve this.

The Chapter is split into four parts. Section 3.1 starts out with explaining the methodology of the proposed windowing method. Section 3.2 presents the proposed framework. The two-class classification results using three different EEG emotion recognition datasets are presented in Section 3.3 and the results for multi-class classification are presented in Section 3.4. Analysis and comparisons with state-of-the-art methods follows in Section 3.5 and finally, a short summary is given in the end of the chapter in Section 3.6.

3.1 Mutual information based signal reduction

Consider an EEG signal $x_{i,j}(t)$, where $i = 1, .., N$ denote the number of channels, $j = 1, ..., L$ denote the number of data samples, and $t = 1, ..., M$ denote the length of the signal. Let $\mathbf{X}_j \in \mathrm{R}^{N \times M}$ denote a data matrix for the jth sample. Hence, the dataset $\mathcal{D}$ consists of matrices $\mathbf{X}_1, ..., \mathbf{X}_L$ together with labels $\mathbf{y} = (y_1, ..., y_L)$, where $y_j \in [1, ..., C]$, and C denotes the number of emotion classes. The dataset is represented as $\mathcal{D} = \{\mathbf{X}_j, y_j\}$.

The proposed mutual information adaptive windowing method for data reduction is an iterative method, given in Algorithm 2. First, let the maximum and minimum window sizes be denoted as W_{max} and W_{min} respectively. The maximum and minimum window sizes indicate the limits of the window. The minimum window size has to be larger than zero, i.e., $W_{min} > 0$, and the maximum window size has to be less than or equal to the length of the signal, i.e., $W_{max} \leq M$. In addition, a change constant c is introduced. The change constant is used to change the size of the current window in an iterative fashion, and should be chosen to be a multiple of the sampling rate.

Next, the window size W is set to be W_{min}, and all possible combinations of signals of size W_{min} are found. That is, consider a new data matrix $\mathbf{X}_{(k,j)}^{W_{min}} \in \mathbb{R}^{N \times W_{min}}$, where $k = 1, 2, ...K_{W_{min}}$, and $K_{W_{min}}$ is the number of different possible reduced data matrices with signal length W_{min}.

The mutual information, denoted by $MI_{(k,:)}^{W_{min}}$, is the mutual information between $\mathbf{X}_{(k,:)}^{W_{min}}$ and $\mathbf{y}_{:}$, where replacing subindex j with : means that the mutual

Algorithm 2 Determine signal window using mutual information

Require: Data $\mathbf{X}_j$, where $j = 1, ..., L$ is the number of samples of size $N \times M$.
$\mathbf{y} = [y_1, ..., y_L]$ - the emotion labels corresponding to the samples.
W_{min} - minimum window size.
W_{max} - maximum window size.
c - change constant.
f = sampling rate.
Let $W = W_{min}$
while $W \leq W_{max}$ **do**
For all $\mathbf{X}^W_{(k,j)}$, where
$\mathbf{X}^W_{(k,j)} = \mathbf{X}_j(:, a_k : b_k)$, $\forall\, b_k - a_k = W$, $a_k < b_k$, $b_k \leq M$, and $a_{k+1} = a_k + f$
where $k = 1, ..., K_W$, and K_W is the number of possible reduced sample matrices
for $k = 1 : K_W$ **do**
$MI^W_{(k_W)} = I(\mathbf{X}^{W_{min}}_{(k,:)}, \mathbf{y})$
end for
$W = W + c$
end while
$M_{k_{MI},W_{MI}} = max(MI)$, where W_{MI} is the window size with highest average mutual information, and k_{MI} contains information about best window of size W_{MI} for all data samples.
The reduced data samples $\tilde{\mathbf{X}}_j = \mathbf{X}^{W_{MI}}_{(k_{MI},j)}$.

information is calculated over all samples $j = 1, ..., L$. The mutual information was first introduced as part of information theory by Shannon in [91]. The mutual information between the new reduced matrix and labels is given by:

$$\begin{aligned} MI^{W_{min}}_{(k,:)} &= I(\mathbf{X}^{W_{min}}_{(k,:)}, \mathbf{y}_:) \\ &= H(\mathbf{y}_:) - H(\mathbf{y}_:|\mathbf{X}^{W_{min}}_{(k,:)}), \end{aligned} \tag{3.1}$$

where

$$H(\mathbf{y}_:) = -\sum_{j=1}^{L} p(y_j) \log p(y_j) \tag{3.2}$$

$$H(\mathbf{y}_:|\mathbf{X}^{W_{min}}_{(k,:)}) = -\sum_{j=1}^{L} \frac{1}{L} \sum_{c=1}^{C} p(y_c|\mathbf{x}_j) \log p(y_c|\mathbf{x}_j). \tag{3.3}$$

The exact conditional probability, $p(.)$, is unknown, but can be estimated using Parzen Window density estimation, i.e.,

$$p(y|\mathbf{x}) = \frac{\sum_{p \in P^y} \exp(-\frac{(\mathbf{x}-\mathbf{x}_p)\Sigma_X^{-1}(\mathbf{x}-\mathbf{x}_p)}{2h^2})}{\sum_{c=1}^{C} \sum_{p \in P^c} \exp(-\frac{(\mathbf{x}-\mathbf{x}_p)\Sigma_X^{-1}(\mathbf{x}-\mathbf{x}_p)}{2h^2})}\ , \tag{3.4}$$

where P^y denotes all samples in class y, P^c denotes all samples in the class $y = c$, Σ is the covarience matrix for each class, and h is the width of the Parzen window.

Next, the size of the window is increased by the change constant c. The default change constant is chosen to be equal to the sample rate (i.e. sampling frequency)[1]. Similarly, all possible combinations of signals of size $W = W_{min} + c$ are found and the mutual information $MI_{(k,:)}^{W_{min}+c}$ between new data matrices $\mathbf{X}_{(i,;)}^{W_{min}+c}$ and emotional labels $\mathbf{y}$ calculated, $k = 1, 2, ...K_{(W_{min}+c)}$. This process is repeated until the window size is greater than or equal to W_{max}. Iterating this process assures that all possible signal time locations are considered.

Finally, choosing the reduced signal matrix is achieved in two steps. First, for the sake of simplicity later on, the window size is chosen to be uniform over all data samples. That is, from all tested window sizes the one with the highest average mutual information is chosen and denoted as W_{MI}. Second, the data matrix, $\mathbf{X}_{q_{MI}}^{W_{MI}}$, where $q_{MI} \in [1, ..., K_{W_MI}]$, for which has the length W_{MI} and the highest mutual information is identified as the reduced data matrix. The reduced data matrix with the highest mutual information is assumed to consist of signals with the greatest emotion intensity, and is chosen for further analysis. The new data samples are given by

$$\tilde{\mathbf{X}}_j = \mathbf{X}_{(k_{MI},j)}^{W_{MI}}, \tag{3.5}$$

for all $j = 1, ..., L$.

3.2 Emotion recognition framework

To classify emotion using EEG signals, a framework as shown in Figure 3.1 is proposed. This framework consists of five steps. First, the raw EEG signals are pre-processed. Second, the proposed mutual information based windowing method is used for signal reduction. The reduced signals are extracted and used in the later steps. Third, the feature extraction aims to reduce the data while still accurately and completely describing the data set. The goal of the next step, feature selection, is to select a subset of relevant features. There are multiple methods that can be used for feature extraction and selection. Finally, the selected features are used to classify the signals.

In the proposed emotion recognition framework pre-processing consists of three steps: artifact removal using *iWASOBI*, bandpass frequency filtering, and averaging to a common reference. The overview of these methods is given in Section 2.2.1.

[1]In this thesis the change constant is chosen to be equal to 128, as the sampling frequency for all used datasets is $128Hz$. However, one can try different options for c, for example multiples of sampling frequency.

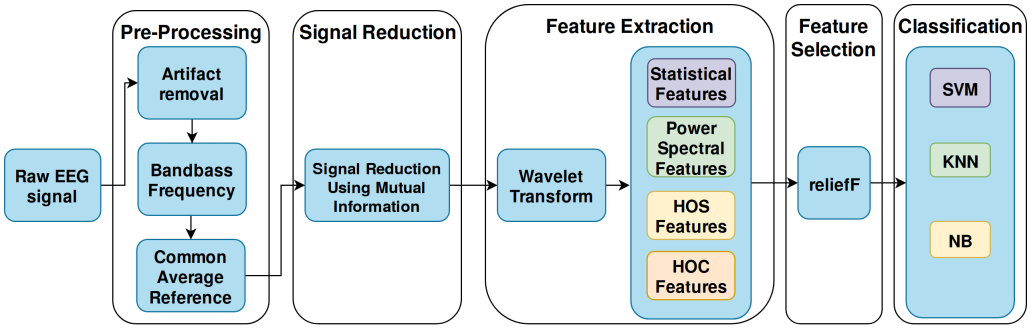


Figure 3.1: The framework proposed for EEG emotion recognition.

After pre-processing, signals are reduced, and the wavelet transform with "db" wavelet is used to extract the different frequency bands of the signal as in Section 2.2.1. Following this, the features are extracted using the methods explained in Section 2.2.2. The statistical features, power spectral features, HOC, and HOS features will be extracted from delta (δ, $0-4$ Hz), theta (θ, $4-8$ Hz), alpha (α, $8-14$ Hz), beta (β, $14-32$ Hz), and gamma (γ, $32-64$ Hz) frequency bands for signals from all 32 electrodes s_i , $i = 1, ..., N$.

For all feature extraction methods, for a single signal the features f_r, where $r = 1, ..., R$ is the number of features extracted using a certain method, can be written as

$$\begin{aligned} FV_{s_i} = &[f_{(1,\delta_{s_i})}, f_{(2,\delta_{s_i})}, ..., f_{(R,\delta_{s_i})}, f_{(1,\theta_{s_i})}, f_{(2,\theta_{s_i})}, ..., f_{(R,\theta_{s_i})}, \\ &f_{(1,\alpha_{s_i})}, f_{(2,\alpha_{s_i})}, ..., f_{(R,\alpha_{s_i})}, f_{(1,\beta_{s_i})}, f_{(2,\beta_{s_i})}, ..., f_{(R,\beta_{s_i})}, \\ &f_{(1,\gamma_{s_i})}, f_{(2,\gamma_{s_i})}, ..., f_{(R,\gamma_{s_i})}] \ . \end{aligned} \tag{3.6}$$

Note that the order of in which the features are extracted is kept the same for all signals (i.e., δ, θ , α, β, and γ), and therefore for each trial, the frequency vector that includes all signals can be written as

$$FV = \ [FV_{s_1}, FV_{s_2}, ..., FV_{s_N}] \ . \tag{3.7}$$

The total number of features depends on the chosen feature extraction method, however for the excisting dataset the number of features extracted are a lot higher than the number of samples. To avoid overfitting, feature selection methods are used to reduce the number of features to be less than (or equal to) the number of samples used for classification. The method used for feature selection in this chapter is *reliefF*. The preliminary tests were also done using mRMR and differential evolution based feature selection, however the reliefF gave the best results. In addition, in [6], overview of mRMR is given together with SVM classifier. In chapter 4 the

mRMR feature selection is explored together with Gaussian process classifiers, where the same conclusion is reached: reliefF gives better results especially for multiclass classification.

Finally, a classification method is chosen. The methods compared in this chapter include SVM, kNN, and NB. These methods were chosen as they are the commonly used classification methods used for emotion recognition. The results using the proposed framework and for three emotion recognition datasets are given in the next section.

3.2.1 Experimental Setup and Parameter Selection

For the model, all three raw datasets were preprocessed following [47] to keep the pre-processing consistent. First, the data was downsampled to 128 Hz, the artifacts were removed using *iwasobi*, a bandpass frequency filter from $4 - 45Hz$ was applied, and the common average reference applied.

For the next step, mutual information based signal reduction, the following parameters were used for all examples. Similarly to [54] and [66], the Parzen window width parameter was chosen to be equal to $\frac{1}{\log(n)}$, where n is the number of samples. The minimum signal window size was chosen to be 5 seconds, as it was decided that shorter signals would not give enough information. To make sure the signals wont be too long, the maximum signal window size was chosen to be 10 seconds, and the change constant c was chosen to be equivalent to 1 second. The level 5 wavelet transform was applied using "db5" wavelet.

The SF, HOC, HOS, and PSD features were extracted. To reduce the amount of features, reliefF feature selection was used with the number nearest neighbours between 2 and 10. Finally, the KNN, SVM, and NB classification was performed using automatic parameter selection on MATLAB.

3.3 Results: Two-class Classification

To see how the proposed signal reduction and the emotion recognition framework perform, three EEG emotion recognition datasets, DEAP [47], MAHNOB [94], and DREAMER [39], are used for classification. First, the two-class classification results are presented for all three datasets separately. The comparison between the results using the reduced signals and full signals is given. In addition, the four different feature extraction methods are compared. A range of features are selected, and it is shown how the number of features will reflect the accuracy of the emotion recognition. Finally, the comparison for three different classifiers is given.

There are a few points to note regarding EEG-based emotion recognition.

First, in this chapter, the concentration lies in subject-dependent emotion recognition. Hence, the classifiers are trained for all participants (subjects) separately. Second, the emotion recognition datasets are relatively small, and the available number of trials to train the classifier for each participant is limited. With a small dataset, avoiding over-fitting can be difficult. To overcome this, the number of selected features is chosen to be smaller than the number of trials. In addition, leave-one-out classification is performed.

The leave-one-out cross validation, uses one observation as validation data, and trains the model using the remaining observations. The two-class classification is performed using all three datasets, and for valence and arousal dimensions separately. The two-classes are positive and negative for both valence and arousal.

Finally, it is noted, that to implement the classification step, MATLAB Statistics and Machine Learning Toolbox is used. This toolbox is used with the aim of finding suitable parameters for the emotion recognition problem.

3.3.1 DEAP dataset

The indepth overview of the DEAP dataset [47] is given in Section 2.3.1. For all examples, the proposed framework using reduced data gives better results than using non-reduced data for both valence and arousal classification.

For both reduced and non-reduced data, four different feature extraction methods are used and compared. To avoid overfitting, features are selected using ReliefF algorithm [85], where the number of features is selected to be between 30 and 39. The classification is performed using different number of features, and the results are compared. Furthermore, three different classification methods are used and compared .

The results for reduced data are shown in Table 3.1 for valence and in Table 3.3 for arousal. Similarly, the results in using entire signals are given in Table 3.2 and Table 3.4 for valence and arousal, respectively. The results are shown for all combinations of feature extraction methods, number of features and classification methods. The best classification method for each feature extraction method and number of features selected are denoted in bold. The highest accuracy per number of features is highlighted in grey. The average for each method is given in the final column with the highest highlighted.

In the following, a detailed overview and discussion of the results is given. The discussion is presented under different feature extraction methods.

Table 3.1: Results on using reduced signals for valence using DEAP dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	79.06	78.89	79.72	82.81	80.39	79.20	78.89	77.87	78.26	80.89	79.60	82.81
kNN	**84.38**	**84.92**	**84.06**	**83.75**	82.66	**84.69**	**84.14**	**84.30**	81.80	82.73	**83.74**	**84.92**
NB	82.03	82.73	82.50	82.73	**83.44**	82.27	82.66	82.73	**82.97**	**82.89**	82.70	83.44

(a) Average accuracy using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	85.55	85.70	85.94	85.55	84.92	85.08	86.56	86.09	87.03	87.11	85.95	87.11
kNN	**89.61**	**89.61**	**89.06**	**89.38**	**88.98**	**88.05**	**88.98**	**88.20**	**88.59**	**89.06**	**88.95**	**89.61**
NB	86.02	86.48	86.95	86.95	87.19	87.03	86.95	87.42	86.88	87.11	86.90	87.42

(b) Average accuracy using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	69.92	71.88	71.17	70.78	70.39	72.73	71.25	70.08	70.78	71.09	71.01	72.73
kNN	**76.17**	73.44	**75.16**	**74.84**	72.19	**75.63**	**74.53**	**74.45**	73.20	72.58	74.22	**76.17**
NB	75.39	**74.61**	74.38	74.61	**74.77**	73.75	74.38	74.06	**73.91**	**73.91**	**74.38**	75.39

(c) Average accuracy using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	75.94	74.49	78.16	77.19	79.22	77.66	77.58	77.97	77.58	76.09	77.19	79.22
kNN	**83.52**	**82.50**	**82.73**	**83.98**	**81.87**	**83.05**	**82.66**	**84.14**	**82.89**	**81.80**	**82.91**	**84.14**
NB	81.25	81.72	81.64	81.17	81.64	81.48	81.88	81.80	80.86	80.94	81.44	81.88

(d) Average accuracy using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.2: Results in using entire signals for valence using DEAP dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	75.68	75.15	71.87	72.09	68.20	70.35	75.92	71.68	70.82	73.36	72.51	75.92
kNN	**77.03**	**77.73**	**77.11**	**78.05**	**77.19**	76.09	**78.44**	76.72	**77.58**	**77.03**	**77.30**	**78.44**
NB	76.33	76.09	75.78	76.17	76.17	**76.33**	76.25	**76.80**	75.94	76.25	76.21	76.80

(a) Average accuracy using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	80.00	81.17	80.23	81.02	81.48	80.78	79.38	80.70	79.38	80.47	80.46	81.48
kNN	**81.41**	**83.98**	**81.88**	**82.42**	**82.81**	**82.66**	**84.38**	**82.66**	81.64	**83.75**	**82.76**	**84.38**
NB	80.78	80.94	81.48	81.33	81.95	81.64	82.58	82.42	**82.50**	83.28	81.89	83.28

(b) Average accuracy using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	67.03	69.14	68.13	66.72	67.42	67.27	65.94	66.64	68.13	67.27	67.37	69.14
kNN	**75.86**	**72.66**	**72.73**	**75.78**	**72.58**	**74.14**	**72.27**	**73.05**	**74.69**	**71.88**	**73.56**	**75.86**
NB	71.80	71.88	71.02	70.63	70.78	69.84	70.23	70.31	68.52	69.53	70.45	71.88

(c) Average accuracy using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	75.89	71.19	75.55	75.98	75.76	75.78	75.55	75.54	72.86	73.67	74.78	75.98
kNN	79.14	78.52	77.89	78.20	**79.77**	76.48	78.28	77.66	**78.91**	**79.84**	78.47	**79.84**
NB	**79.53**	**79.22**	**79.77**	**79.45**	79.69	**79.61**	**79.14**	**78.98**	78.83	79.53	**79.38**	79.77

(d) Average accuracy using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.3: Results in using reduced signals for arousal using DEAP dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	78.24	76.63	80.00	78.88	77.40	77.91	76.39	75.13	75.39	77.94	77.39	80.00
kNN	**82.03**	**81.72**	**81.64**	**81.80**	**80.86**	**80.70**	**81.56**	**80.78**	**81.09**	**81.80**	**81.40**	**82.03**
NB	77.66	77.34	77.50	77.19	77.03	77.42	77.03	76.95	76.64	76.64	77.14	77.66

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	82.97	84.45	83.44	83.83	83.36	83.83	84.77	84.84	84.69	84.61	84.08	84.84
kNN	**89.77**	**89.38**	**89.38**	**89.53**	**89.84**	**89.53**	**88.83**	**89.30**	**89.84**	**89.14**	**89.45**	**89.84**
NB	85.55	85.55	85.47	86.72	86.09	85.63	85.78	86.33	86.17	86.48	85.98	86.72

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	72.66	71.41	72.58	71.64	70.70	71.56	71.56	70.47	69.77	71.56	71.39	72.66
kNN	75.23	75.55	**74.84**	75.70	**76.17**	**75.78**	**76.09**	**75.00**	**74.84**	**74.69**	**75.39**	76.17
NB	**76.33**	**75.70**	74.69	**75.78**	**76.17**	75.16	75.23	74.61	74.38	74.45	75.25	**76.33**

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	AVG	MAX
SVM	78.67	79.33	78.66	78.13	78.33	78.11	78.40	78.40	77.50	84.61	79.01	84.61
kNN	**84.69**	**85.94**	**85.86**	**83.91**	**85.55**	**86.95**	**85.08**	**85.55**	**85.70**	**87.34**	**85.66**	**87.34**
NB	81.48	81.41	81.72	81.56	82.11	81.88	81.41	81.95	82.11	86.41	82.20	86.41

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.4: Results in using entire signals for arousal using DEAP dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	MEAN	MAX
SVM	71.60	67.55	67.34	69.20	68.87	73.26	65.97	67.79	72.24	71.64	69.55	73.26
kNN	**77.03**	**76.80**	**75.63**	**74.53**	**75.78**	**75.47**	**77.11**	**76.72**	**76.56**	**74.53**	**76.02**	**77.11**
NB	73.75	73.91	72.89	73.44	73.36	72.81	73.13	73.36	72.27	73.05	73.20	73.91

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	MEAN	MAX
SVM	80.47	81.09	**81.72**	81.80	81.48	81.41	**82.73**	80.23	81.56	81.09	81.36	82.73
kNN	**83.28**	**83.44**	81.48	**82.73**	**83.83**	**82.42**	82.50	**82.50**	**82.42**	**83.13**	**82.77**	**83.83**
NB	81.80	81.09	**81.72**	81.25	81.17	81.09	80.55	81.17	81.95	80.78	81.26	81.95

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	MEAN	MAX
SVM	71.48	68.59	68.91	67.19	67.97	68.91	68.83	68.59	69.53	68.05	68.81	71.48
kNN	**73.52**	**72.66**	70.63	**72.03**	**71.33**	**72.27**	**75.00**	**72.42**	**72.19**	**72.81**	**72.49**	**75.00**
NB	71.95	71.25	**71.64**	71.25	70.47	71.09	70.78	70.00	71.33	70.23	71.00	71.95

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	30	31	32	33	34	35	36	37	38	39	MEAN	MAX
SVM	75.68	73.80	73.36	74.59	76.64	74.77	70.31	75.41	74.30	77.19	74.61	77.19
kNN	78.13	**80.39**	79.38	**80.23**	**79.22**	**80.39**	**80.00**	**80.39**	**79.14**	**81.25**	**79.85**	**81.25**
NB	**79.45**	78.91	**79.45**	78.75	78.98	78.75	78.59	79.22	78.91	79.45	79.05	79.45

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Higher Order Crossing Feature Extraction

The HOC method was chosen because it was shown to give good results when Ekman's picture set was used in [77]. Using the framework given in section 3.2 where the features are HOC features, the experiments were run using both, reduced signals and all signals.

Table 3.1a shows the results for average accuracy of two-class valence classification using reduced data and HOC feature extraction for three different classifiers and number of features ranging between 30 and 39. The highest accuracy using SVM, namely 82.81%, is achieved using 33 features. The average accuracy over all sets of features is 79.60%. The highest valence classification accuracy is achieved using kNN classifier, where the average accuracy over all number of features is 83.74%. The highest accuracy is reached using 31 features, namely 83.75%. For all different number of features, kNN over-performed the SVM classifier. The third and final method for classification is the NB classifier. NB classifier outperformed both kNN and SVM for 34, 38 **and** 39 features. However, its maximum accuracy over all features and overall average accuracy for valence using reduced data is lower than kNN. Its average accuracy over all features is 82.70%, and maximum accuracy is 83.44%, achieved using 34 features.

The two-class reduced signal arousal classification was performed similarly to valence classification, and the results are shown on Table 3.3a. For classifying arousal, the highest accuracy reached using SVM classifier is 80.00% with 32 features. The overall average accuracy over all number of features is 77.39%. Similarly to valence, two-class arousal classification is highest using kNN classifier. Furthermore, the kNN classifier gives the best results for HOC arousal classification for all number of features. The highest accuracy, 82.03%, is reached with 30 features. Finally, NB classifier is used for classifying arousal, with average accuracy of 77.14% over all number of features. Similarly to kNN classifier, its best result, namely 77.66% is achieved with 30 features. The overall accuracy of arousal is slightly higher when compared to valence, however for arousal kNN classifier achieves the best performance for all number of features.

To compare how reducing the signal affects the accuracy of the emotion recognition framework, all three classifiers were used for both valence and arousal two-class classification using the whole signal and HOC features. Similarly to classifying reduced signals, the features selected range between 30 and 39. Overall, the accuracy reached is considerably higher using reduced signals.

The valence classification results for using entire signals are shown in Table 3.2a. The average accuracy using HOC features and SVM classifier is 72.51% over all number of features, with the highest accuracy, 75.92%, reached using 36

features. The average increase in accuracy is 7.09% when reduced signals are used. Similarly, the optimal results using kNN classifier are also achieved using 36 features. However the accuracy reached was somewhat higher, namely 78.44%. The average accuracy over all features is 77.30% when kNN classifier is used and 76.21% when NB classifier is used. For NB, the maximum accuracy reached is 76.80%. Hence, the use of mutual information windowing for signal reduction results in an average increase of 6.44% in accuracy when using kNN, and an average increase of 6.46% in accuracy when using NB.

The results for classifying arousal using all data and HOC feature extraction can be seen in the Table 3.4a. The SVM classifier resulted in the average two-class classification accuracy of 69.55% when using all data. The highest accuracy reached using entire signals is 73.26%. Similarly for results in using reduced signals , the highest overall accuracy is reached using kNN classifier, however the accuracy is a lot lower. The average accuracy over all number of features is 76.02%, with a maximum of 77.11% using 36 features. Finally, the NB classifier resulted in average accuracy of 73.20%. The maximum accuracy reached is slightly higher than when using SVM to classify arousal, namely 73.91%.

The increase in accuracy is the highest using SVM, 7.84%. Furthermore, reducing the signal resulted in the increase in accuracy of 5.38% using kNN and 3.94% using NB classifiers. For all three classifiers, the increase in accuracy is highly noticeable for both arousal and valence classification.

Statistical Features Feature Extraction

The best overall accuracy, for both reduced data and all data, were achieved using statistical features for both valence and arousal classification. The complete set of results are given for both valence and arousal in Table 3.1b and Table 3.3b for reduced data, and in Table 3.2b and Table 3.4b for entire data.

The highest accuracy reached for valence using SVM classifier and reduced signals is 87.11% with 39 features. The average accuracy reached is a lot higher than using HOC features, namely 85.95%. Using kNN, the highest accuracy of 89.61%, is reached using both 30 and 31 features. The NB classifier performed similarly to SVM classifier, the highest accuracy reached for NB is 87.42%, and the overall average accuracy is 86.90%.

The accuracy of arousal using reduced data and SF are also higher than using HOC features. The average accuracy for all three classifiers are 84.08%, 89.45%, and 85.98% for SVM, kNN and NB, respectively. The maximum accuracy for SVM is achieved using 37 features, and for NB is 33 features, where the accuracy is 84.84% and 86.72%. However, for all number of features, the highest accuracy is achieved

using kNN classifier. The overall maximum, 89.84%, is reached using both 34 and 38 features.

Similarly for HOC results, the accuracy of valence using entire signals is considerably lower. The two-class arousal classification using SF and SVM. On average, the accuracy increases 5.49% when reduced signal is used. The average accuracy using all data and SVM classifier over all features is 81.36%. The optimal accuracy is obtained using 36 features, reaching 82.73%. The kNN classifier reaches higher accuracy than SVM with 83.83%. The average accuracy over all number of features is 82.77%, which is 6.19% lower than using reduced signals. The smallest rise in accuracy is observed when using NB classifier, on average 5.01% over all features. The maximum accuracy reached using all signals and NB classifier is 81.95% using 38 features. Over all features, the mean accuracy observed is 81.26%.

When reduced signals is used to classify arousal, SF gives the best results for all different number of features, however using entire signals two main trends are observed. First, the overall average accuracy for all three classifiers is relatively close together, with the difference between NB (lowest) and kNN (highest) being only 1.51%. When using reduced data, the difference between the highest average accuracy (kNN) and the lowest average accuracy (SVM) reached over 4%. The second noticeable difference between using reduced signals and entire signals is that kNN does not outperform SVM and NB for all number of features. For arousal classification, both SVM and NB outperform kNN using 32 features. In addition SVM also gives better results using 36 features.

For valence classification, the difference between the average results using different classifiers is higher than when classifying arousal, namely 2.3%, whereas when using reduced data, the difference is 3%.

Power Spectral Entropy Feature Extraction

The PSE method was chosen due its good performance in extracting features from EEG for imagined left and right-hand movements in [114]. Our results show that PSE feature extraction is outperformed by all other feature extraction methods we considered, however, the number of features extracted is smallest for PSE method.

The two-class valence classification using PSE features reached the maximum accuracy of 72.73%. This accuracy is reached using 35 features. The lowest accuracy is achieved using 30 features, namely 69.92%. This is also the lowest valence result achieved using reduced data when all feature extraction methods are compared. The average accuracy over all number of features is 71.01%.

Similarly to HOC and SF results, the highest accuracy is obtained using kNN classification. However, the highest average accuracy over all number of features is

reached using NB classifier. The maximum accuracy reached using kNN and NB are 76.17% and 75.39%, respectively. For both classifiers, the maximum accuracy is reached using 30 features. The average accuracy over all features using NB classifier, 74.38, is only marginally better than using kNN classification, where the highest average accuracy over all features is 74.22%. Overall, the two classifiers performed very similarly when classifying valence. The results for two-class reduced signal classification for valence are shown in Table 3.1c.

Arousal classification using reduced signals give very similar results to valence classification, and can be seen in Table 3.3c. Using SVM and PSE features to classify arousal, the average accuracy over all number of features is 71.39%. The maximum accuracy, 72.66%, is reached using 30 features. The kNN and NB classifiers had very similar average accuracy, where kNN performed marginally better. The accuracy is 75.39% for kNN and 75.25% for NB. However, the maximum accuracy of NB outperformed kNN slightly. Similarly to SVM, the maximum accuracy of NB, 76.33%, is achieved using 30 features. The maximum accuracy of kNN is achieved using higher number of features, namely 34. The maximum accuracy of kNN is only marginally lower than NB, reaching 76.17%.

When using PSE features, reducing data does increase the accuracy of the classification, but not as much as using SF or HOC features. When classifying valence using PSE features and entire signals, the average accuracy is 67.37%, 73.56%, and 70.45% for SVM, kNN and NB, respectively. Comparing to using reduced signals, the increase in average accuracy is the lowest for kNN, only 0.66%. The increase in accuracy for SVM and NB is more significant. Reducing the data increases the accuracy of SVM classifier by 3.64% and NB classifier by 3.93%. The highest accuracy classifying valence using entire signals is achieved with kNN classifier, namely 75.86% using 30 features. The SVM and NB classifier achieved the best results using 31 features where the accuracy is 69.14% and 71.88%, respectively. It is interesting to note, that the best results for all three classifiers are achieved using small number of features. Table 3.2c shows the results achieved for all signal valence classification results.

Arousal classification using PSE features, entire signals, and SVM classifier achieved the best performance using 30 features with accuracy of 71.48%. The average accuracy of SVM over all features is 68.81%. This is noticeably lower than when using reduced signals by 2.58%, however the increase in accuracy is smaller than when HOC and SF are used. The highest accuracy and highest average accuracy are achieved for entire signals to classify arousal with kNN classifier. The average accuracy over all features is 72.49%, where the best results, 75.00%, are reached using 36 features. Comparing to reduced data, the whole signal results are on average

2.9% lower. The highest increase in accuracy is achieved using NB when reduced data and whole data is compared. The average accuracy using NB is 71.00% over all features. This is 4.25% lower than using reduced data. The full set of results is shown in Table 3.4c.

It is interesting to note, that using PSE features results in smaller overall increase in accuracy for both valence and arousal than HOC and SF. In addition, considering the small number of features extracted compared to other feature extraction methods, the PSE features achieve good performance.

Higher Order Spectral Feature Extraction

A good accuracy for two-class classification using HOS feature extraction was achieved in [34]. Motivated by this, HOS is chosen as the final feature extraction method used in this chapter. The results using the HOS feature extraction method explained in Section 2.2.2 can be seen in Tables 3.1d and 3.3d for reduced signals and Tables 3.2d and 3.4d for entire signals.

The results in using reduced signals for classifying valence give an average accuracy over all number of features of 77.19% when SVM classifier is used. The maximum accuracy reached using SVM classifier is 79.22%. Again, the highest average accuracy is achieved using kNN classifier. Its accuracy reached maximum of 84.14% using 37 features, while the average over all features is 82.91%. The NB classifier also gives good results, with 81.44% average over all features and maximum of 81.88% using 36 features.

The classification accuracy of arousal (shown in Table 3.3d) is on average higher than classification of valence using reduced signals. The average accuracy over all features reached 79.01%, 85.66%, and 82.20% for SVM, kNN, and NB, respectively. An interesting result is that HOS features achieve better accuracy when using higher number of features. For all the classifiers, the best results have been achieved using 39 features. As for most other methods considered, the highest accuracy using HOS features is also achieved using kNN classification, namely 87.34%. The accuracy of SVM and NB are respectively 84.61% and 86.41%, using 39 features, which are a lot higher than the average accuracy of the method.

Table 3.2d shows the two class classification results using HOS features and entire signals for all three classifiers. It can be seen that when classifying valence, the average accuracy over all features is 74.28% when SVM classifier is used. The maximum accuracy, 75.98%, is achieved using 33 features. The kNN classifier gives the highest maximum result, i.e., for 39 features, the accuracy achieved is 79.84%. However, the highest average accuracy over all features is achieved using NB classifier, namely 79.38%. The maximum achieved accuracy for NB is 79.77%. It can be seen,

that all three classifiers behave very similarly using HOS features.

When comparing two-class valence classification using reduced signals and entire signals, the increase is most noticeable for kNN classifier, namely 4.45%. For SVM and NB classifiers, the increase in accuracy is not as high, but also significant. For SVM reduced signals increase the accuracy on average by 2.41%. The NB classifier achieved the smallest increase in accuracy when using reduced signals, namely 2.06%.

Similarly to results using reduced signal, when classifying arousal, the highest accuracy is achieved using the highest number of features, i.e. 39. The maximum accuracy is 77.19%, 81.25%, and 79.45% for SVM, kNN, and NB, respectively. The highest average accuracy, 79.85% is achieved using kNN classifier. Similarly to valence, the highest increase in accuracy is achieved using kNN classifier, with 5.81% increase in accuracy when reduced signals are used. The SVM gives the accuracy of 74.61% on average over all number of features. Comparing to results using reduced signals, the increase in accuracy is 4.41% when mutual information signal reduction is used. Finally, the smallest increase in accuracy, 3.16%, is achieved when using NB classifier. However, the overall average accuracy is still higher than when using SVM classifier. NB classifier resulted in an average accuracy of 79.05%. All two-class arousal classification results are shown in Table 3.4d.

Comparing the results achieved using reduced signals to entire signals, it can be seen that the reduced signals give better results for both valence and arousal. It is noted that for arousal classification HOS features give better results than HOC features. However, for valence classification, HOC features achieve better performance. In addition, the overall increase in accuracy is higher for arousal when HOS features are used.

3.3.2 MAHNOB dataset

Similarly to DEAP dataset, different classification methods are trained to compare their performance using the reduced signals and entire signals. For all methods, the reduced signals give higher accuracy for both valence and arousal.

Likewise, the classification methods were trained using four different feature extraction methods, including HOC, HOS, SF, and PSE. Using each of these methods, features were selected. Since the number of observations per subject is 20, the number of features selected was between 10 and 19. The classification methods used were SVM, kNN, and NB. All results using reduced signals are shown in Table 3.5 for valence, and Table 3.7 for arousal. In addition, the results using entire signals are given in Table 3.6 for valence and Table 3.8 for arousal. In addition, the accuracy is calculated using leave-one-out cross validation.

Table 3.5: Results in using reduced signals for valence using MAHNOB dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	85.60	85.60	85.00	84.40	85.80	86.60	84.80	84.80	85.40	85.40	85.34	86.60
kNN	**93.60**	**90.00**	**92.20**	**91.60**	**90.60**	**90.80**	**90.00**	**90.60**	**90.20**	**90.40**	**91.00**	**93.60**
NB	87.00	86.80	86.80	88.60	88.40	87.80	87.20	88.00	88.00	88.40	87.70	88.60

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	**93.40**	92.40	93.00	91.20	91.40	90.00	91.60	92.40	93.20	91.40	92.00	93.40
kNN	93.20	**93.20**	**94.00**	**93.00**	**93.80**	**93.60**	**93.60**	**92.80**	**94.60**	**94.40**	**93.62**	**94.60**
NB	91.60	91.80	92.00	91.60	91.80	91.60	91.20	92.40	92.80	93.20	92.00	93.20

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	87.40	86.40	86.40	87.40	87.00	87.40	87.00	89.20	88.00	85.80	87.20	89.20
kNN	**91.20**	**91.40**	**90.20**	**89.40**	89.60	**90.60**	89.40	90.00	89.40	89.80	**90.10**	**91.40**
NB	89.00	90.00	**90.20**	89.20	**90.00**	89.60	**90.40**	**90.20**	**90.40**	**90.40**	89.94	90.40

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	87.40	88.60	89.60	88.20	89.80	89.00	89.40	90.40	90.40	90.60	89.34	90.60
kNN	**91.40**	**92.40**	**91.80**	**90.60**	**91.80**	**93.00**	**91.40**	**93.60**	**93.80**	**93.40**	**92.32**	**93.80**
NB	89.60	88.60	88.40	89.40	88.80	88.00	86.80	88.20	88.00	88.60	88.44	89.60

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.6: Results in using entire signals for valence using MAHNOB dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	62.40	64.00	64.00	63.80	63.80	62.00	63.60	64.60	63.80	62.20	63.42	64.60
kNN	**73.80**	**72.00**	**78.00**	**74.60**	**74.20**	**72.40**	67.80	**73.40**	**72.40**	**73.00**	**73.16**	**78.00**
NB	71.80	71.60	73.00	74.00	72.80	**72.40**	**70.40**	70.60	70.20	70.80	71.76	74.00

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	86.00	86.20	85.20	83.40	84.60	84.00	86.40	87.60	87.20	88.20	85.88	88.20
kNN	**88.20**	**89.20**	**91.00**	**89.20**	**89.60**	**89.20**	**90.40**	**89.20**	**88.60**	**88.80**	**89.34**	**91.00**
NB	85.00	84.20	85.40	84.60	84.80	85.60	85.20	84.80	85.40	85.40	85.04	85.60

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	66.20	68.80	69.60	69.60	68.80	66.80	68.40	65.40	65.60	66.80	67.60	69.60
kNN	**76.20**	**77.40**	**77.60**	**75.60**	**75.40**	**75.20**	**75.00**	**70.80**	**74.40**	**77.40**	**75.50**	**77.60**
NB	71.80	71.80	69.40	71.00	69.80	70.40	69.40	68.60	68.80	68.80	69.98	71.80

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	73.60	72.00	73.00	69.80	70.80	71.40	71.60	69.20	70.20	68.20	70.98	73.60
kNN	**78.40**	**77.80**	**76.20**	**76.80**	**78.00**	**75.80**	71.60	75.00	**76.20**	**75.60**	**76.14**	**78.40**
NB	76.80	75.20	75.20	75.20	75.00	74.80	**73.60**	**75.80**	75.20	73.20	75.00	76.80

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.7: Results in using reduced signals for arousal using MAHNOB dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	87.00	86.60	89.00	87.80	86.80	88.00	87.60	87.20	90.20	88.20	87.84	90.20
kNN	**91.20**	**91.20**	**92.60**	**92.80**	**91.60**	**90.40**	**93.00**	**89.20**	**91.60**	**92.23**	**91.58**	**93.00**
NB	89.60	89.00	89.40	90.20	88.80	89.40	89.00	88.80	89.40	89.40	89.30	90.20

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	89.60	91.40	**92.00**	91.40	91.60	91.00	90.20	91.40	90.60	91.00	91.02	92.00
kNN	**93.60**	91.60	91.40	**93.60**	**93.40**	**93.20**	**92.40**	**92.40**	**94.00**	**92.40**	**92.80**	**94.00**
NB	90.20	**92.40**	90.60	91.00	91.40	90.40	92.00	92.00	93.20	92.00	91.52	93.20

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	87.40	87.40	88.00	86.80	87.40	86.20	85.40	86.40	86.80	85.80	86.76	88.00
kNN	89.60	88.20	87.60	86.80	85.60	88.60	87.60	86.20	85.20	88.40	87.38	89.60
NB	**90.60**	**90.20**	**91.20**	**90.80**	**90.40**	**90.20**	**89.60**	**89.00**	**90.60**	**90.60**	**90.32**	**91.20**

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	88.80	89.40	89.60	89.40	90.20	91.40	90.00	90.40	91.00	**91.80**	90.20	91.80
kNN	**89.60**	**90.27**	**91.00**	**91.40**	**93.60**	**92.60**	**92.00**	**91.00**	**93.60**	91.20	**91.63**	**93.60**
NB	88.40	88.20	88.60	87.40	86.80	88.20	88.20	87.80	89.20	88.40	88.12	89.20

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.8: Results in using entire signals for arousal using MAHNOB dataset for two class classification (Accuracy %). The results highlighted in bold in the feature columns show the highest accuracy over different classifiers per number of features, whereas the results highlighted in bold in the average and maximum columns show the highest overall accuracy over all features and classifiers. Finally, the results highlighted with dark grey show the highest accuracy for specific classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	66.00	66.40	66.80	65.60	66.40	67.00	67.00	69.40	65.80	68.00	66.84	69.40
kNN	**75.40**	**76.00**	**78.80**	**77.60**	**77.40**	**79.00**	**77.60**	**74.80**	**78.00**	**77.93**	**77.25**	**79.00**
NB	74.20	74.60	75.40	74.60	75.00	74.40	74.60	74.60	75.40	74.00	74.68	75.40

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	87.40	87.80	87.40	88.00	88.40	88.40	89.60	**89.00**	87.40	**89.40**	88.28	89.60
kNN	**90.20**	**90.20**	**89.40**	**92.20**	**89.40**	**88.20**	**89.80**	88.80	**88.40**	89.20	**89.58**	**92.20**
NB	86.60	84.80	85.80	83.60	86.20	86.00	85.20	85.40	86.60	86.00	85.62	86.60

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	70.20	70.20	69.80	68.80	67.80	67.40	68.40	67.80	67.80	68.80	68.70	70.20
kNN	**79.80**	**76.60**	**77.60**	**78.00**	**77.60**	**77.60**	**76.40**	**75.80**	**75.40**	**76.20**	**77.10**	**79.80**
NB	75.80	74.40	74.80	76.00	73.80	74.20	74.20	73.60	74.00	70.80	74.16	76.00

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	10	11	12	13	14	15	16	17	18	19	AVG	MAX
SVM	81.20	81.00	84.40	85.80	82.80	86.00	86.60	87.80	85.40	86.40	84.74	87.80
kNN	**89.00**	**86.20**	**85.40**	**88.80**	**86.40**	**86.80**	**88.40**	**88.60**	**89.60**	**90.00**	**87.92**	**90.00**
NB	82.00	81.40	83.00	81.60	82.60	83.20	83.40	83.00	83.00	81.80	82.50	83.40

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Higher Order Crossing Feature Extraction

The results for two-class valence classification using reduced signals and HOC features are shown in Table 3.5a. For SVM, the average accuracy over all number of features is 85.34%. The highest accuracy using SVM classifier is achieved using 15 features, 86.6%. However, the results using SVM for classification with HOC features are very similar over all number of features. The difference in accuracy between the best (using 15 features) and worst (using 13 features) is only 2.2%. The best results are achieved used kNN classifier, that is on average 91.0% over all features, and with maximum of 93.6% using 10 features. For two-class valence classification, the NB classifier is more accurate than SVM but less accurate than kNN. On average over all number of features, the accuracy of NB classifier is 87.7%, with the maximum of 88.6% using 13 features.

The two-class classification for arousal using HOC features is shown in Table 3.7a. In general, the accuracy of arousal classification is higher than valence. For SVM, the accuracy of arousal is 87.84% on average. The maximum is reached for 18 features, with accuracy 90.2%. Similarly to valence, the kNN classifier gives the most accurate results when HOC features are used. The average over all number of features using kNN classification is 91.58%, with the highest accuracy of 93.00% achieved using 16 features. Likewise, the NB classifier performs better than SVM, with accuracy of 89.3% on average and with maximum 90.2%, reached using 13 features.

The accuracy using entire signals is a lot lower than when using reduced signals. The results using entire signals and HOC features for valence are 63.42%, 73.16%, and 71.76% for SVM, kNN and NB, respectively. The maximum accuracy using SVM is 78.0% and achieved using 12 features. On average, the increase in accuracy when using reduced signals is 21.92%. For HOC features this is the largest increase. For kNN, the average increase in accuracy when classifying valence is 17.84% when reduced signals are used. Similarly to using reduced signals, the kNN achieves the highest accuracy when HOC features are used. The smallest increase in accuracy is achieved using NB classifier, namely 15.94% when using reduced signals. The maximum accuracy when using entire signals is 74.0% achieved using 13 features. The complete results are shown in Table 3.6a.

The two-class arousal classification using HOC and entire signals resulted in average accuracy of 66.84% using SVM. The maximum accuracy 69.40% is reached using 17 features. The average increase in accuracy is 21% when reduced signals are used for classifying arousal using SVM. When classifying arousal using entire signals, the highest accuracy was achieved using kNN classifier. The average accuracy over all features is 77.25% and the maximum accuracy using kNN is 79.0%. The

maximum accuracy is reached using 15 features. Overall, using reduced signals increases the accuracy of arousal classification by 14.33%. Finally, the NB classifier results are better than when SVM classifier is used but worse than kNN classifier. The average classification accuracy using NB is 74.68%, with maximum accuracy of 75.4% achieved with using both 12 and 18 features. The average increase in accuracy for NB classification when reduced signals are used is 14.62%. The complete results are shown in Table 3.8a.

Overall, the two-class classification accuracy of arousal is higher than accuracy of valence for MAHNOB dataset using HOC features. More importantly, the accuracy when reduced signals are used is a lot higher than when original signals are used.

Statistical Features Feature Extraction

Similarly to the results on DEAP, statistical features give the best overall results for both all data and reduced data. The results on reduced data are shown in Table 3.5b for valence and Table 3.7b for arousal. Similarly, results in using entire signals are given in Tables 3.6b and 3.8b for valence and arousal, respectively.

For the two-class valence classification using reduced signals (see Table 3.5b), SF and SVM give the average accuracy over all number of features of 92.00%. The maximum accuracy, 93.40% is achieved using 10 features. This is notable as the lowest number of features gives the highest accuracy. The overall best results are achieved using kNN classifier, where the average accuracy achieved is 93.62%. The highest accuracy using SF is achieved with 18 features, namely 94.60%. The kNN classifier outperforms SVM and NB classifiers for number of features, except for 10 features, where the highest accuracy is achieved using SVM classifier. The NB classifier performs very similarly to SVM. The average classification accuracy of NB over all number of features is, similarly to SVM, namely 92.0%. However, the maximum accuracy is slightly lower, i.e., 93.20% achieved using 19 features.

Table 3.7b shows the results for the two-class arousal classification using reduced signals. The SVM classification gives an average accuracy of 91.02% over all features. The highest accuracy, 92.00%, is achieved using 12 features. Similarly to valence classification, the best results when classifying arousal with reduced signals are achieved for kNN classifier. The maximum accuracy for arousal is achieved using 18 features, namely 94.0%, and the overall average accuracy is 92.80%. For arousal, the NB classifier achieved noticeably higher than for SVM. The average accuracy reaches 91.52%, whereas the maximum accuracy is 93.20%. Similarly to kNN classification, the maximum is reached using 18 features.

The two-class valence classification results are less accurate when entire signals are used. The SF results using entire signals to classify valence are shown

in Table 3.6b. The average accuracy using SVM is 85.88% and the maximum accuracy 88.20% is achieved using 19 features. The average increase in accuracy when using SVM and reduced data is 6.1%. Similarly to using reduced signals, the highest accuracy is achieved using kNN classification. However, for kNN the accuracy increase is 4.3% when using reduced signals. For entire signals, the average accuracy is 89.34%, and the maximum accuracy is 91.00%. The NB classifier gives the weakest results when using entire signals, with an average of 85.04% and a maximum of 85.60%. Comparing the results in using entire signals and reduced signals, using the latter results in average increase in accuracy of 7.0% for NB classifier.

The arousal classification using SF and entire signals are shown in Table 3.8b. When using whole data, the average accuracy is 88.28%, 89.58%, and 85.62% for SVM, kNN, and NB respectively. The highest accuracy was achieved using 10 and 13 features and kNN classifier, namely 92.2%. For arousal, using the reduced signals increases the accuracy of SVM classification on average by 2.74%, kNN classification by 3.46%, and NB by 5.9%. This is lower than for valence, but still significant.

When SF and HOC feature extraction are compared, reducing the signal has higher impact on HOC features. In addition, average accuracy using SF has less variation between classifiers. That is, for HOC features, the difference between the highest and lowest averages are over 10%, i.e., highest average 77.25% (using kNN) and lowest average was 66.84% (using SVM). However, when using SF the maximum and minimum average between classifiers are 85.64% (using NB) and 89.58% (using kNN) therefore, the difference between classifiers is less than 5%.

Furthermore, the overall difference in accuracy between using reduced and entire signals is a lot smaller for SF than for HOC features. The lowest increase for HOC was 11.4%, whereas the highest increase in accuracy for SF is 7.8%.

Power Spectral Entropy Feature Extraction

The two-class valence classification using PSE features are shown in Table 3.5c. The highest accuracy achieved using SVM classification is 89.20% using 17 features. The average accuracy using PSE features and SVM is 87.2%. Similarly to HOC and SF, the highest accuracy is achieved using kNN features. The average accuracy reached using kNN classification is 90.10%. The highest accuracy is achieved using 11 features, namely 91.40%. NB classifier gives better results than SVM but worse than the kNN classifier. The average accuracy reached is 89.94% and the maximum accuracy is 90.40%. The maximum accuracy is reached using both 18 and 19 features.

Using reduced signals for two-class arousal classification, the PSE results are the only results that achieved higher accuracy using NB classifier than using kNN classifier. When using SVM and kNN classifiers, the average accuracy of all features

to classify arousal is 86.76% and 87.38% respectively. However, for NB classifier the average accuracy is 90.32%. Similarly to average accuracy, the highest accuracy is observed when using NB classifier. The maximum accuracy using kNN is 89.60% and achieved using 10 features. For both, SVM and NB classifier, the maximum accuracy is achieved using 12 features. For SVM and NB classifiers, the maximum accuracy is 88.00% and 91.20% respectively. The complete results are shown in Table 3.7c.

Table 3.6c shows the whole data results for two class valence classification. The average accuracy over all features is 67.6%, 75.5%, and 69.98% for SVM, kNN, and NB, respectively. Therefore, the average increase in accuracy when using reduced signals over all features is 19.6% for SVM, 14.60% for kNN, and 19.96% NB classifier. In addition, the maximum accuracy of all classifiers is a lot lower when using entire signals. The maximum accuracy 69.60% is achieved using both 12 and 13 features for SVM classifier. For NB, the maximum accuracy of 71.80%, is achieved using 12 features. Finally, the highest maximum accuracy is achieved using 11 features, namely 77.60%.

Similarly to valence, the results classifying arousal are a lot lower when using entire signals. All set of results for arousal using entire signals are shown in Table 3.8c. The average classification accuracy over all features using all data and SVM classification is 68.70%. The maximum accuracy, 70.20%, is achieved using both 10 and 11 features. For all signal arousal classification, the highest accuracy is achieved using kNN classifier. The average accuracy is 77.10% with the maximum of 79.80% achieved using 10 features. Finally, the NB classifier gives an average accuracy of 74.16% when classifying arousal. The maximum accuracy using NB is achieved using 13 features, namely 76.00%.

When comparing results using entire signals to reduced signals, the latter results in 18.06%, 10.28%, and 16.16% increase in accuracy using SVM, kNN, and NB classifiers, respectively. The high increase in accuracy results in NB classifier achieving the highest accuracy with reduced signals.

Higher Order Spectral Feature Extraction

The two-class valence classification using reduced signals and HOS features are shown in Table 3.5d. The SVM average accuracy using HOC features is 89.34%, with a maximum of 90.60% achieved using 19 features. Similarly to all other valence results using reduced signals, the kNN classifier gives the most accurate results. The maximum accuracy using kNN classifier is achieved using 18 features, namely 93.80%. The average accuracy over all number of features is 92.32%. The NB classifier gives the smallest average accuracy compared to SVM and kNN classifiers. The average accuracy using NB over all number of features is 88.44%, with a maximum accuracy

of 89.60% achieved using 10 features.

Table 3.7d shows the results for two-class arousal classification using reduced signals and HOS features. The average accuracy using SVM classification is 90.20% over all number of features. The highest average accuracy is achieved using kNN classifier, namely 91.63%. The lowest average accuracy is achieved using NB classification, 88.12%. The maximum accuracy achieved is 91.80%, 93.60%, and 89.20% for SVM, kNN, and NB, respectively. For both, SVM and NB classifier, the maximum accuracy is achieved using 18 features, however, for kNN, the maximum is achieved using 19 features.

To compare the two-class valence classification performed using entire signals, see Table 3.6d. The average accuracy of valence using entire signals and HOS features is 70.98%, 76.14%, and 75.00% for SVM, kNN, and NB, respectively. Hence, the average increase in accuracy is 13.2% for SVM and 13.44% for NB classifiers. The highest increase in accuracy is achieved using kNN classifier, namely 21.34%.

Similarly to valence, two-class arousal classification is performed using entire signals. The arousal classification gave better results than valence classification using whole data, however even for arousal, the reduced data outperformed the whole signal classification. The complete results are shown in Table 3.8d. Using SVM classifier, the average accuracy over all number of features is 84.74%, with the maximum of 87.80% achieved using 17 features. The highest average accuracy using entire signals, 87.92%, is achieved using kNN classifier, and finally, NB gives the average accuracy of 82.50%.

For all signal arousal classification, the results using entire signals are higher, however the difference between accuracy is not as big as for valence. The increase in accuracy is 5.46% for SVM, 3.707% for kNN, and 5.62% for NB classifier. In general, the HOS features give the second best results for two-class valence classification for both reduced signals and entire signals, outperforming HOC and PSE features.

3.3.3 DREAMER dataset

Finally, the DREAMER dataset was used for two-class classification. This dataset differs from the previous two datasets by having smaller number of subjects and trials per subject. Furthermore, the trials in DREAMER dataset are recorded using a low cost and portable EEG recording device that is able to record 14 channels.

Similarly to previous datasets, the two-class classification was performed and compared using reduced signals and entire signals. The number of features used varied between 14 and 17. The SVM, kNN, and NB classifications were performed using HOC, SF, PSE and HOS features. The results are shown in Tables 3.9 and 3.10 for valence and arousal using reduced signals, and Tables 3.11 and 3.12 for entire

Table 3.9: Results in using reduced signals for valence using DREAMER dataset (Accuracy %).

# of Features	14	15	16	17	AVG	MAX
SVM	73.67	**73.67**	**73.67**	**74.4**	73.85	74.4
kNN	69.32	67.39	71.5	70.05	69.56	71.5
NB	**75.6**	73.43	73.43	**74.4**	**74.22**	**75.6**

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	85.51	86.71	86.23	87.68	86.53	87.68
kNN	**89.37**	**88.65**	**89.61**	**88.89**	**89.13**	**89.61**
NB	82.85	84.54	83.82	84.54	83.93	84.54

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	76.81	76.33	75.12	76.81	76.27	76.81
kNN	73.67	73.43	70.53	75.85	73.37	75.85
NB	**78.99**	**78.99**	**77.05**	**78.26**	**78.32**	**78.99**

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	**77.05**	74.64	74.40	71.98	74.52	77.05
kNN	76.09	76.33	76.33	66.67	73.86	76.33
NB	**77.05**	**77.78**	**79.47**	**78.50**	**78.20**	**79.47**

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

signals.

Higher Order Crossing Features

Table 3.9a shows the results using DREAMER dataset and HOC features. The average accuracy using SVM classifier is 73.85%, with the maximum accuracy 74.4% achieved using 17 features. The average accuracy of kNN classifier is lower than SVM with accuracy 69.56%. When classifying valence using reduced signals, the highest accuracy is achieved using NB classifier. The average accuracy of NB is 74.22% and the maximum accuracy is 75.6% achieved using 14 features.

For two-class arousal classification using reduced signals, the highest average accuracy is achieved using SVM classifier, namely 74.70%. The maximum accuracy

Table 3.10: Results in using reduced signals for arousal using DREAMER dataset (Accuracy %).

# of Features	14	15	16	17	AVG	MAX
SVM	**75.12**	**73.67**	**75.85**	**74.15**	**74.70**	**75.85**
kNN	72.71	72.71	72.95	73.67	73.01	73.67
NB	74.15	73.19	72.22	70.29	72.46	74.15

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	91.30	89.86	**91.30**	**91.79**	**91.06**	91.79
kNN	**92.03**	**90.58**	89.86	90.34	90.70	**92.03**
NB	87.44	84.06	84.30	83.57	84.84	87.44

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	76.09	75.60	78.02	76.09	76.45	78.02
kNN	76.33	**77.78**	**78.99**	76.81	**77.48**	**78.99**
NB	**77.29**	75.36	76.57	**77.29**	76.63	77.29

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	74.40	74.15	69.57	72.95	72.77	74.40
kNN	75.60	77.78	**77.78**	72.22	75.85	77.78
NB	**78.50**	**78.02**	77.29	**78.02**	**77.96**	**78.50**

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.11: Results in using entire signals for valence using DREAMER dataset (Accuracy %).

# of Features	14	15	16	17	AVG	MAX
SVM	56.52	54.83	**56.52**	54.35	55.56	56.52
kNN	52.90	52.17	50.48	51.45	51.75	52.90
NB	**58.70**	**57.00**	**56.52**	**56.52**	**57.19**	**58.70**

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	75.60	**77.05**	**77.54**	**74.88**	**76.27**	**77.54**
kNN	**76.57**	74.15	76.09	72.22	74.76	76.57
NB	74.88	74.88	77.29	74.64	75.42	77.29

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	61.35	62.56	61.84	62.08	61.96	62.56
kNN	64.25	**64.98**	**65.46**	62.56	64.31	65.46
NB	**65.70**	64.73	64.98	**65.70**	**65.28**	**65.70**

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	63.29	63.04	65.46	59.18	62.74	65.46
kNN	68.84	70.29	67.87	71.26	69.57	71.26
NB	**69.81**	**72.46**	**71.74**	**74.64**	**72.16**	**74.64**

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

Table 3.12: Results in using entire signals for arousal using DREAMER dataset (Accuracy %).

# of Features	14	15	16	17	AVG	MAX
SVM	59.66	61.11	61.11	57.49	59.84	61.11
kNN	57.49	56.28	56.52	59.18	57.37	59.18
NB	**63.29**	**63.53**	**63.04**	**64.01**	**63.47**	**64.01**

(a) Average accuracy for arousal using HOC feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	72.22	71.01	70.05	68.84	70.53	72.22
kNN	75.36	74.88	**76.33**	72.46	74.76	76.33
NB	**78.02**	**78.26**	75.60	**76.33**	**77.05**	**78.26**

(b) Average accuracy for arousal using SF feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	61.35	62.08	60.14	61.11	61.17	62.08
kNN	64.98	**68.36**	65.94	**67.15**	66.61	68.36
NB	**68.60**	66.67	**66.18**	66.18	**66.91**	**68.60**

(c) Average accuracy for arousal using PSE feature extraction and SVM, kNN, and NB classifiers.

# of Features	14	15	16	17	AVG	MAX
SVM	68.60	63.77	64.73	63.29	65.10	68.60
kNN	69.81	71.98	71.50	68.84	70.53	71.98
NB	**75.85**	**73.43**	**72.46**	**70.77**	**73.13**	**75.85**

(d) Average accuracy for arousal using HOS feature extraction and SVM, kNN, and NB classifiers.

using SVM, 75.85%, is achieved using 16 features. The average accuracy achieved using kNN and NB is 73.01% and 72.46%, respectively. The maximum accuracy using kNN, 73.67%, is achieved using 17 features, and the maximum accuracy of 74.15% is achieved using 14 features. The complete results are shown in Table 3.10a.

To show that reducing the data improves the classification accuracy, the comparisons were done using all signal data. In Table 3.11a, the two-class classification results using SVM, kNN, and NB classifiers are shown for valence. Comparing to the results in using reduced signals, the accuracy using entire signals is a lot lower for HOC features. The highest accuracy achieved for two-class valence classification, using SVM classifier, is 56.52%. This accuracy is achieved using 16 and 14 features. Using kNN classifier, the highest accuracy is reached using 14 features, namely 52.90%. The highest accuracy of 58.70% is achieved using NB classifier with 14 features. The average accuracy over all features is also the highest for NB, namely 57.19%. The average accuracy over all features for SVM and kNN classifiers only reached 55.56% and 51.75%, respectively.

Similarly to valence classification, two-class classification is performed for arousal using entire signals as shown in Table 3.12a. The average accuracy achieved using SVM, kNN, and NB classifiers are 59.84%, 57.37%, and 63.47%, respectively. The maximum accuracy, 61.11%, is reached for SVM classifier using 15 and 16 features. For both kNN and NB classifier, the maximum accuracy is achieved using 17 features. The highest accuracy reached is 59.18% and 64.01% for kNN and NB classifiers, respectively.

Comparing the reduced signal and all signal results, it can be seen clearly, that the reducing the signal increases the accuracy for all features and classifiers. The increase in accuracy is higher for classifying valence, namely 17.71%, than when classifying arousal, namely 13.16%, when reduced signals are used. The average increase in accuracy using different classifiers is 16.58% for SVM, 16.73% for kNN, and 13.01% for NB.

Statistical Features

Using SF to classify valence using reduced signals, the average accuracy achieved for SVM classifier is 86.53%, and the maximum accuracy is achieved using 17 features. The highest accuracy for SF is achieved using kNN classifier. The average accuracy over all number of features is 89.13%, and the maximum accuracy reached is 89.61% using 16 features. The NB classifier resulted in the lowest accuracy when SF are used. The maximum accuracy using NB classifier is 84.54% achieved using 17 features. The average accuracy using NB classifier is 83.93%. The complete results are shown in Table 3.9b.

The two-class arousal classification results using reduced reduced signals are shown in Table 3.10b. The highest average accuracy is achieved using SVM classifier, namely 91.06%. The average accuracy achieved using kNN and NB classifiers are 90.70% and 84.84%, respectively. The maximum accuracy, 92.03%, is achieved using kNN classifier and 14 features. Using SVM and NB classifiers, the maximum accuracy achieved for two-class arousal classification is 91.79% and 87.44%, respectively. The highest accuracy using SVM classifier is achieved using 17 features. Similarly to kNN, the maximum accuracy for NB is achieved using 14 features.

To compare, the two-class valence classification results using entire signals are shown in Table 3.11b. The average accuracy for valence classification using SVM, kNN and NB classification is 76.27%, 74.76%, and 75.42%, respectively. The highest accuracy, 77.54%, is reached using 16 features and SVM classifier. The highest accuracy using kNN and NB classifiers are achieved using 14 and 16 features, respectively. The maximum accuracy for kNN classifier was 76.57% and the maximum accuracy for NB classifier is 77.29%.

Similarly, the two-class arousal classification results are shown in Table 3.12b. The highest overall accuracy, 78.26%, is achieved using NB classifier and 15 features. For SVM and kNN, the highest accuracy is achieved using 14 features and 16 features, respectively. The highest accuracy is 72.22% using SVM and 76.33% using kNN. The average accuracy when classifying arousal is 70.53%, 74.76%, and 77.05% for SVM, kNN, and NB respectively.

Similarly for HOC features, the increase in accuracy using SF is significant when reduced signals are used. The average increase in accuracy over both valence and arousal, and all number of features is 15.40% using SVM classifier, 15.16% using kNN classifier, and 8.15% using NB classifier. The average increase over all classifiers is higher for arousal, namely 14.76%, whereas for valence the average accuracy is 11.05%.

Power spectral Entropy Features

The PSE features outperform HOC features but worse than SF when classifying valence using reduced signals. All results for PSE classification using reduced data are shown in Table 3.9c. The average accuracy over all number of features is 76.27% using SVM classifier and 73.37% using kNN classifier. The maximum accuracy achieved is 76.81% and 75.85% for SVM and kNN, respectively. For both classifiers, the maximum is reached using 17 features. The highest accuracy is achieved using NB classifier, namely 78.32% on average. The maximum accuracy, 78.99%, is achieved using 15 features.

Table 3.10c shows the arousal classification results using reduced signals and

PSE features. The average accuracy using SVM is 76.45% over all number of features, and the maximum accuracy 78.02% is achieved using 16 features. Similarly to SVM, the highest accuracy using kNN classifier is also achieved using 16 features. The highest accuracy is 78.99% and the average accuracy over all number of features is 77.48%, achieved using kNN. The NB classifier gives the highest accuracy using 14 features, namely 77.29%. The average accuracy using NB classifier is 76.63%.

The two-class valence classification results using a entire signals are shown in Table 3.11c. The average accuracy using SVM classifier and entire signals is 61.96%. Its maximum accuracy, 62.56%, is achieved using 15 features. The average accuracy using kNN classification is 64.31%, with maximum accuracy achieved using 16 features, namely 65.46%. The best results using PSE features and entire signals are achieved using NB classifier with average of 65.28% and maximum of 65.70% with 14 and 17 features.

Similarly to valence, the two-class classification was performed for arousal using entire signals. The average accuracy over all number of features is 61.17%, 66.61%, and 66.91% for SVM, kNN, and NB classifiers, respectively. The maximum accuracy for both SVM and kNN classifiers are achieved using 15 features. The maximum accuracy reached using SVM is 62.08% and using kNN is 68.36%. The highest accuracy is achieved using NB classifier and 14 features, namely 68.60%. The complete results are shown in Table 3.12c.

Similarly to HOC and SF results, the increase in accuracy is significant when reduced signals are used. The increase in accuracy using reduced signals is highest using SVM, namely 14.80%. The increase in accuracy using kNN and NB classifier is 9.96% and 11.38%, respectively. The increase in accuracy is 12.14% for two-class valence classification. The increase is slightly lower for two-class arousal classification, namely 11.96%.

Higher Order Spectral Features

For two-class valence classification using HOS features and reduced signals, the average accuracy is 74.52%, 73.86%, and 78.20% for SVM, kNN, and NB classifiers, respectively. The highest accuracy is reached using 16 features and NB classifier, namely 79.47%. The highest accuracy using SVM classifier, 77.05%, is reached using 14 features. The kNN gives its best results using both 15 and 16 features. The maximum accuracy reached is 76.33%. The complete results are shown in Table 3.9d.

The two-class arousal classification using reduced signals are shown in Table 3.10d. The average accuracy using SVM, kNN, and NB classifier is 72.77%, 75.85%, and 77.96%, respectively. The maximum accuracy is achieved using 14 features for both SVM and NB classifiers. The highest accuracy is 74.40% for SVM

and 78.50% for NB classifier. For kNN classifier, the highest accuracy, 77.78%, is achieved using 16 features.

To compare, the valence results using entire signals are shown in Table 3.11d. When using entire signals, the overall accuracy is lower than when using reduced signals. The average accuracy over all features using SVM classifier is 62.74%, and the maximum accuracy, 65.46%, is achieved using 16 features. The kNN classifier gives better results than SVM classifier. The average accuracy is 69.57%, with maximum accuracy of 71.26%. The maximum is achieved using 17 features. The best results for two-class valence classification are achieved using NB classifier. The average accuracy over all features is 72.16%. The maximum accuracy reached is 74.64% using 17 features.

Similarly to valence classification, two-class arousal classification using HOS features and entire signals was performed. Results in using entire signals for arousal are shown in Table 3.12d. The maximum accuracy using entire signals is 68.60%, 71.98%, and 75.85% for SVM, kNN and NB, respectively. The highest accuracy is reached for SVM and NB classifier using 14 features. For kNN classifier, the highest accuracy is achieved using 15 features. The average accuracy over all features using SVM is 65.10%. For kNN and NB, the average accuracy is higher, namely 70.53% and 73.13%, respectively.

Overall, the classification accuracy is higher using reduced signals. For SVM, the average increase in accuracy is 9.72%. For kNN and NB, the increase is slightly lower, 4.80% and 5.43%, respectively. The overall increase in accuracy when using HOS features and reduced signals is 7.37% for valence and 5.94% for arousal.

3.4 Results: Multi-class classification

In addition to the two-class classification, three-class classification was performed using statistical features and SVM, kNN and NB classifiers. The three-class classification was performed on DEAP and MAHNOB-HCI datasets. Furthermore, five-class classification is performed on DEAP dataset. Due to the small number of samples, DREAMER dataset is not used for multiclass classification.

Similarly to two-class classification, valence and arousal dimensions are considered separately. The class labels for three-class classification are positive, negative, and neutral for both valence and arousal. For five-class classification, the labels can be thought of as positive, semi-positive, neutral, semi-negative, and negative for both valence and arousal.

As every dataset used gives the best classification accuracy using SF extraction method, and the HOC and HOS feature extraction are computationally

more expensive, the multiclass classification results were only performed using SF. Furthermore, for multiclass classification the number of features is chosen to be 39 for DEAP dataset, and 19 for MAHNOB-HCI. The larger number of features were chosen to increase the information that can be used for classification.

3.4.1 Three-class classification

The three-class classification accuracy for valence and arousal on DEAP dataset is given in Table 3.13. Equally high accuracy is achieved when using both SF and kNN and SVM classifier when classifying valence. The average accuracy reached is 77.58% with standard deviation of 7.34 and 7.25 for SVM and kNN, respectively. The NB classifier gives slightly lower accuracy than the other two classifiers, the accuracy obtained using NB is 69.3% with a standard deviation of 8.45%.

Table 3.13: Three-class classification using DEAP dataset (Accuracy %).

Method	Valence	Arousal
SVM	**77.58**	**81.41**
kNN	**77.58**	76.56
NB	69.3	78.59

(a) Average accuracy for valence and arousal using reduced data.

Method	Valence	Arousal
SVM	**56.64**	57.58
kNN	55.16	54.53
NB	56.33	**59.22**

(b) Average accuracy for valence and arousal using all data.

Similarly to classifying valence, three-class arousal classification was performed using reduced signals. The highest accuracy, 81.41%, is achieved using SVM with standard deviation of 8.18%. The accuracy reached using kNN and NB is 76.56% and 78.59%, respectively. The standard deviation in classifying arousal is 8.18%, 7.8%, and 6.25% for SVM, kNN, and NB classifiers, respectively. The results for both valence and arousal using reduced signals are shown in Table 3.13a.

To compare, the results using all data for three-class classification are given in Table 3.13b. The highest accuracy using entire signals to classify valence is 56.64%, achieved using SVM classifier. The kNN and NB classifiers give accuracy of 55.16% and 56.33%, respectively. The standard deviation is noticeably higher using entire signals. The SVM classifier gives standard deviation of 12.22 when classifying valence. The standard deviations of kNN and NB are lower, namely 9.24% and 10.75%, respectively.

The accuracy of using entire signals is slightly higher for arousal using SVM and NB classifiers. The accuracy of three-class arousal classification using SVM is 57.58%. The lowest accuracy using all data is achieved using kNN, namely 54.53%. NB classifier gives the highest accuracy, 59.53%, for arousal. Similarly to valence

classification, the standard deviation is high when classifying arousal. The SVM classifier has the smallest standard deviation of 9.12% when classifying arousal. The standard deviation of kNN and NB are 10.52% and 10.23%, respectively.

The increase in accuracy for valence classification is significant when reduced signals are used for all three classifiers. The highest increase in accuracy is noted using kNN classifier, namely 22.42%, and the smallest increase in accuracy, 12.97%, is noted when using NB classifier. Using reduced signals and SVM resulted in increased accuracy of 20.94%.

The overall increase in accuracy is slightly higher for arousal. The increase in accuracy using reduced signals and SVM classifier is 23.83%. The kNN classifier resulted in 22.03% increase in accuracy. Furthermore, the NB classifier gives 19.37% increase in accuracy using reduced signals to classify arousal.

To verify the results further, MAHNOB-HCI dataset is used for three-class classification. The results are shown in Table 3.14. Similarly to previous results, it can be seen that reducing the signal improves the classification accuracy. The results on valence using reduced signals are 86.2%, 88.8%, and 75.2% for SVM, kNN, and NB, respectively. The SVM has the largest standard deviation of 10.23%, and the smallest standard deviation is achieved using kNN, namely 6.17%. Furthermore, the NB classifier resulted in a standard deviation of 9.63%.

Table 3.14: Three-class classification using MAHNOB-HCI dataset (Accuracy %).

Method	Valence	Arousal
SVM	86.2	86.4
kNN	**88.8**	**87.0**
NB	75.2	76.6

(a) Average accuracy for valence and arousal using reduced data.

Method	Valence	Arousal
SVM	**81.2**	**78.2**
kNN	73.8	75.2
NB	66.8	71.4

(b) Average accuracy for valence and arousal using all data.

The arousal classification using reduced data gives the best results with kNN classifier, namely 87.0%, with standard deviation of 7.5%. The SVM accuracy is 86.4% with standard deviation 9.19, and NB classifier gave the results of 76.6% with standard deviation 8.75%. The complete results are shown in Table 3.14a.

To compare, the results using entire signals with SVM, kNN, and NB classifiers are obtained for both valence and arousal and are shown in Table 3.14b. The accuracy of valence is 81.2% using SVM. The kNN and NB give the accuracy of 73.8% and 66.8%, respectively. The standard deviation using all signal is higher for all classifiers. For SVM, the standard deviation is 9.71%. The highest standard deviation is noted using kNN classifier, namely 13.29%. Finally, the standard deviation using NB classifier is 10.4%.

The arousal classification gives slightly better results than valence classification using entire signals. The accuracy observed using SVM, kNN, and NB classifiers are 78.2%, 75.2%, and 71.4%, respectively. Similarly to valence classification, the standard deviation is noticeably higher using all signals. The standard deviation are 11.08%, 11.75%, and 8.1% for SVM, kNN, and NB classifiers, respectively.

Overall, using the reduced signals increases the three-class classification accuracy for both valence and arousal. The increase in accuracy using SVM to classify valence is 5%. For kNN and NB the increase is higher, namely 15% and 8.4%, respectively. The increase when classifying arousal is slightly lower than for valence. The average increase in accuracy for arousal is 8.2%, 11.8%, and 5.2% for SVM, kNN, and NB, respectively.

3.4.2 Five-class classification

Finally, five-class classification using both reduced signals and entire signals are performed on DEAP dataset. The results are shown in Table 3.15. The highest accuracy for five-class valence classification is achieved using kNN classifier, namely 65.08%. The SVM and NB classifiers reached the accuracy of 62.73% and 54.92%, respectively. The standard deviations are 9.4%, 11.04%, and 7.23% for SVM, kNN, and NB, respectively.

Table 3.15: Five-class classification using DEAP dataset (Accuracy %).

Method	Valence	Arousal
SVM	62.73	**66.64**
kNN	**65.08**	66.33
NB	54.92	59.84

(a) Average accuracy for valence and arousal using reduced data.

Method	Valence	Arousal
SVM	**45.86**	**46.64**
kNN	40.23	43.13
NB	39.06	42.42

(b) Average accuracy for valence and arousal using all data

Overall, the arousal accuracy is little higher then valence accuracy. The highest accuracy for five-class arousal classification is 66.64%, and is achieved using SVM classifier with standard deviation of 10.83%. The kNN classifier gives an accuracy of 66.33% with standard deviation of 5.04%, and the NB classifier gives an accuracy of 59.84% with standard deviation of 7.78%. All results in using reduced signals for both valence and arousal are shown in Table 3.15a.

To compare, the five-class classification is performed using entire signals and the results are shown in Table 3.15b. The accuracy on valence classification using entire signals are 45.86%, 40.23%, and 39.06% for SVM, kNN, and NB, respectively. The standard deviation is highest, namely 11.07%, using NB classifier. The standard

deviations for SVM and kNN are 9.58% and 9.97%, respectively.

Similarly to reduced signals, the results in using entire signals for arousal are slightly higher. The five-class classification for arousal gives accuracy of 46.64% using SVM classifier, 43.13% using kNN classifier, and 42.42% using NB classifier. The standard deviations are 9.58%, 10.12%, and 11.32% for SVM, kNN, and NB, respectively. Table 3.15b shows five-class classification results using all data for both valence and arousal.

Overall, reduced signals give better results for five-class classification than entire signals. For valence classification, the increase in accuracy using reduced signals is 16.87% for SVM, 24.85% for kNN, and 15.86% for NB. In addition, the increase in accuracy in five-class arousal classification is 20.0% using SVM. Furthermore, the increase in accuracy for kNN and NB is 23.2% and 17.42%, respectively.

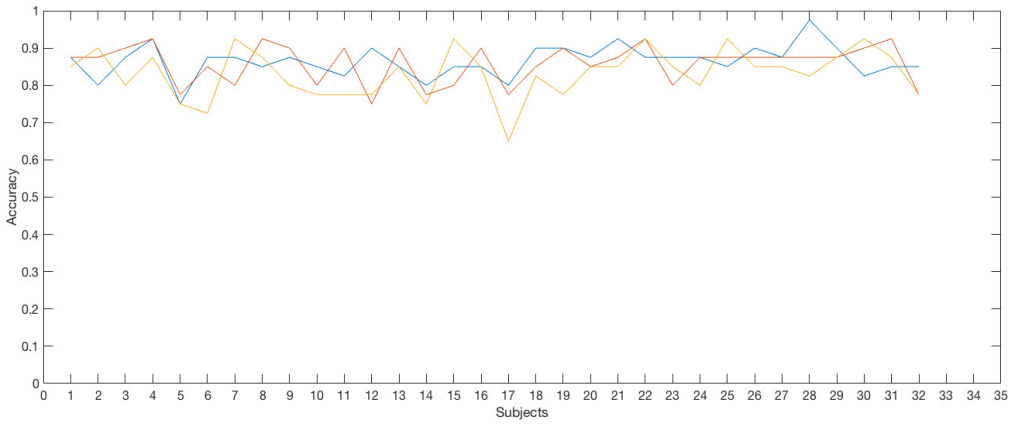
3.5 Analysis and comparison with other emotion recognition systems

The average accuracy for both two-class and multi-class classification has been shown to improve for all sample datasets. Figures 3.2- 3.10 show how the accuracy varies between single subjects for valence and arousal using different classifiers. For all figures, the KNN classifier is denoted in blue, the NB classifies in orange, and the SVM classifier in yellow.

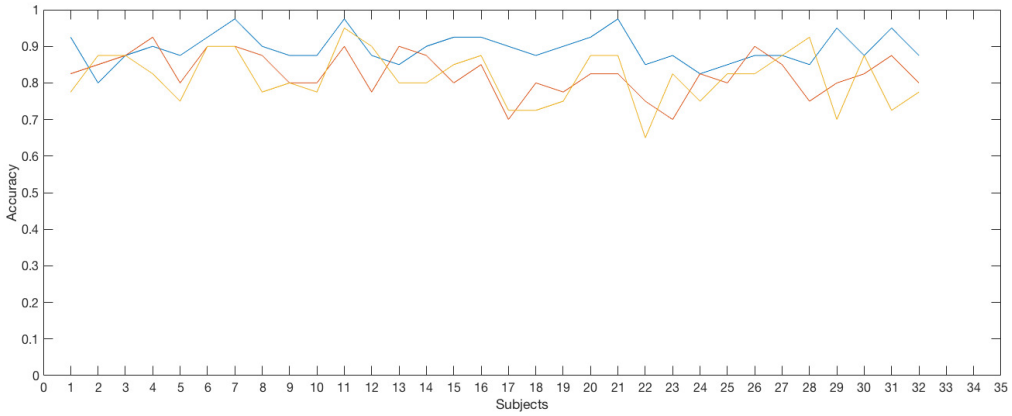
The DEAP dataset had the largest accuracy difference between subjects, especially when dealing with multi-class classification. Figure 3.2 shows the two-class classification for arousal (Figure 3.2a) and valence (Figure 3.2b). For two-class classification, the lowest accuracy 65% was recorded for SVM classifier for both valence and arousal.

As expected, the results for the multi-class classification using DEAP dataset vary more across subjects. On Figure 3.3a, the lowest accuracy for arousal can be seen to be 55% using SVM classifier. However for three-class valence classification, Figure 3.3b, the lowest accuracy, 50%, was recorded for NB classifier. Similarly to three-class classification, five-class five-class classification accuracy for each subject can be seen on Figure 3.4a for arousal and on Figure 3.4b for valence. The lowest accuracy for five-class classification recorded for arousal was 45% using NB classifiers, however, the lowest for valence reached 30% using KNN.

The two-class classification accuracy using MAHNOB dataset is given on Figure 3.5a for arousal and Figure 3.5b for valence. The two-class classification accuracy for both valence and arousal is 80% or above for all subjects. Similarly to DEAP dataset, the three-class classification accuracy varies more across subjects.



(a) Two-class classification accuracy of the proposed model for each subject using DEAP dataset when classifying arousal. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.



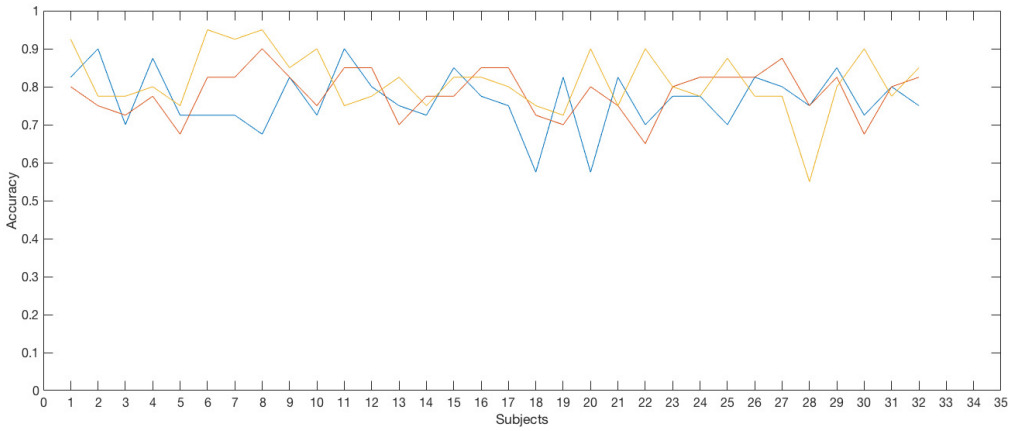
(b) Two-class classification accuracy of the proposed model for each subject using DEAP dataset when classifying valence. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

Figure 3.2: Two-class classification accuracy for DEAP dataset

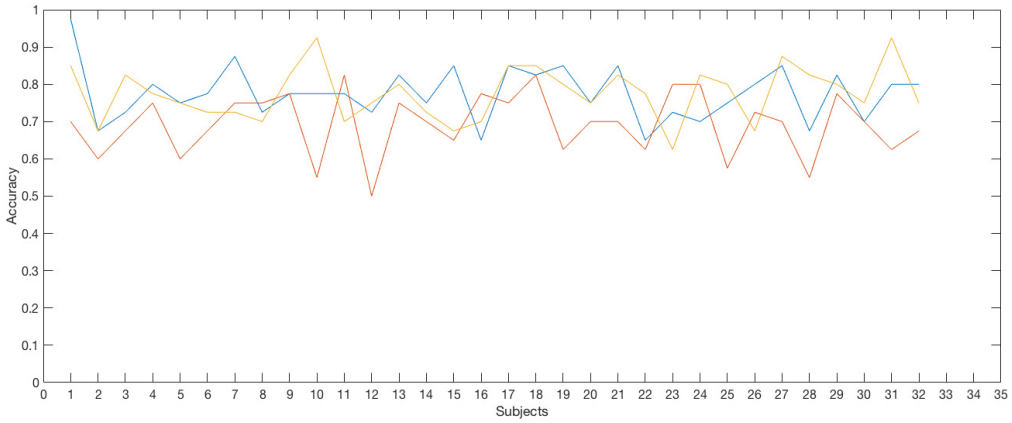
Figure 3.6a shows the accuracy of arousal for each subject. The lowest accuracy recorded is 60% for arousal.The lowest accuracy recorded for valence is 55%, the classification accuracy for all subjects can be see on Figure 3.6b.

Finally, the two-class classification accuracy for every subject separately can be seen on Figure 3.10 using DREAMER dataset. For both, valence and arousal, the lowest recorded accuracy was 80%. The accuracy for valence is shown on Figure 3.7a for arousal and on Figure 3.7b for valence.

Accuracy may not always be the best way to evaluate classification performance, especially when the dataset is imbalanced. To analyze the classifiers even further, receiver operating characteristic (ROC) curve is used to visualize the perfor-



(a) Three-class classification accuracy of the proposed model for each subject using DEAP dataset when classifying arousal. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.



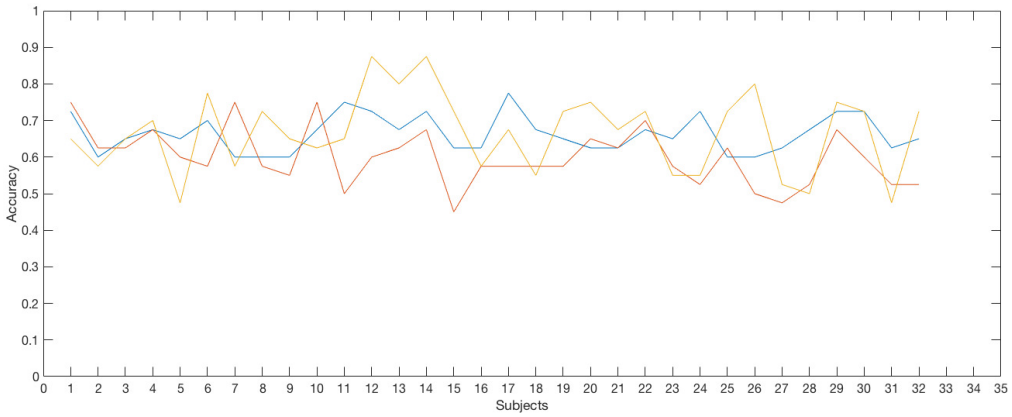
(b) Three-class classification accuracy of the proposed model for each subject using DEAP dataset when classifying valence. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

Figure 3.3: Three-class classification accuracy for DEAP dataset

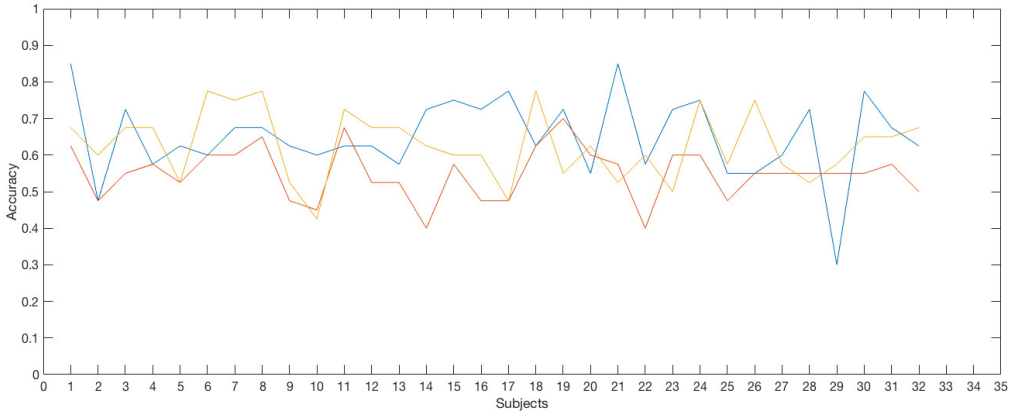
mance of the classifiers. The ROC curve shows classifier performance as a trade off between specificity and sensitivity, giving a good estimate on how well the classifiers separate the classes.

The ROC curves shown for two class classification in Figures 3.8 to 3.7 use statistical features and are used to compare all three classifiers. Using the DEAP dataset the ROC curves in classifying valence and arousal are shown in Figure 3.8a and Figure 3.8b, respectively. Similarly, Figure 3.9a and Figure 3.9b show the curves using the MAHNOB-HCI dataset, and Figure 3.10a and Figure 3.10b show the curves for DREAMER dataset.

These figures show that the ROC curves of all classifiers are very close to



(a) Five-class classification accuracy of the proposed model for each subject using DEAP dataset when classifying arousal. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

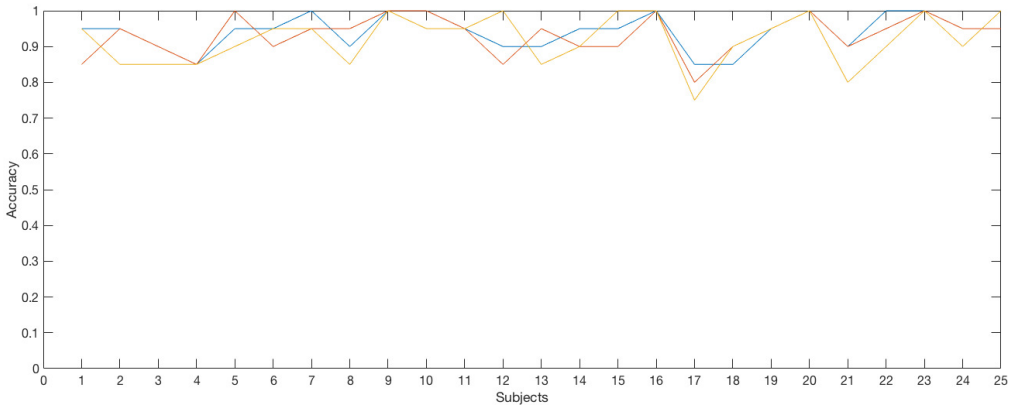


(b) Five-class classification accuracy of the proposed model for each subject using DEAP dataset when classifying valence. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

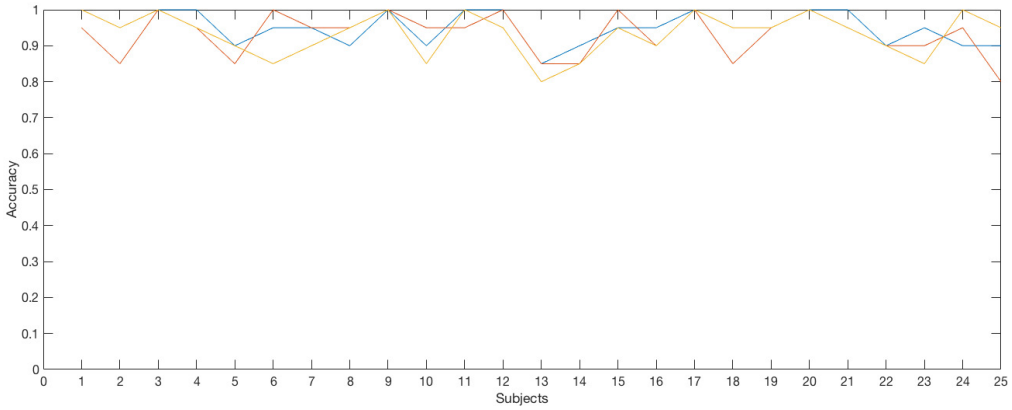
Figure 3.4: Five-class classification accuracy for DEAP dataset

the upper left corner, indicating a good separation between the classes. An area under a ROC curve of 1 represents an ideal test, where the two classes are perfectly separated. However, if the variable under study cannot distinguish between the two groups (there is no difference between the distributions) the area will be equal to 0.5.

For valence classification on DEAP Dataset, the kNN, SVM, and NB classifiers result in the area under the ROC curve of 0.8328, 0.8859, and 0.8481, respectively. The area under the ROC curve for arousal is little higher, namely 0.8897, 0.9102, and 0.9333 for kNN, SVM and NB, respectively. Hence, all three classifiers give very good results for two-class classification. However, even though the accuracy of kNN classifier is higher, SVM and NB classifiers separate the classes better according to



(a) Three-class classification accuracy of the proposed model for each subject using MAHNOB dataset when classifying arousal. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.



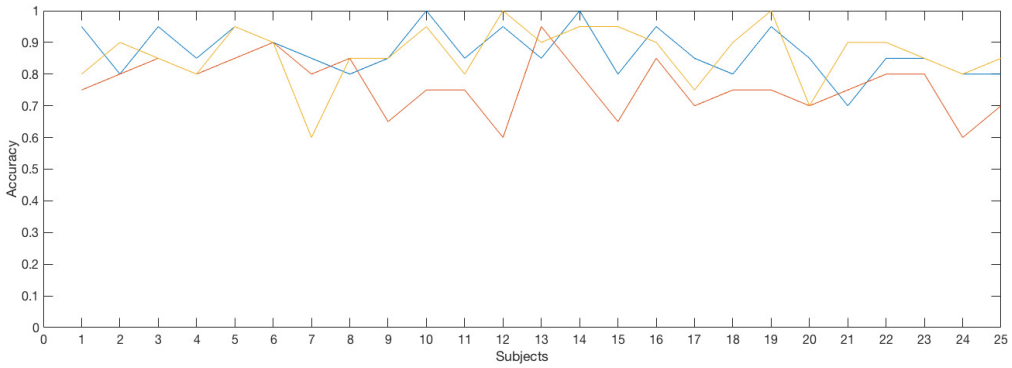
(b) Two-class classification accuracy of the proposed model for each subject using MAHNOB dataset when classifying valence. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

Figure 3.5: Two-class classification accuracy for MAHNOB dataset

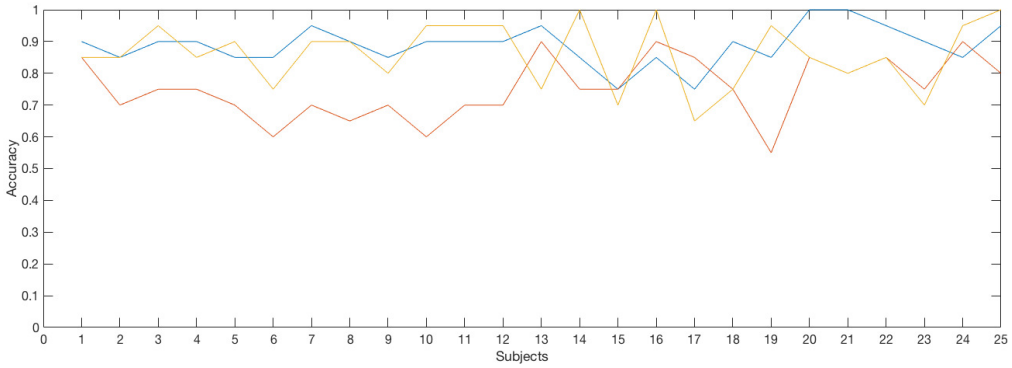
ROC curves.

The results on MAHNOB-HCI results are better than the results on DEAP. For valence the area under the ROC curve is 0.9578 for kNN, 0.9417 for SVM, and 0.9277 for NB. Similarly to DEAP dataset, the arousal results are slightly better. Using statistical features, the kNN, SVM and NB classifiers give the area under the ROC curve of 0.9465, 0.9608, and 0.9222, respectively. In general, when the area under the ROC curve is over 0.8, the classifier is considered as good. For MAHNOB dataset, the area under the ROC curve for all classifiers is above 0.9, and the classification is considered excellent.

Similarly to MAHNOB-HCI dataset, excellent results are achieved using DREAMER dataset for both valence and arousal classification. The area under the



(a) Three-class classification accuracy of the proposed model for each subject using MAHNOB dataset when classifying arousal. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.



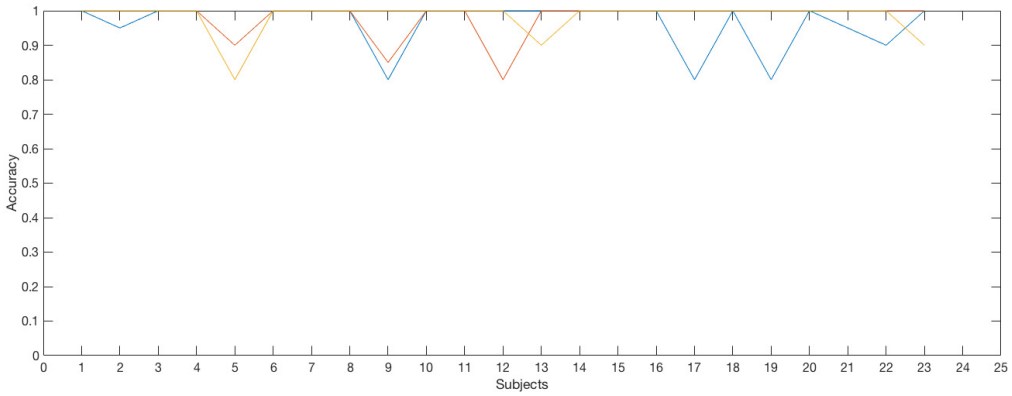
(b) Three-class classification accuracy of the proposed model for each subject using MAHNOB dataset when classifying valence. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

Figure 3.6: Three-class classification accuracy for MAHNOB dataset

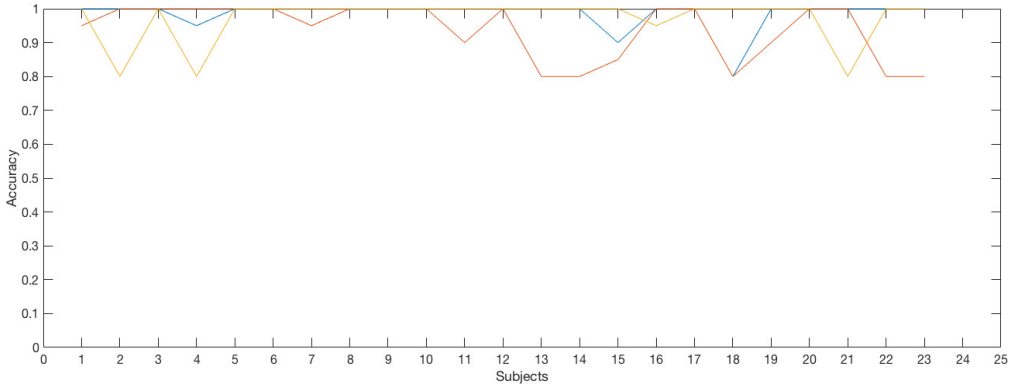
ROC curve is 0.9333 for kNN, 0.9542 for SVM, and 0.8948 for NB classifiers. As before, the arousal results are better, resulting in area under the curve of 0.9455, 0.969, and 0.9105 for kNN, SVM, and NB, respectively.

The ROC curves presented here correspond only to the reduced signal models using statistical feature extraction, as it gives the best accuracy. In addition, the ROC curves were generated for all datasets using all feature extraction and classification methods, and are shown in Appendix A, B, and C for DEAP, MAHNOB, and DREAMER, respectively.

The overall trend as observed from the experimental results indicates that reducing data increases the accuracy of EEG-based emotion recognition. As both DEAP and MAHNOB-HCI are publicly available, many state-of-the-art emotion recognition methods have used these datasets to validate their methods. This makes



(a) Two-class classification accuracy of the proposed model for each subject using DREAMER dataset when classifying arousal. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

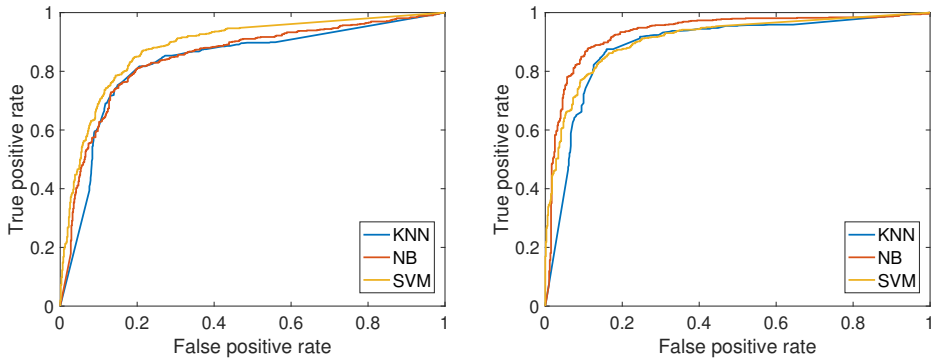


(b) Two-class classification accuracy of the proposed model for each subject using DREAMER dataset when classifying valence. The accuracy of KNN classifier is denoted in blue, the accuracy of NB classifier is denoted in orange, and the accuracy of SVM classifier in yellow.

Figure 3.7: Two-class classification accuracy for DREAMER dataset

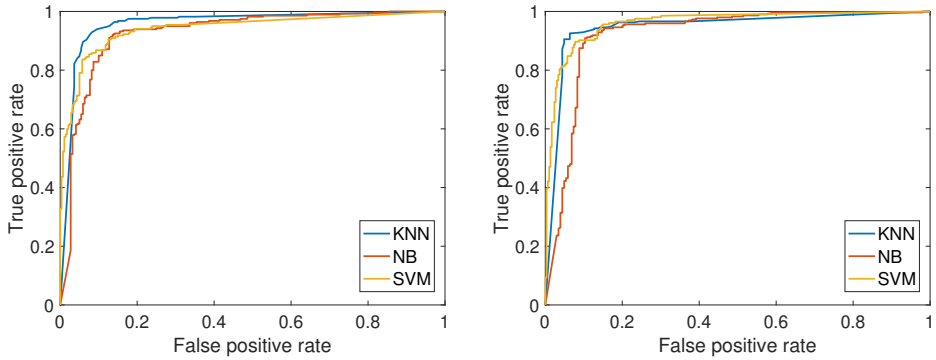
it possible to compare the proposed method with other emotion recognition methods. The results on DEAP dataset compared to other state-of-the-art methods are shown in Table 3.16. All results given in the table are averages over all subjects.

The two-class classification proposed in [6], [86] and [57] have been used to compared to the method proposed in this thesis. All three aforementioned papers use the same dataset (DEAP dataset) making the comparisons more reliable as the data recordning conditions are the same. The mRMR-SVM method for emotion classification proposed in [6] and evaluated on the DEAP dataset give the two-class classification accuracy of 73.14% for valence and 73.06% for arousal. In addition, two-class classification proposed in [86] gives accuracy of 76.9% and 69.1% for valence and arousal, respectively. Furthermore, the Deep Belief Network in [57] gives an



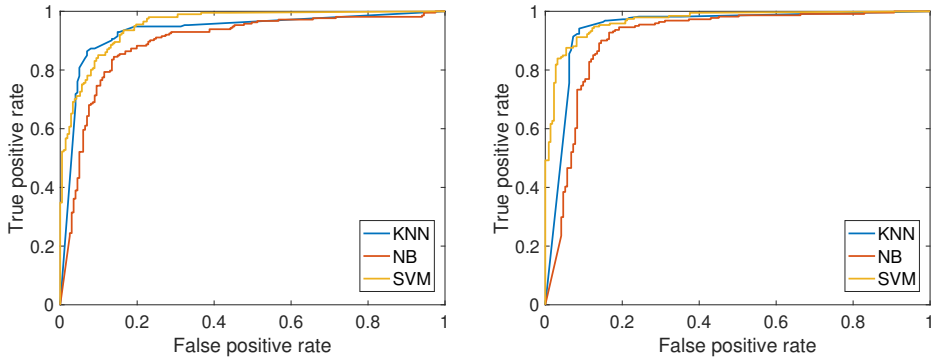
(a) ROC curve for valence classification. (b) ROC curve for arousal classification.

Figure 3.8: Results on DEAP dataset.



(a) ROC curve for valence classification. (b) ROC curve for arousal classification.

Figure 3.9: Results on MAHNOB dataset



(a) ROC curve for valence classification. (b) ROC curve for arousal classification.

Figure 3.10: Results on DREAMER dataset

average valence accuracy of 58.49% and an average arousal accuracy of 64.3%. The method proposed in this thesis achieves accuracy of 89.61% and 89.84% for valence

Table 3.16: Using DEAP dataset to compare proposed method with other state-of-the-art methods.

Method	No. of classes per dimension	Valence	Arousal
Reduced Data, Statistical Features SVM,	2	87.11%	84.84%
	3	**77.58%**	**81.41%**
	5	62.73%	**66.64%**
Reduced Data, Statistical Features kNN,	2	**89.61%**	**89.84%**
	3	**77.58%**	76.56%
	5	**65.08%**	66.33%
Reduced Data, Statistical Features NB,	2	87.42%	86.72%
	3	69.3%	78.59%
	5	54.92%	59.84%
mRMR-SMV, [6]	2	73.14%	73.06%
	3	62.33%	60.70%
	5	45.32%	46.69%
K-PCA, segment-to-response features, RBF-SVM, [86]	2	76.9%	69.1%
Deep Belief Network, [57]	2	58.49%	64.3%

and arousal, respectively, outperforming all other methods.

The three-class classification in [6] gives accuracy of 62.33% for valence and 60.7% for arousal. The three-class classification accuracy using reduced signals, statistical features and kNN classification is 77.58% and 76.56% for valence and arousal, respectively. The SVM gives 77.58 accuracy for valence and 81.41% accuracy for arousal. Furthermore, the three-class classification using NB classifier resulted in valence and arousal accuracy of 69.3% and 78.59%, respectively.

Finally, the five-class classification results achieved in this thesis are compared to the five-class classification results in [6]. In [6], the five-class classification results are 45.32% and 46.69% for valence and arousal, respectively. The method proposed in this thesis gives significantly better results for all three classifiers considered in this thesis. The accuracy for valence is 62.73%, 65.08%, and 54.92% for SVM, kNN, and NB, respectively. The accuracy for arousal is 66.64%, 66.33%, and 59.84% for SVM, kNN, and NB, respectively.

The two-class classification results on MAHNOB-HCI dataset are compared to the results published in [117] and [46] which use the same raw dataset. The results are shown in Table 3.17. The two-class valence accuracy achieved in [117] is 55.72% and accuracy achieved in [46] is 67.5%. The method proposed in this

thesis outperformed both methods. The accuracy of valence achieved in this thesis is 94.6%.

Table 3.17: Classification using MAHNOB-HCI dataset compared to other state-of-the-art methods.

Method	No. of classes per dimension	Valence	Arousal
Reduced Data, Statistical Features, kNN	2	94.6	94.00%
	3	86.00%	87.20%
Power Spectral density, SVM, [117]	2	55.72%	60.23%
Power spectral density, lateralization, Nave Bayes, [46]	2	67.5%	70.0%

Similarly to valence, the proposed method outperforms the other two methods when classifying arousal. In [117], the accuracy achieved is 60.23% and the accuracy achieved in [46] is 70.0%. However, the proposed method achieves an accuracy of 87.2% for arousal.

In addition, comparing the results on DEAP and MAHNOB-HCI dataset, it can be seen that using MAHNOB-HCI dataset gives more superior results than using DEAP dataset. There are two explanations for this. First, DEAP used music videos, whereas MAHNOB-HCI used clips from motion pictures. It can be argued, that the stimuli used in MAHNOB-HCI are more effective in evoking emotional responses. Second, both datasets are noisy and the pre-processing may not have been as effective on DEAP dataset.

The difference between datasets makes it difficult to compare the proposed method to state-of-the-art methods that used different datasets. It will be noted that for the proposed framework, both three-class classification and five-class classification reached 92.50% accuracy for a single subject classification, and the framework reached 99.9% for both DEAP and MAHNOB-HCI dataset for a number of subjects. However, to give a more representative performance of our framework, the results in Sections 3.3 and 3.4 are presented as average accuracy over all subjects. Overall, reducing the EEG data yields significant improvement in emotion recognition. Thus, the proposed mutual information windowing has been shown to be effective.

3.6 Summary

In this chapter, a signal reduction method based on mutual information is proposed. In addition, the proposed method is integrated into subject-dependent emotion recognition framework. Using multiple different EEG emotion recognition datasets, it is shown that signal reduction improves the two-class classification when compared with results using all signal.

In addition, an overview of different feature extraction and classification methods are presented. An in-depth comparison between models using various feature extraction techniques, different number of features and classification methods is given for both reduced signal and all signal for two-class classification. Furthermore, three-class classification is performed on both DEAP and MAHNOB-HCI datasets, and five-class classification is performed on DEAP dataset. It is shown that reducing the signals also improves subject-dependent multiclass classification.

Finally, the proposed method is analysed and compared to other state-of-the-art methods. It is shown that the method proposed gives superior results outperforming other methods that used the same publicly available datasets.

Chapter 4

Subject-dependent and subject-independent classification using Gaussian processes

There are two approaches to EEG-based emotion recognition: subject-independent and subject-dependent. The difference between these approaches lies in how the classifiers are trained. In subject-dependent approach, a new classifier is trained for each subject (participant) separately. Due to the nature of emotions, the subject-dependent approach is used more often and results in higher accuracy. However, for the subject-independent approach, one classifier is trained for all subjects. When trying to classify emotions subject-independently, some loss of accuracy is expected as predictions are often based on an unknown subject. However, being able to accurately classify emotions using subject-independent approach, would open up many new doors, as one could make prediction based on people whose EEG recordings are previously unseen.

Conventional classification methods used for emotion classification include k-nearest neighbours (kNN), support vector machine (SVM), and even neural networks (NN). In addition, transfer learning has been used for EEG-based emotion recognition, especially in the subject-independent case. However, the subject-dependent emotion recognition datasets are often very small, and the conventional methods tend to struggle with very small training sets, particularly when dealing with multi-class classification. In addition, the datasets lie in high dimensional input space, which can cause over-fitting. Furthermore, these methods lack flexibility to capture all nuances of subject-independent classification. In this chapter, it is shown that accurate and

reliable multiclass classification can be performed using small sample size, and high dimensional datasets that are currently available for EEG-based emotion recognition. In addition, it is shown that the subject-independent classification can give accurate results, that outperform some published subject-dependent models.

Gaussian process (GP) classification [81] is introduced in this chapter. It is a non-linear Bayesian classifier that gives probabilistic outputs for classification problems and it has been shown to work well with small amount of data. Although GP classification has been used in multiple areas for classification problems, it is unexplored for EEG-based emotion recognition.

The contributions of the work this chapter are highlighted. In this chapter, GP classification is introduced for EEG-based emotion recognition. The method is explained and compared to conventional classification methods. It is shown that the GP classification has many advantages when compared to a number of other classifiers. The GP classification is first applied to conventional subject-dependent emotion recognition and compared with other state-of-the-art methods to highlight clearly the improvement in accuracy for both two-class classification and multiclass classification. Furthermore, the increase in accuracy is shown for two different features selection methods.

In addition a novel framework is proposed for subject-independent emotion recognition together with increased number of features extracted. The feature extraction proposed makes use of both wavelet transform (WT) and empirical mode decomposition (EMD) to compute the set of statistical features. This thesis shows that using GP classification achieves higher recognition rate for two- and three-class classification when compared to the conventional classifiers for subject-dependent data. Furthermore, the increased number of features proposed for subject-independent classification is shown to give good results compared to different state-of-the-art methods. Using three publicly available EEG-based emotion recognition datasets (DEAP, MAHNOB, and DREAMER) this chapter shows that the proposed framework generalises well.

The chapter is organised as follows. Section 4.1 gives the details of the subject-dependent EEG-based emotion recognition and the methods proposed for the framework. In Section 4.2 the subject-independent emotion recognition framework is introduced. The illustrative examples for emotion recognition using DEAP, MAHNOB-HCI, and DREAMER datasets are given in Section 4.3 for both subject-dependent and subject-independent classification together with comparison and analysis with existing state-of-the-art methods. Finally, the analysis is given in Section 4.4, and summary of the chapter is given in Section 4.5.

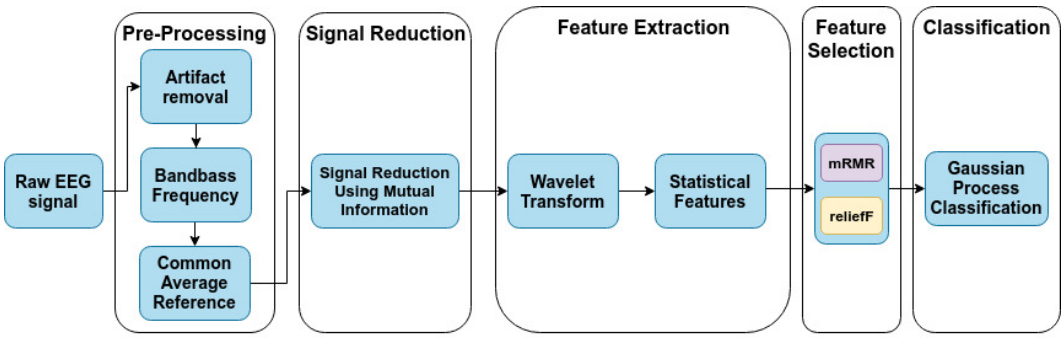


Figure 4.1: Proposed EEG-based Emotion Recognition model used for subject-dependent classification.

4.1 Subject-dependent emotion recognition

The subject-dependent emotion recognition framework proposed is a continuation to the framework proposed in chapter 3 for subject-dependent EEG-based emotion recognition. Similarly to the framework proposed in Section 3.2, the framework proposed in this section has pre-processing, signal reduction, feature extraction, feature selection, and classification steps. Furthermore, the signal reduction is performed using mutual information based signal reduction proposed in Section 3.1. However, in this chapter, the mRMR feature selection is compared to reliefF method (see section 2.2.3), and both feature selection methods are used with GP classification.

The aim of this section is to clearly highlight the usefulness of the GP classification. It is shown, that the GP classifier improves the quality of two-class classification, however, the more important result of GP classification is the increase in multi-class classification accuracy. Being able to classify emotion from very few data samples is an important result.

Figure 4.1 shows an overview of the subject-dependent EEG-based emotion recognition system. The emotion recognition involves EEG signal processing, feature extraction and selection, and GP for binary classification.

The raw EEG signals are first pre-processed, reduced using mutual information windowing method, and the statistical features are extracted. The statistical features extracted are mean, variance, first and second order difference, and normalised first and second order difference. The mRMR and reliefF feature selection methods are then used.

Feature selection is an important step in EEG-based emotion recognition, as the datasets are small, number of input dimensions is high, and most classification models are prone to over-fitting. Feature selection gives a simple solution, by reducing the number of input features. As a general rule, the number of input dimensions

should be smaller than the number of samples.

ReliefF method is a good choice for feature extraction, as is shown in Chapter 3. The ReliefF algorithm has high efficiency and can be applied to both discrete and continuous data. However, the ReliefF algorithm can be computationally expensive. This becomes more important in subject-independent classification, when the number of features is very high. In addition, Relief algorithms can fail to remove redundant features.

The mRMR feature selection method uses both relevance criterion and redundancy criterion, as explained in Section 2.2.3. It checks for the unneeded or similar features, resulting in a new set of features without redundancies. In addition, it is computationally more efficient than ReliefF. However, in certain situations the mRMR algorithm may underestimate the usefulness of features as it has no way to measure interactions between features which can increase relevancy which can lead to poor performance in cases where the features are individually useless, but are useful when combined.

Gaussian process (GP) classification [81] is a non-linear Bayesian classifier that gives probabilistic outputs for classification problems and it has been shown to work well with small amount of data. Although GP classification has been used in multiple areas for classification problems, it is unexplored for EEG-based emotion recognition.

4.1.1 Gaussian processes for binary classification

In this section the signal feature vectors obtained using the above feature extraction techniques will be referred to as input vectors. This to avoid confusion with the GP literature, where the latent functions are distributions over what is referred to as the latent *feature* space.

GPs are nonparametric models that are predominantly used for supervised learning tasks, including classification. Following the Bayesian paradigm, the GP is a prior distribution over the function space, which in the classification setting maps features to class labels. It is a favourable method for classification as it: allows for flexible learning of the classifier through the choice of kernel and its hyperparameters; gives a robust classification model by automatically regularising through the use of prior information (which can be seen as L-2 regularisation); allows for automatic relevance determination which can be used for integrated input selection; and is fully probabilistic, which allows for measures of uncertainty [43, 81, 109].

In GP regression the likelihood is often assumed to be Gaussian. This leads to an analytically tractable marginal likelihood and posterior process over functions. However, this likelihood is only suitable for continuous target data. GP

classification uses alternative likelihoods for discrete labels, most often this is the Bernoulli likelihood. As this results in analytically intractable posterior inference, it is necessary to then use either deterministic approximate inference procedures such as the Laplace approximation, Expectation propagation or variational inference, or stochastic approximate inference procedures such as Monte Carlo.

The binary classification is defined as follows. Consider a training dataset $D = \{\mathbf{X}, \mathbf{y}\}$, where X is a $n \times m$ input matrix, n is the number of samples and m the number of inputs. Hence, $\mathbf{X}$ is a matrix of vector inputs $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$, and $\mathbf{y}$ is the target vector containing sample labels, which in binary classification problem consists of values, $y_i \in \{-1, 1\}$, $i = 1, ..., n$. The aim is to find a function such that when given a new input vector $\mathbf{x}_*$, the model predicts a new label y_*. In the case of classification, the predictions take the form of class probabilities $p(y_* = 1|\mathbf{x}_*, D)$.

GP classification models $p(y|\mathbf{x})$ as a Bernoulli distribution given a fixed $\mathbf{x}$ [53]. The success probability, $p(y = 1|\mathbf{x})$, is related to an unconstrained latent function $f(\mathbf{x})$ which is mapped to the unit interval by a sigmoid transformation. Let $p(y = 1|\mathbf{x}) = \text{sig}(f(\mathbf{x}))$, where $\text{sig}(\cdot)$ is a sigmoid function. The most common sigmoid functions used for GP classification are *probit* and *logistic* functions. The latter, $\text{sig}_{logit}(z) = (1 + \exp(z))^{-1}$, is used in this work.

In GP classification models, Bayesian inference is performed on the latent function $f(\cdot)$. Let $f_i = f(\mathbf{x}_i)$, and $\mathbf{f} = [f_1, ..., f_n]^T$ denote the realizations of latent functions. Given these latent functions, the target values are independent Bernoulli variables. Thus, the likelihood depends on the latent function f only through the observed inputs, and is given by

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i). \tag{4.1}$$

Without loss of generality, a zero-mean GP prior over the latent function $f(\cdot)$ is used. The choice of positive definite covariance function $k(\cdot, \cdot|\boldsymbol{\theta})$ reflects the family of functions that are modelled, where $\boldsymbol{\theta}$ is the set of kernel hyperparameters whose variability further controls the behaviour of the functions modelled and must be inferred from the data. The covariance function can be defined by elements, where each element is written with the shorthand notation $k_{ij} = k(\mathbf{x_i}, \mathbf{x_j}|\boldsymbol{\theta})$, and $\mathbf{K}$ is the Gram matrix that denotes the kernel function evaluated across all training samples. The posterior distribution over latent function values $\mathbf{f}$ at the observed $\mathbf{x}$ for given hyper parameters $\boldsymbol{\theta}$ becomes

$$p(\mathbf{f}|D, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{p(y|\mathbf{X}, \boldsymbol{\theta})}. \tag{4.2}$$

As stated before, the aim of the classification problem is to be able to predict a label y_* for given new test input $\mathbf{x}_*$. To find the predictive distribution, first the latent function distribution is computed by marginalisation:

$$p(f_*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}_*) = \int p(f_*|\mathbf{f}, \mathbf{X}, \boldsymbol{\theta}, \mathbf{x}_*) p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}) \, \mathrm{d}\mathbf{f}. \tag{4.3}$$

This is followed by computing the expectation, giving the predictive distribution (which is a Bernoulli distribution)

$$p(y_*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}) = \int p(y_*|f_*) p(f_*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}_*) \, \mathrm{d}f_*. \tag{4.4}$$

The approximations are needed, as the posterior distribution $p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta})$, marginal likelihood $p(\mathcal{D}|\boldsymbol{\theta})$, and predictive distribution $p(y_*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x})$ are all analytically intractable.

The main approximation approaches used are the Laplace's approximation (LA), Expectation propagation (EP), variational approximations and Markov chain Monte Carlo (MCMC) sampling. An overview of using EP, variational approximation, and MCMC methods can be found in [68, 71]. LA is used due to its fast running time.

The LA is used in this paper is based on the Gaussian approximation to the posterior, i.e., $q(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{A})$. Following this, approximate Gaussian posterior is introduced to (4.3), giving rise to the approximate posterior $q(f_*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}_*) = \mathcal{N}(f_*|\mu_*, \sigma_*^2)$, where mean μ_* and variance σ_* are given by

$$\mu_* = \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{m} \tag{4.5}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_* - k_*^T (\mathbf{K}^{-1} - \mathbf{K}^{-1} \mathbf{A} \mathbf{K}^{-1}) \mathbf{k}_*. \tag{4.6}$$

More in-depth discussion of GP classification and its approximations can be found in [53, 71, 81]. Here LA is used to find $\mathbf{m}$ and $\mathbf{A}$. LA is based on the second order Taylor approximation of the unnormalised log posterior [81]. This approximation works by placing the mean $\mathbf{m}$ at the mode (MAP estimate) and equating the covariance $\mathbf{A}$ to the negative Hessian of the log posterior density at $\mathbf{m}$.

The computational complexity of GP classification is higher than other classification methods looked in this thesis, namely $O(n^3)$. However, the increase in accuracy will also be shown to higher, especially for multi-class classification.

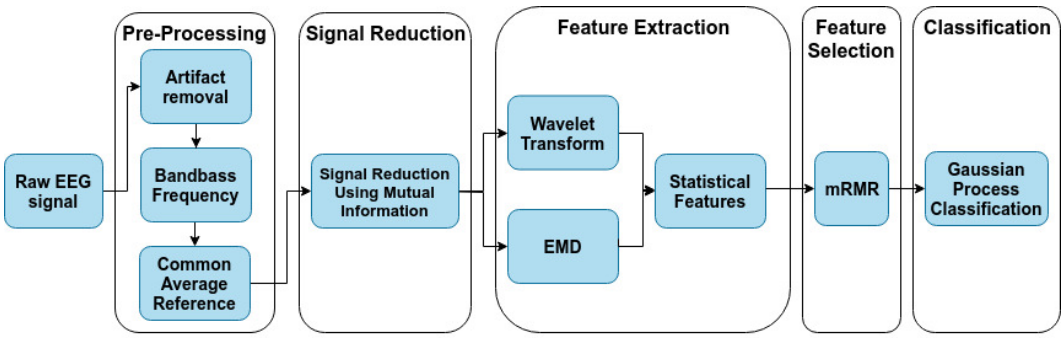


Figure 4.2: Proposed EEG-based Emotion Recognition model used for subject-independent classification.

4.2 Subject-independent framework

For the subject-independent framework (shown in Figure 4.2), the pre-processing, feature selection and classification steps are the same as in subject-dependent framework. However, to make accurate predictions in the subject-independent case additional features are needed. Hence, additional measures are calculated in the feature extraction step.

Feature extraction

For subject-independent classification, the reduced signal is split into evenly spaced sections and the features are calculated separately from each section. The features extracted for the subject-independent classification can be considered as four independent sets of features. The first set of statistical features are extracted from the frequency decompositions and the second set is extracted from EMD. In addition, power spectral density (PSD) is calculated from frequency decompositions and EMD. Extracting the statistical features from the PSD gives the feature sets three and four. A schematic overview of these features is shown in Figure 4.3.

The frequency decomposition into α, β, γ and δ waves is performed using wavelet transforms similarly to the subject-dependent case. However, instead of "db4" wavelet, "sym8" wavelet is used to capture the chaotic nature of the EEG signal. Additional features are found using EMD. EMD is used to decompose the signal in the time domain, and the statistical features extracted from IMFs.

Furthermore, the PSD is calculated from α, β, γ and δ frequency bands, and IMFs calculated using EMD. PSD describes the power present in the signal as a function of frequency.

To create the final feature vector, the statistical features are calculated from the frequency bands and IMFs as well as from the PSD of the above-mentioned

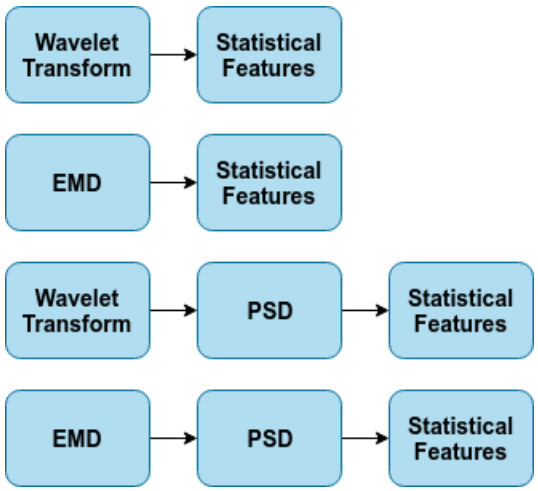


Figure 4.3: Feature extraction for subject-independent framework.

frequency bands and PSD of the IMFs. The statistical features are those used in Section 4.1 with the addition of minimum and maximum of amplitudes of the signal, mode, mean normalised frequency, mean excluding outliers, mean absolute deviation, skewness, entropy, mobility, complexity, occupied bandwidth, and peak magnitude to RMS ratio of the signal. Robust estimates of mean and variance instead of the exact mean and variance of the signals have been used as features. Furthermore, the residual error of the signal against the modelled changes has been calculated and used as a feature.

4.3 Experimental results

The experimental results are obtained for both subject-dependent and subject-independent classification using DEAP, MAHNOB-HCI and DREAMER datasets. The classification is performed for valence and arousal separately.

For subject-dependent classification, data labelling is also performed subject dependently. All datasets used include the participants self assessment ratings. Using this, the two-class subject-dependent classification is performed by splitting the trials as positive and negative for both valence and arousal. Similarly, for three-class classification, the trials are labelled positive, neutral, and negative. Finally, five-class classification uses positive, positive-neutral, neutral, negative-neutral, and negative classes.

The subject-dependent two-class classification is performed for all three datasets. The three-class classification is performed for MAHNOB and DEAP, and five-class classification is only performed using DEAP dataset. This is due to the size of the DREAMER and MAHNOB dataset being small, i.e., the number of trials per

subject is low. However, as all trials can be used together for subject-independent classification, the two and three-class classification is performed for all three datasets.

It is shown that the two-class subject-dependent classification results in a small increase in accuracy. A lot of the state-of-the art methods can successfully classify emotion into two classes. However, the importance of this work lies in multi-class classification results. It is shown that using GP classification for three and five-class classification increases the accuracy significantly.

In our framework, the GP classification model is implemented using LA. The kernel used for all classification is a combined Gaussian and Rational Quadratic kernel with added white noise. The combined kernel was chosen to provide flexibility and robustness to the model.

4.3.1 Results for subject-dependent emotion recognition

In the subject-dependent emotion recognition, the classification model was trained for all participants separately. As the number of trials per subject is small, leave-one-out cross validation was performed. The results are compared to the models using the same datasets. For all methods, the average accuracy over all subjects is given.

Results on DEAP dataset

Table 4.1 shows that GP classification using statistical features on DEAP performed better than the methods in [6], in [104], and [80]. In [80], the best results are found when using statistical features and kNN classification. Using the same framework up to the classification stage shows that GP classification increases the accuracy, raising both valence and arousal classification to over 90% on DEAP.

The average accuracy for valence and arousal using mRMR feature selection is 91.25% and 92.66%. The respective variance is 3.61 for valence and 3.76 for arousal. Similarly reliefF feature selection is used with GP classification. The accuracy of valence is 91.64% with variance of 5.14. The accuracy of arousal using reliefF feature selection is 90.62% with variance 4.98.

For two-class classification, the high accuracy can be reached quite easily using multiple different methods, and therefore the increase in accuracy for two-class classification is not extremely high. To validate the choice of GP classification and proposed model even further, multiclass classification is performed.

Tables 4.2 and 4.3 show three and five-class classification results using GP classification. The accuracy of three-class classification with mRMR feature selection is 81.71% for valence and 81.64% for arousal. The respective variance is 9.71 and 6.30. Furthermore, using reliefF feature selection, the accuracy of valence and arousal is 86.64% and 86.09%, with variance 3.59 and 6.90 respectively.

Table 4.1: Subject-dependent results using DEAP dataset for two-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification (SF+mRMR)	2	**91.25%**	**92.66%**
GP classification (SF+reliefF)	2	**91.64%**	**90.62%**
mRMR-SVM [6]	2	73.14%	73.06%
CNN Model [104]	2	81.406%	73.36%
SF-kNN [80]	2	89.61%	89.84%

Table 4.2: Subject-dependent results using DEAP dataset for three-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification (SF+mRMR)	3	**81.71%**	**81.64%**
GP classification (SF+reliefF)	3	**86.64%**	**86.09%**
mRMR-SVM [6]	3	62.33 %	60.7%
SF-kNN [80]	3	77.58 %	81.41%

Similarly to two and three-class classification, the accuracy of five-class classification is higher using GP classifier. The increase in accuracy is not as high using mRMR feature selection. The accuracy of valence is 65.78% and accuracy of arousal is 67.11% with respective variances of 19.19 and 13.68. However, using reliefF classifier, the accuracy of five-class classification is 72.42% with variance 9.06 for valence and 75.00% with variance 10.58 for arousal.

Results on MAHNOB-HCI dataset

To validate the results even further, the subject-dependent two and three-class classification model is trained for MAHNOB-HCI. The two-class classification results are shown in table 4.4. The comparison is made using kNN and statistical features. Even though kNN gave accurate results with 94.6% and 94.0% respectively for valence and arousal in [80], using GP classification and mRMR feature selection, the accuracy of valence is 97.00% and the accuracy of arousal 98.40%. The variance is 0.75 for valence and 0.56 for arousal.

Similarly using GP classification and reliefF feature selection gives an accuracy of 96.60% for two-class valence classification and accuracy of 95.20% for

Table 4.3: Subject-dependent results using DEAP dataset for five-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification (SF+mRMR)	5	**65.78%**	**67.11%**
GP classification (SF+reliefF)	5	**72.42 %**	**75.00%**
mRMR-SVM [6]	5	45.32 %	46.69%
SF-kNN [80]	5	65.08 %	66.64%

two-class arousal classification. The variance is 0.47 and 1.29 for valence and arousal, respectively.

The three-class classification results are shown in Table 4.5. Similarly to DEAP, the increase in accuracy is higher for three-class classification. The accuracy of valence and arousal is 91.20% and 90.60% respectively using GP classification and mRMR feature selection. The respective variance is 1.11 and 1.19. The accuracy using GP classification and reliefF features selection is 92.60% and 90.60% for valence and arousal respectively. The variance for valence is 1.09 and the variance for arousal is 2.61.

Table 4.4: Subject-dependent results using MAHNOB-HCI dataset for two-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification (SF+mRMR)	2	**97.00 %**	**98.40%**
GP classification (SF+reliefF)	2	96.60 %	95.20%
kNN [80]	2	94.60%	94.0%

Results on DREAMER dataset

Finally, the two-class classification is performed using DREAMER dataset. Similarly to the other datasets, there is an increase in accuracy in two-class classification. Using GP classification with mRMR feature selection results in accuracy of 94.93% for valence and 96.86% for arousal. The variance for valence is 0.81 and for arousal is 0.62. Furthermore, the increase in accuracy is noticeable using reliefF feature selection. The accuracy for valence is 92.51% and accuracy for arousal is 94.44%, with variances 0.69 and 0.81 for valence and arousal respectively.

Table 4.5: Subject-dependent results using MAHNOB-HCI dataset for three-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification (SF+mRMR)	3	**91.20 %**	**90.60%**
GP classification (SF+reliefF)	3	92.60 %	90.60%
kNN [80]	3	86.00%	87.20%

Table 4.6: Subject-dependent results using DREAMER dataset for two-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP Classification (SF+mRMR)	2	**94.93%**	**96.86%**
GP Classification (SF+reliefF)	2	92.51 %	94.44%
kNN [80]	2	89.61%	92.03%

4.3.2 Results for subject-independent emotion recognition

For all three datasets, the subject-independent classification is performed by combining the data from all subjects. This data is then split into training and testing sets, where 80% of the data is used for training and the remaining 20% is used for testing.

Splitting the dataset has been done randomly, following which the classifier is trained. This process is repeated 500 times, with changing testing and training sets. The average accuracy reported is the accuracy to which the classifier converges. This is to show that the proposed method is indeed robust and can deal with different inputs, while still give accurate results. Furthermore, for accurate comparison of classifiers, the benchmarks are run using SVM and kNN classification. Similarly to the subject-dependent classification, the classification is performed for valence and arousal separately.

Subject-independent emotion recognition using DEAP dataset

Combining the data from all 32 participants gives 1280 trials in total when using DEAP dataset, making the dataset big enough to split it into training and testing sets. Using the $80-20$ split, the training set has the size of 1024 trials, and the remaining 256 trials are used as the test set. Every time the model is trained the data

is randomly split, giving a fresh training and testing data. The accuracy presented is the average accuracy over all trials.

The two-class subject-independent classification (see Table 4.7) reaches the accuracy of 80.50% for valence and 81.07% for arousal. These results are achieved using GP classification. Using SVM and kNN classifier, the accuracy of valence was 78.78% and 76.80% respectively. The accuracy for arousal using SVM and kNN classifiers are lower, 78.64% and 72.48%, respectively. Similarly to GP classification, the SVM and kNN classification are repeated by randomly splitting the whole dataset into training and testing sets.

Table 4.8 shows the subject-independent three-class classification. It shows, that the most accurate results are achieved using GP classification. The accuracy for valence is 69.5 and the accuracy for arousal is 67.00%. As a comparison, the accuracy of SVM is 68.58% for valence and 65.63% for arousal. The kNN classifier is the weakest, with accuracy of 66.58% and 50.28% for valence and arousal, respectively.

Table 4.7: Subject-independent results using DEAP dataset two-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification	2	**80.50%**	**81.07%**
SVM	2	78.78%	78.46%
kNN	2	76.80%	72.48%

Table 4.8: Subject-independent results using DEAP dataset three-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification	3	**69.50%**	**67.00%**
SVM	3	68.58%	65.63%
kNN	3	66.58%	50.28%

The convergence of two-class classification training means together with 85% and 95% confidence intervals are shown in Figure 4.4 for valence and in Figure 4.5 for arousal, respectively. The confidence interval represents the margin of error, or the amount of uncertainty, around the point estimate (mean). That is, the 85% confidence interval defines a range of values that one can be 85% certain to contain the population mean. For two-class valence, the 85% confidence interval is between 78.19% and 82.73%, and the 95% confidence interval is between 76.99% and 83.94%. Furthermore, the 85% confidence interval for two-class arousal is between 79.38% and 83.83%, and the 95% confidence interval lies between 78.21% and 85.00%.

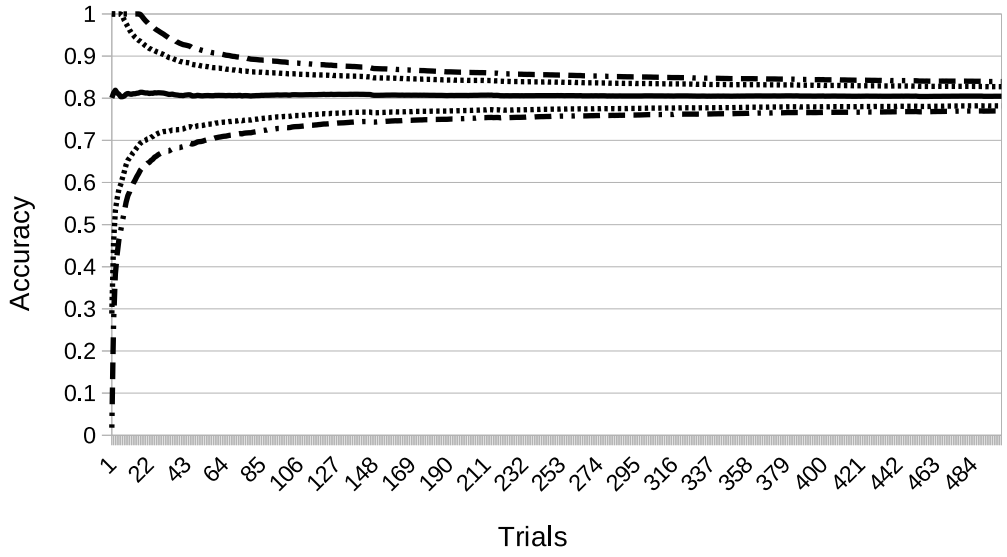


Figure 4.4: The two-class classification of DEAP valence. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

Similarly, Figure 4.6 and Figure 4.7 show the mean accuracy and the 85% and 95% confidence intervals for three-class classification. The confidence for three-class classification is slightly lower, as expected. The 85% confidence interval of three-class valence classification is between 64.44% and 74.05%, and the 95% confidence interval is between 62.84% and 71.09%.

Subject-Independent results using MAHNOB-HCI dataset

To validate the subject-independent classification framework even further, the MAHNOB dataset is used similarly to DEAP dataset. The framework is validated by showing it can be used on a different EEG-based emotion recognition dataset. The MAHNOB-HCI dataset is split into training and testing sets, where the training set size was 80% of the total of 500 data points, and the remainder is left as a testing set. Furthermore, splitting of the dataset is done randomly and the model is trained multiple times allowing different training and testing sets. This would validate the model even further showing that the accuracy of the model would converge and not depend on the data points which it is trained on.

The results for two-class classification are shown in Table 4.9. Good accuracy can be reached using the method presented for all classification methods (kNN, SVM and GP classification). The accuracy of two-class valence classification is

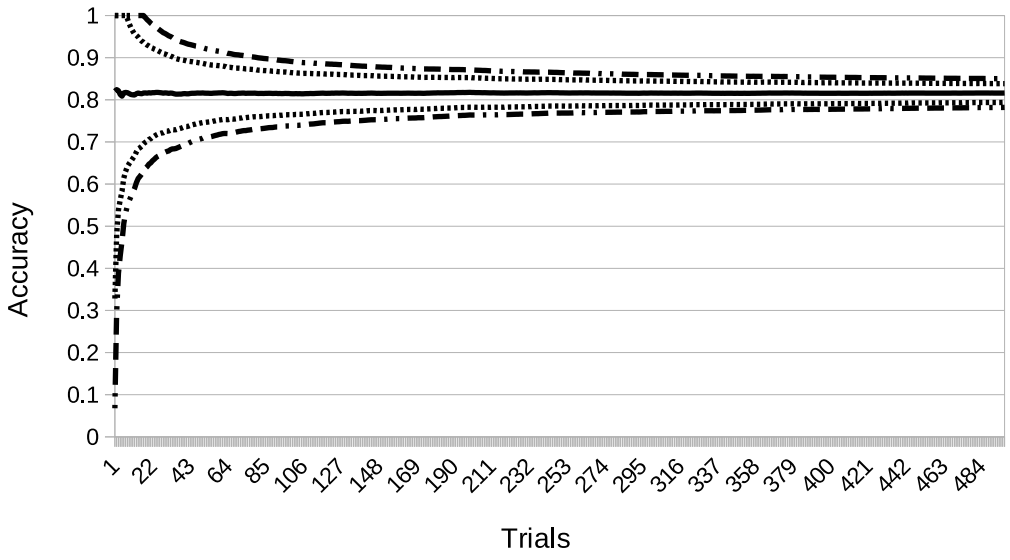


Figure 4.5: The two-class classification of DEAP arousal. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

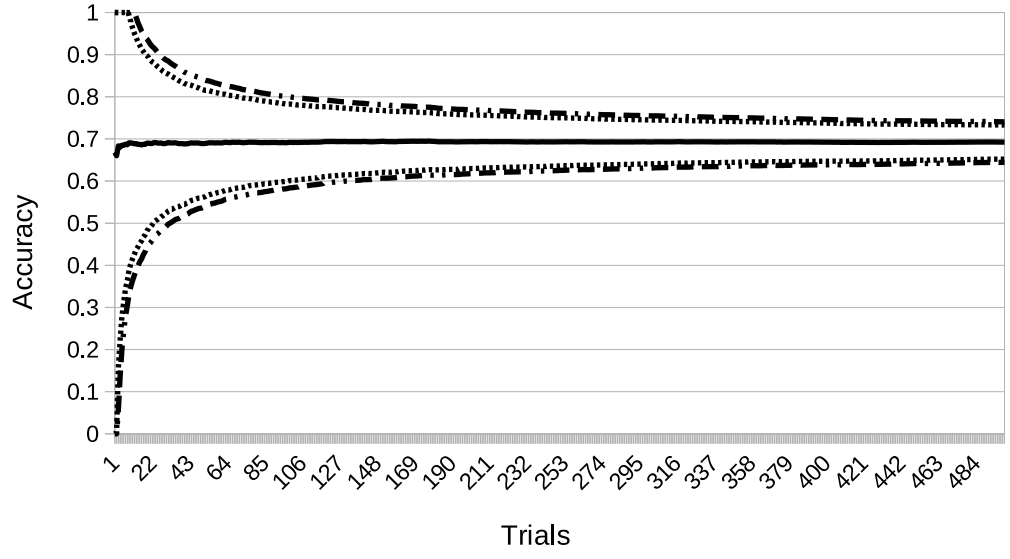


Figure 4.6: The three-class classification of DEAP valence. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

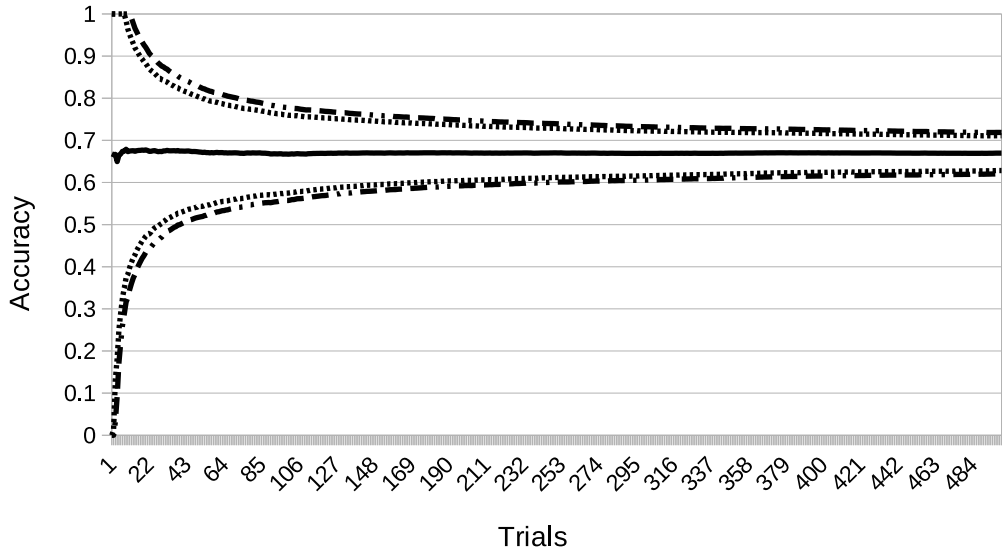


Figure 4.7: The three-class classification of DEAP arousal. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

Table 4.9: Subject-independent results using MAHNOB-HCI dataset two-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification	2	**83.35%**	**82.75%**
SVM	2	80.75%	80.3%
kNN	2	70.4%	72.39%

80.75% when using SVM, and 70.4% when using kNN classifier. However, the highest accuracy for subject-independent classification is reached using GP classification, 83.35%. The convergence of the two-class valence classification is shown in Figure 4.8 together with 95% and 85% confidence intervals. The 98% confidence interval for two-class valence is between 81.21% and 85.48%, and the 95% confidence interval is between 80.08% and 86.61%.

The two-class classification is also performed to classify positive and negative arousal. The GP classification resulted in 82.75% accuracy. The clear convergence of the model can be seen in Figure 4.9. The 85% and 95% confidence intervals for arousal are 80.58% and 84.92%, and 79.44% and 86.06%, respectively. To compare, the kNN and SVM classifiers were also used to classify arousal. The accuracy achieved using SVM and kNN are 81.0% and 72.39%, respectively.

Table 4.10: Subject-independent results using MAHNOB-HCI dataset three-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification	3	**69.12%**	**73.02%**
SVM	3	65.76 %	71.84%
kNN	3	54.64%	61.32%
ANOVA,SVM [110]	3	62.75%	64.74%

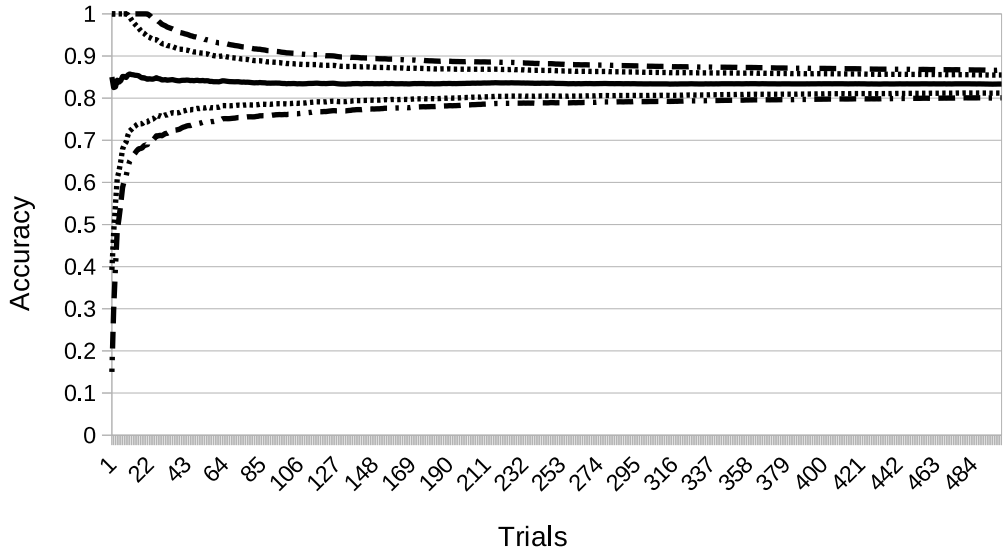


Figure 4.8: The two-class classification of MAHNOB-HCI valence. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

Over both modalities, arousal and valence, the average increase in accuracy is over 2% comparing to SVM and over 11% when comparing to kNN. Furthermore, the increase in accuracy is more significant in classifying valence.

In addition to two-class classification, three-class classification is performed using one-vs-rest model. The convergence plots with 95% and 85% confidence interval for valence and arousal are shown in Figures 4.10 and 4.11 respectively. The 98% confidence interval for valence is 64.31% and 73.94%, and for arousal is 68.39% and 77.64%. Furthermore, the 95% confidence interval is given, where the interval for valence is 65.07% to 73.17%, and for arousal is from 69.13% to 76.91%.

The accuracy of classifying arousal into positive, negative, and neutral is higher than for valence. The accuracy for three-class classification for valence is

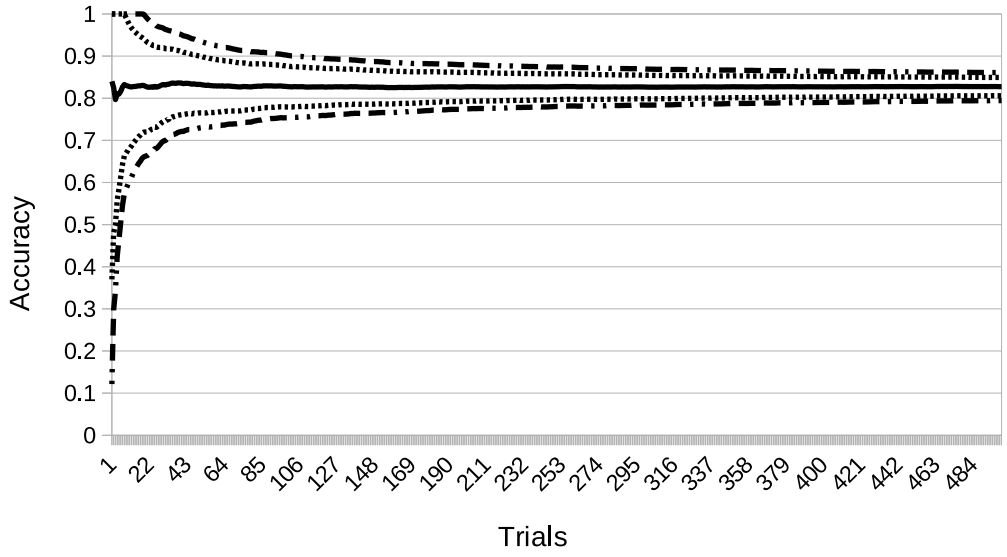


Figure 4.9: The two-class classification of MAHNOB-HCI arousal. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

54.64% when using kNN classifier and 65.76% when using SVM. The classification accuracy is noticeably higher when using GP classification, namely 69.12%.

Overall, the accuracy of arousal is higher for all classifiers. The three-class classification using kNN resulted in 61.32% accuracy, and using SVM resulted in 71.84% accuracy. The highest accuracy was again reached when using GP classification, similarly to all previous examples.

The average increase in accuracy is similarly to MAHNOB-HCI two-class classification example higher when classifying valence. Overall, both two-class classification and three-class classification give better results when classifying arousal. This is probably due to the dataset itself, as this is not evident in other examples, namely subject-dependent examples and examples using DEAP dataset.

Subject-Independent results using DREAMER dataset

To validate the framework even further, DREAMER dataset was used for two-class and three-class subject-independent classification. Similarly to DEAP and MAHNOB datasets, the DREAMER dataset gives good accuracy using GP classification. The two-class classification results are shown in Table 4.11 and the three-class classification results can be found in Table 4.12. Similarly to DEAP and MAHNOB, the SVM and kNN classification results are also included.

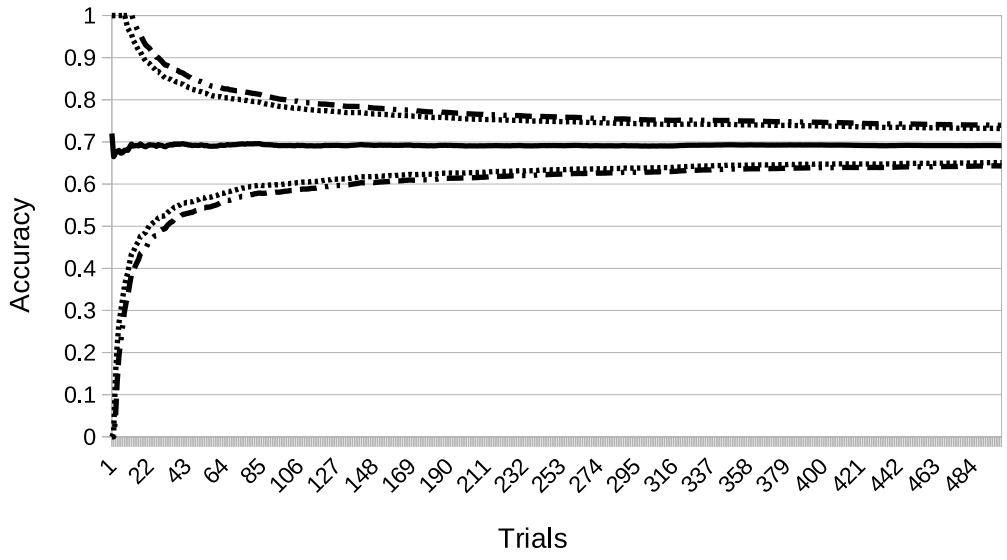


Figure 4.10: The three-class classification of MAHNOB-HCI valence. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

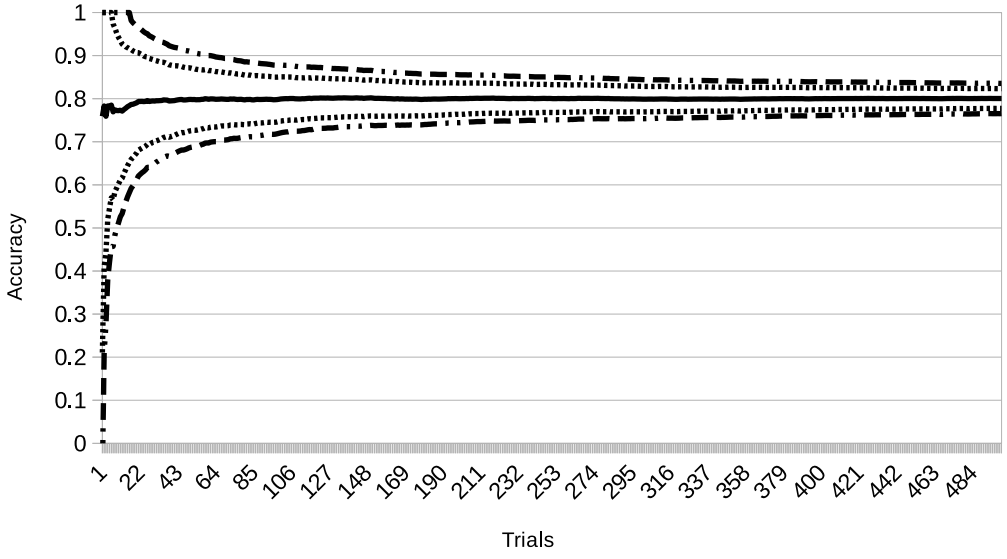


Figure 4.11: The three-class classification of MAHNOB-HCI arousal. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

For subject-independent two-class valence classification, the accuracy using kNN is 75.17% and the accuracy using SVM is 77.69%. The highest accuracy is reached using GP classification, where the two-class classification reaches 80.80%. Similarly, the GP classification gives better results for two-class arousal classification. The accuracy using GP classification is 80.04%, whereas the accuracy of kNN and SVM classifiers result an accuracy of 67.49% and 77.71% respectively. The Figures 4.12 and 4.13 respectively show the convergence plots for two-class valence and arousal classification, with 85% and 95% confidence intervals given on the same plot. The 85% confidence interval when classifying arousal lies between 82.33% and 77.75%, whereas the 95% lies between 83.55% and 76.54%. Likewise, for valence, the 95% confidence interval lies between 84.26% and 77.35, and 85% confidence interval lies between 83.06% and 78.55%.

Table 4.11: Subject-independent results using DREAMER dataset for two-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification	2	**80.80%**	**80.04%**
SVM	2	77.69%	77.71%
kNN	2	75.17%	67.49%

Table 4.12: Subject-independent results using DREAMER dataset for three-class classification.

Method	No. of classes per dimension	Valence	Arousal
GP classification	3	**73.12%**	**71.95%**
SVM	3	70.26%	68.95%
kNN	3	65.02%	65.93%

Similarly to two-class classification, the three-class classification gives good results for all three classifiers. The SVM and kNN classifiers resulted in valence accuracy of 70.26% and 65.02%, respectively. The accuracy of arousal is slightly lower for SVM and kNN, namely 68.95% and 65.93%. However, the GP classification gives the accuracy of 73.12% for valence and 71.95% for arousal, outperforming both SVM and kNN classifiers. The results are shown in Table 4.12.

The convergence plots are shown in Figure 4.14 for valence and Figure 4.15 for arousal. Likewise, the 85% and 95% confidence intervals are given. The 85% confidence interval of arousal lies between 74.53% and 69.38%, whereas the 95% confidence interval lies between 75.89% and 68.01%. Similarly, when classifying valence,

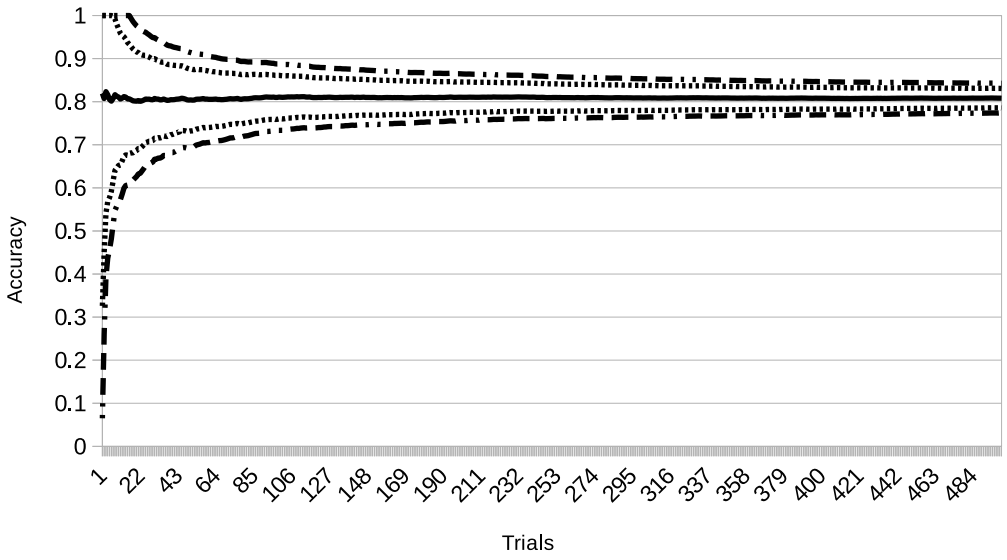


Figure 4.12: The three-class classification of DREAMER valence. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

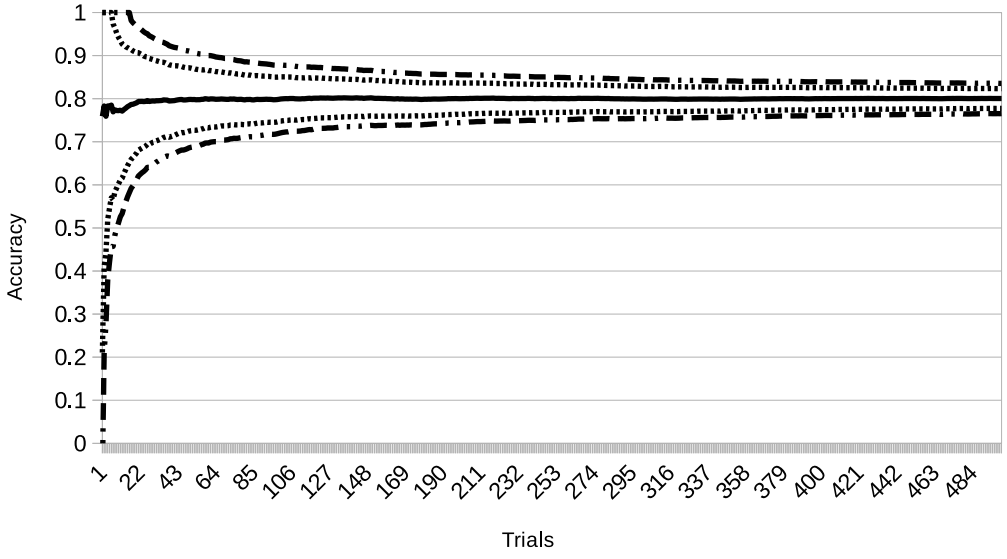


Figure 4.13: The two-class classification of DREAMER arousal. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

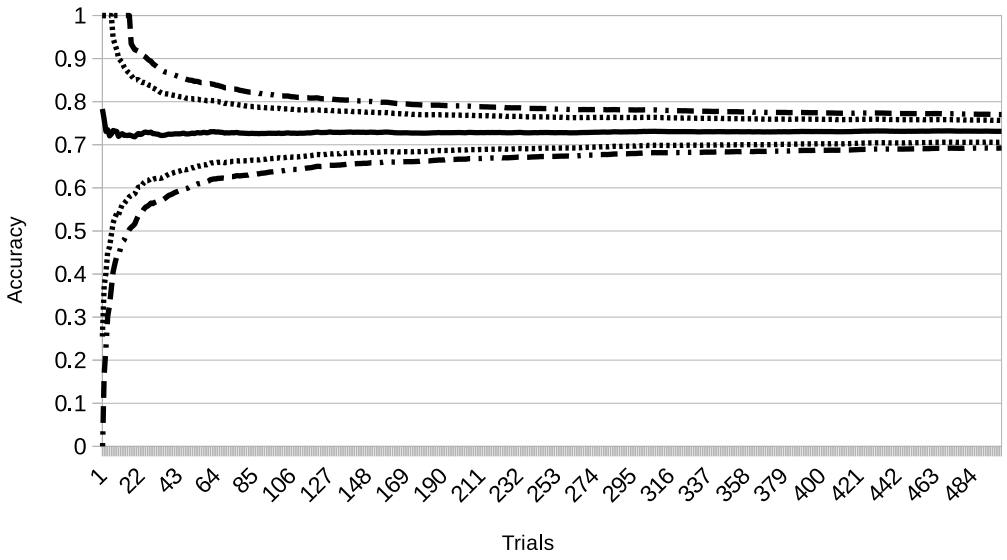


Figure 4.14: The three-class classification of DREAMER valence. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

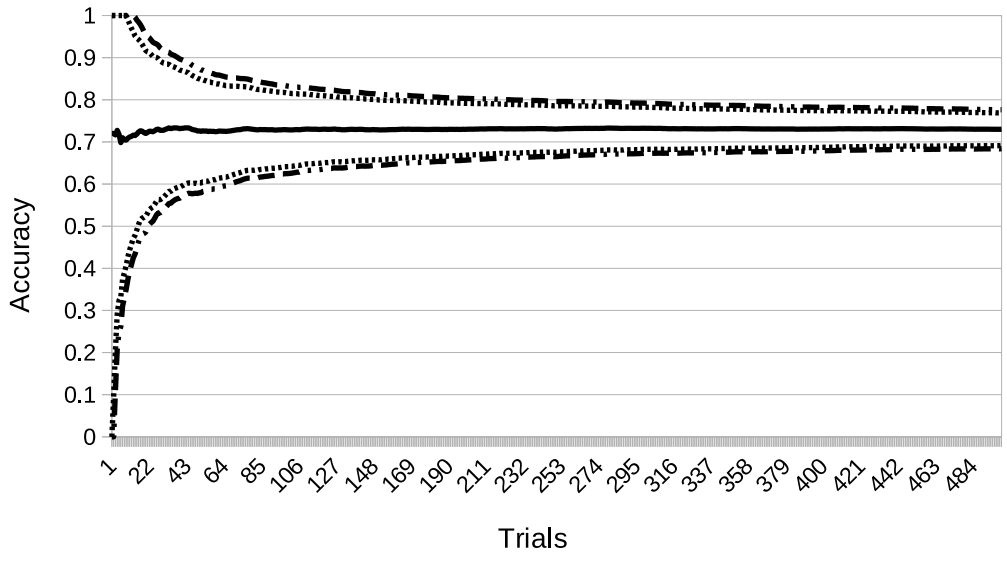


Figure 4.15: The three-class classification of DREAMER arousal. Convergence plot showing the average classification accuracy on the training set denoted by a solid line, together with 85% confidence interval denoted by a dotted line and 95% confidence interval denoted by a dash/dot line.

the 85% confidence interval lies between 75.66% and 70.58, and 95% confidence interval lies between 77.01% and 69.23%.

4.4 Analysis

Due to the nature of emotions, the EEG-based emotion recognition is often performed following subject-dependent principles. The state-of-the-art methods have managed to get high accuracy for two-class classification. However, it has been shown that the framework proposed in this chapter improves the classification accuracy. From the subject-dependent results it can be concluded that GP classification works well with EEG-based emotion recognition, increasing accuracy when compared to state-of-the-art methods. The classification accuracy is over 90% for all three datasets used.

In addition, subject-dependent two-class classification was performed using both mRMR and reliefF feature selection. The two-class classification results are better using mRMR feature selection. However, the multiclass classification achieved higher accuracy using reliefF feature selection. This can be explained by the fact that mRMR is not able to measure interactions between the features. This can create situations where the mRMR algorithm underestimates the usefulness of some features, mainly in cases where the features are individually useless, but are useful when combined.

Comparing the subject-dependent and the subject-independent results, as expected, the latter is less accurate. However, being able to recognise emotions from EEG signals subject-independently opens up a variety of applications. Nevertheless, the subject-independent model still gives better results than the subject-dependent model presented in [6]. Hence, the two-class and three-class classification performs well with both subject-dependent and subject-independent models. Furthermore, when comparing to state-of-the-art methods, the results are comparable or better for both subject-independent and subject-dependent models.

For the subject-independent results, it is difficult to make comparisons with other published methods, as there are not as many studies as there are the ones using subject-dependent model. This is why in addition to GP classification, comparative results using kNN and SVM classifiers are presented.

In addition, in [58] the highest average accuracy for two-class classification using leave-one-subject-out and DEAP dataset is 59.06%, which is significantly worse than the result of our proposed framework of 80.5% and 81.07%. It has to be noted that the comparison may not be completely accurate as in this chapter the test set is a random sample of 20% of the data, whereas in [58] a leave-one-subject-out is used.

In [10], the training accuracy reached for two-class valence classification is 71.20%. To make an accurate comparison between the two models is complicated, as the datasets used in [10] and this paper are different. This is because the nature of emotion recognition using EEG signals is dependent on multiple factors including the stimulus used (the datasets in this paper used video segments where as in [10] pictures are used). Furthermore, the training accuracy was significantly higher than their testing accuracy, thus suggesting that the model is overfitting, which is naturally avoided using our Gaussian process. The training and testing accuracy was obtained when the model is applied to the training and testing sets, respectively. If the difference between training and testing accuracy is large, i.e., the model learned rules specifically for the training set, the rules do not generalise well.

Overall, it can be concluded that GP classification is a good option for the recognition. The experimental results indicate that its use increases the accuracy of the emotion recognition when compared to state-of-the-art methods. The GP classification has shown to give more accurate results for EEG-based emotion recognition.

In addition, the increase in accuracy is higher for three-class classification (over 5%) than for two-class classification (1.5%). This gives evidence to suggest that the GP classification performs better for a higher number of classes than other classifiers used (i.e., SVM and kNN).

The main problem with GP classification is that the training time is higher than when using some simpler classifiers, e.g., kNN, but similar to using SVM. The computational complexity of GP classification is $O(n^3)$. On the other hand, where most other commonly used classifiers are prone to overfitting, the Bayesian approach in GP classification minimises overfitting of the model.

4.5 Summary

In this chapter, GP classification has been introduced for EEG-based emotion recognition. The results show that the average accuracy of subject-dependent classification increases when using GP classification compared to state-of-the-art methods using SVM [6], and CNN [104] classifiers. The results are presented using both mRMR and reliefF feature selection.

Furthermore, a framework for subject-independent classification is presented with a novel feature extraction method using increased number of features. This is used for classification using GP classifier, kNN, and SVM. From the results achieved in this chapter, the subject-independent classification model gives the best results when using GP classifier. This agrees with the subject-dependent classification

model.

The experiments on both subject-dependent and subject-independent classification were conducted on three datasets (DEAP, MAHNOB-HCI, and DREAMER) to validate the proposed framework. The results achieved are alike for two-class and three-class classification.

The work in this chapter is based on the work published in [80], with some extensions, including validating the results using a third dataset, DREAMER, and presenting results using reliefF feature selection. Overall, it can be concluded that GP classification is a good option for the emotion recognition. The experimental results indicate that its use increases the accuracy of the emotion recognition when compared to state-of-the-art methods. The GP classification has shown to give more accurate results for EEG-based emotion recognition problems.

Chapter 5

Conclusions and discussion

5.1 Conclusion

The goal of this thesis has been to introduce new ideas to the area of EEG-based emotion recognition. The thesis explored, analysed, and compared different signal processing methods suitable for EEG-based emotion recognition. The author has proposed a mutual information based signal reduction algorithm. In addition, a subject-dependent emotion recognition framework has been proposed making use of the mutual information based signal reduction, and it has been shown to improve the accuracy when compared to the state-of-the-art methods. Furthermore, Gaussian process classification has been introduced for the purpose of EEG-based emotion recognition. A framework for subject-independent emotion recognition has been presented, showing that using the methods presented in this thesis, result in improved accuracy. The results presented in this thesis are obtained using three different publicly available datasets, showing that the frameworks proposed generalise well.

5.1.1 Mutual information based adaptive windowing

EEG-based emotion recognition datasets are often small, multidimensional, and noisy. Furthermore, in the creation of the datasets, often lengthy videos are used as stimuli, resulting in long signals. It has been suggested that during these video clips, the intensity of the emotions can change. It has been proposed that the accuracy of the EEG-based emotion recognition can be increased by identifying the section of the signal where the emotion intensity is highest. This has been achieved by proposing a mutual information based algorithm for signal reduction.

The algorithm introduced in Chapter 3 works by iteratively computing the mutual information between different-length EEG signals at different time locations and emotion labels. The signal with the highest mutual information is used for

extracting the features for emotion classification. It has been shown that reducing the signal increased the accuracy of the EEG-based emotion recognition significantly. Several feature extraction methods as well as classification methods have been compared using both reduced signal and entire signal. The increase in accuracy has been evident for all feature extraction and classification methods. The increase in accuracy has also been evident in multiclass emotion recognition.

The main features compared were statistical, HOS, HOC, and PSE features. The statistical features have been shown to give the most accurate results. It has also been shown that PSE features did not perform as well as other features for subject-dependent emotion recognition. These results have been consistent over all three datasets (DEAP, MAHNOB, and DREAMER). Due to the small size of the datasets, to validate the results, leave-one-out cross validation was used. In addition, the experimental results were analysed using ROC curves. The curves verified the conclusions drawn. The three classifiers compared give good results for all datasets using reduced signals. However, when using entire signal, often kNN classifier under performs against NB and SVM classifiers.

5.1.2 EEG-based emotion recognition using Gaussian process classification

In Chapter 4, the GP classification has been introduced for EEG-based emotion recognition. It has been demonstrated, that the average accuracy of subject-dependent classification increases when using GP classification compared to state-of-the-art methods using SVM [6, 80], kNN [80], and CNN [104] classifiers.

A framework for subject-independent classification is presented with a novel feature extraction method using increased number of features. This was used for classification using GP classifier, kNN, and SVM. From the results achieved in this thesis, the subject-independent classification framework gives the best results when using GP classifier. This agrees with the results of the subject-dependent classification.

The experiments on both subject-dependent and subject-independent classification were conducted on three datasets (DEAP, MAHNOB-HCI, and DREAMER) to validate the proposed frameworks. The results achieved are alike for two- and three-class classification. Overall, the GP classification has proven to be a good option for the EEG-based emotion recognition. The experimental results indicate that its use increases the accuracy of the emotion recognition when compared with the state-of-the-art methods. The GP classification has been shown to give more accurate results for EEG-based emotion recognition.

As expected, the subject-independent emotion recognition does not give

as good results as the subject-dependent emotion recognition. Nevertheless, the subject-independent framework still gives better results than the subject-dependent model presented in [6]. Hence, the two- and three-class classification performs well with both subject-dependent and subject-independent models. Furthermore, when compared with the state-of-the-art methods, the results are comparable or better for both subject-independent and subject-dependent emotion recognition.

5.2 Limitations and future work

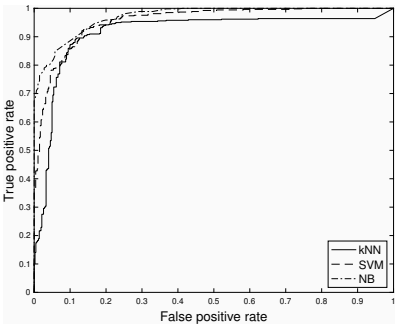
The main limitations of the EEG-based emotion recognition are related to the training time, and the number of classification classes. For a single participant, full training using the proposed signal reduction algorithm requires about 12 hours. This was somewhat overcome by parallelising the code, so that the training was performed for all subjects simultaneously. In addition, due to the size of the datasets, the current number of classes in the classifier is limited.

The direction of the future work is closely linked to these limitations. One of the main objectives of the future work is to improve the training time of the models. Currently, the EEG-based emotion recognition training time is relatively long, and infeasible for many applications. However, for a lot of the applications (e.g., virtual reality and gaming), the training can be done off-line, and more important is the fast classification time. The eventual goal would be the ability to train and use the proposed models in real time. This could be improved by better optimisation of the code, and increased computing power.

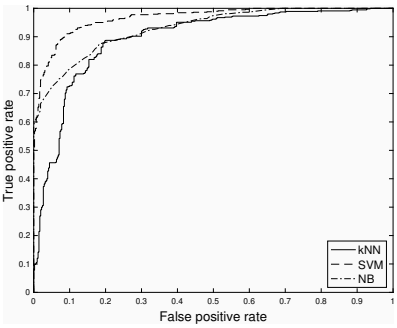
Another prime direction is to work towards increasing the number of classes. The main aim here is to work towards creating larger datasets, as the number of samples in current EEG datasets is relatively small. Having more samples can increase the accuracy, especially when dealing with multi-class emotion recognition.

Appendix A

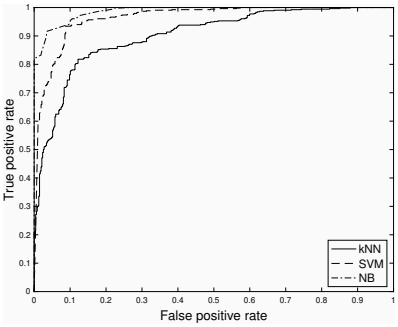
Subject-Dependent Emotion Recognition ROC Curves using DEAP dataset



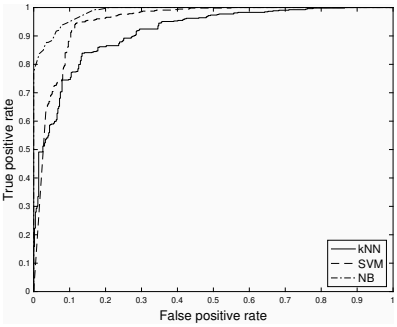
(a) ROC curve using reduced signal and HOC features.



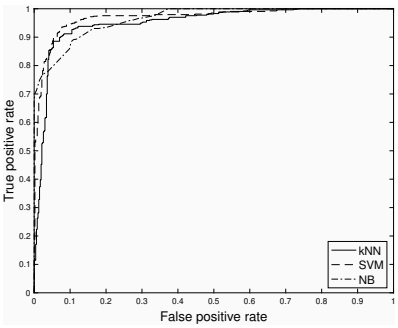
(b) ROC curve using all signal and HOC features.



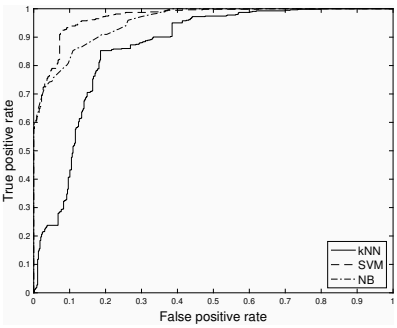
(c) ROC curve using reduced signal and HOS features.



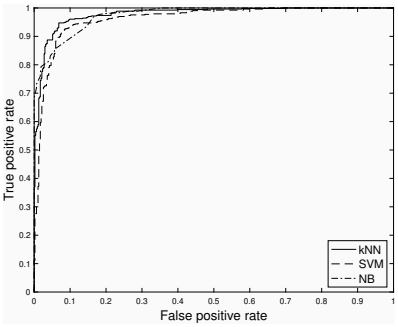
(d) ROC curve using all signal and HOS features.



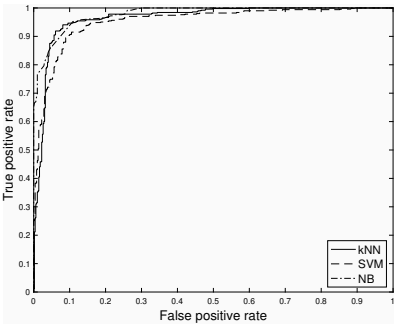
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

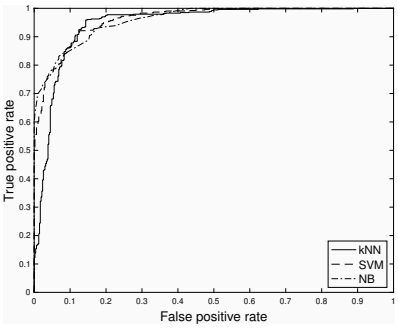


(g) ROC curve using reduced signal and SF.

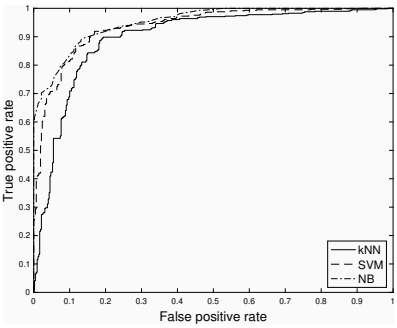


(h) ROC curve using all signal and SF.

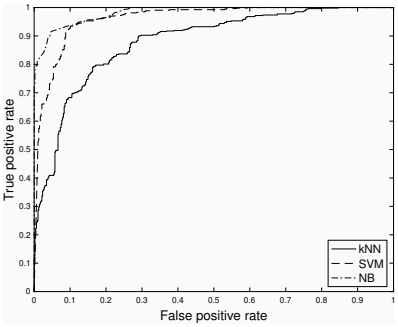
Figure A.1: Results using DEAP dataset and 30 features for classification of Arousal.



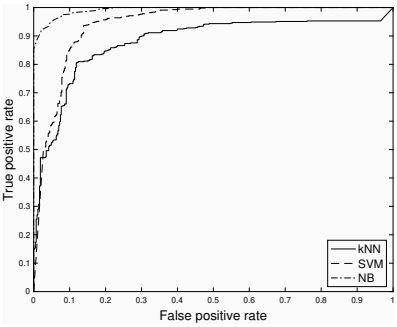
(a) ROC curve using reduced signal and HOC features.



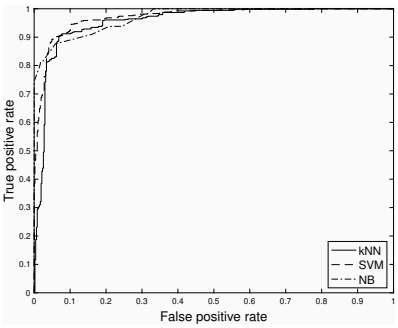
(b) ROC curve using all signal and HOC features.



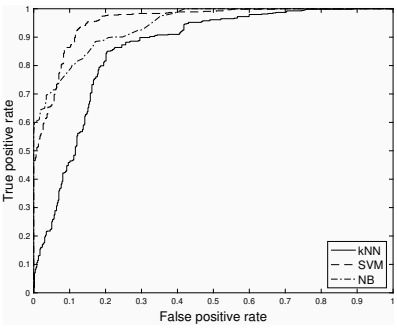
(c) ROC curve using reduced signal and HOS features.



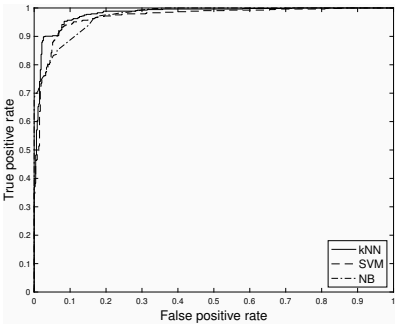
(d) ROC curve using all signal and HOS features.



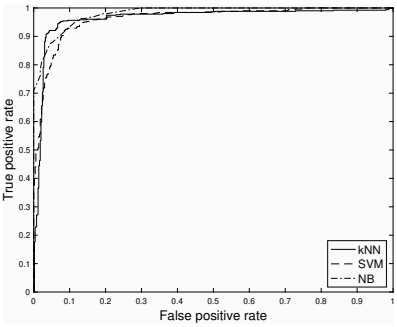
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

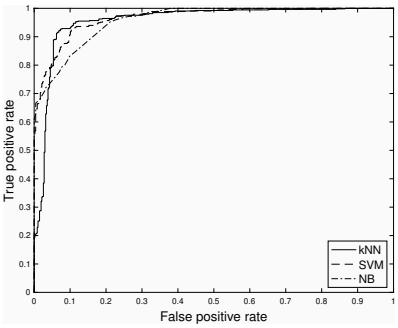


(g) ROC curve using reduced signal and SF.

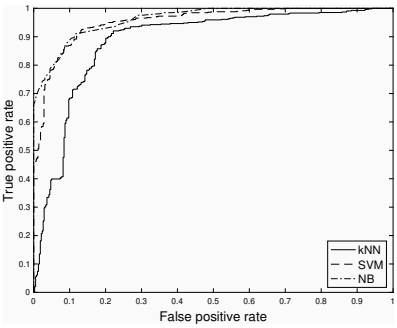


(h) ROC curve using all signal and SF.

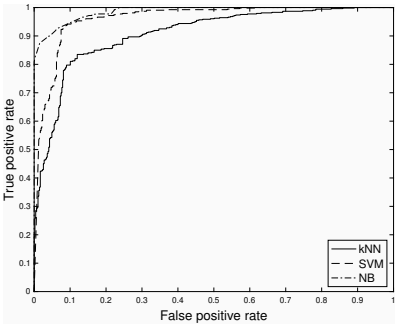
Figure A.2: Results using DEAP dataset and 31 features for classification of Arousal.



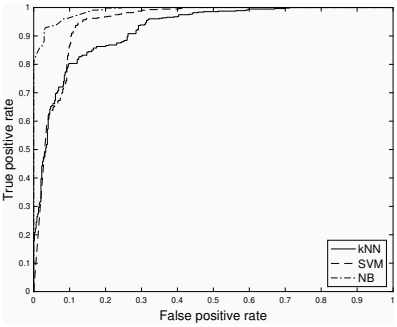
(a) ROC curve using reduced signal and HOC features.



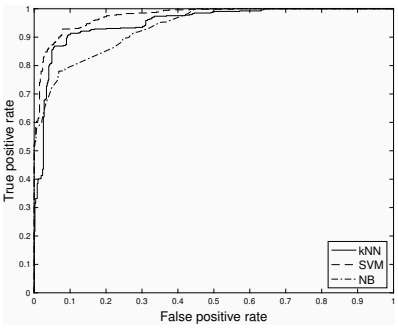
(b) ROC curve using all signal and HOC features.



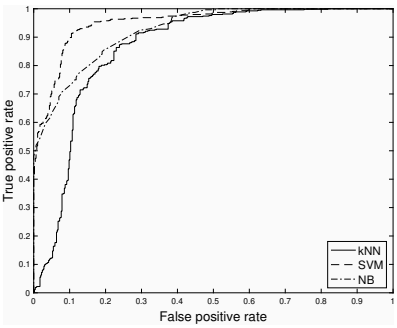
(c) ROC curve using reduced signal and HOS features.



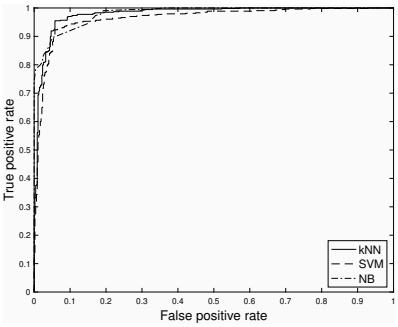
(d) ROC curve using all signal and HOS features.



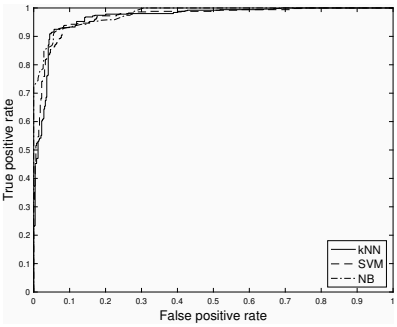
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

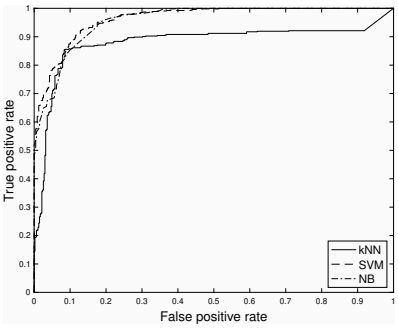


(g) ROC curve using reduced signal and SF.

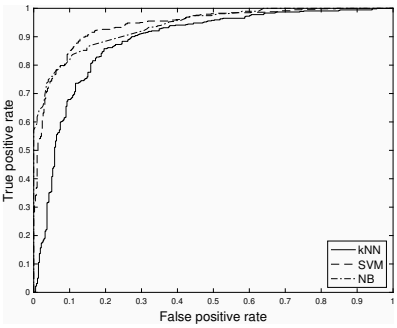


(h) ROC curve using all signal and SF.

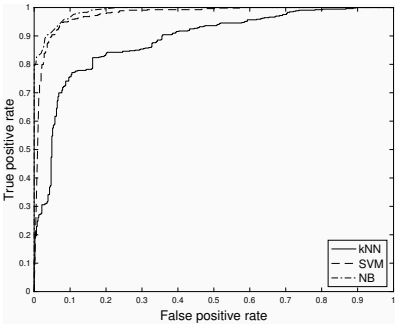
Figure A.3: Results using DEAP dataset and 32 features for classification of Arousal.



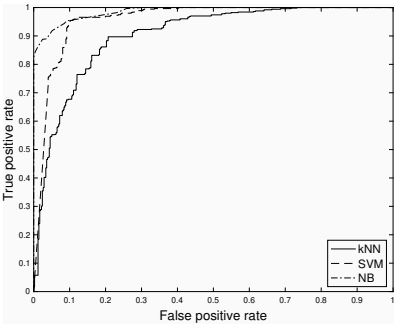
(a) ROC curve using reduced signal and HOC features.



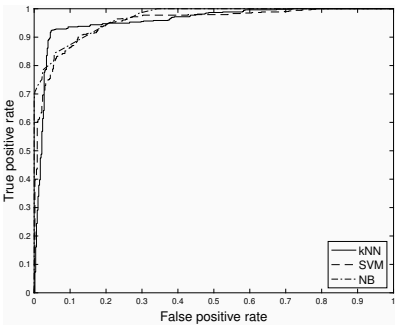
(b) ROC curve using all signal and HOC features.



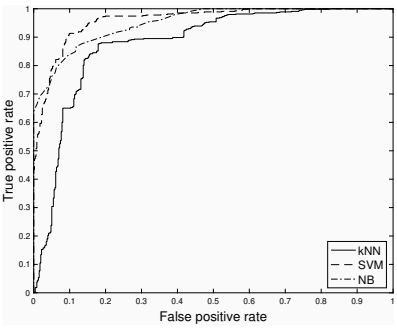
(c) ROC curve using reduced signal and HOS features.



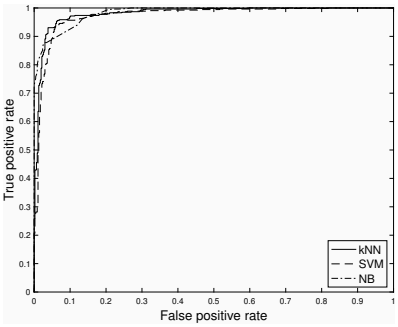
(d) ROC curve using all signal and HOS features.



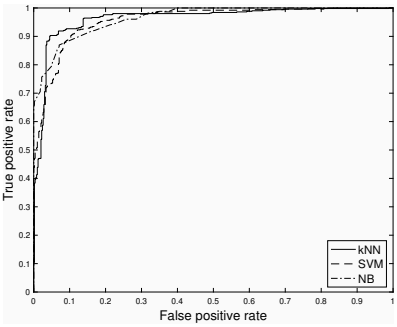
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

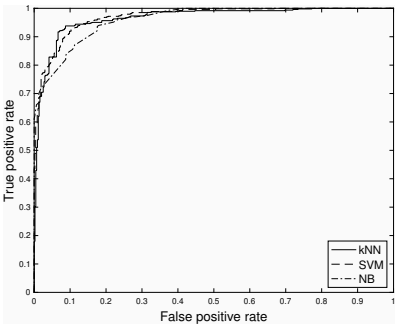


(g) ROC curve using reduced signal and SF.

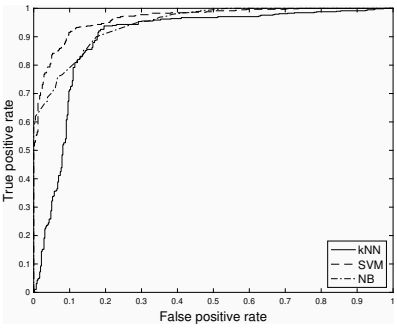


(h) ROC curve using all signal and SF.

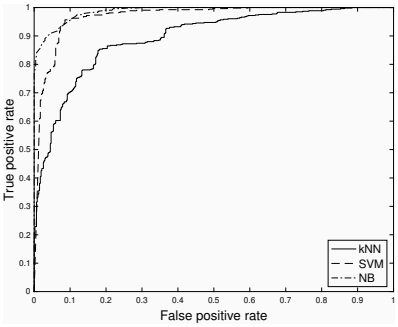
Figure A.4: Results using DEAP dataset and 33 features for classification of Arousal.



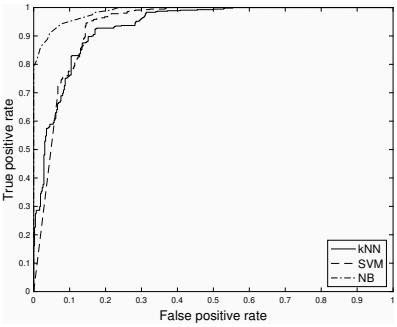
(a) ROC curve using reduced signal and HOC features.



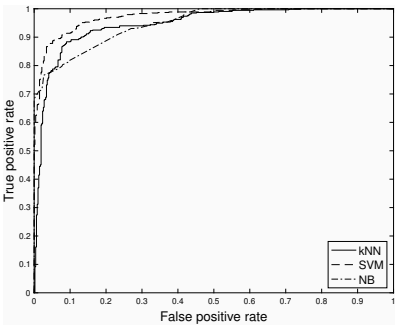
(b) ROC curve using all signal and HOC features.



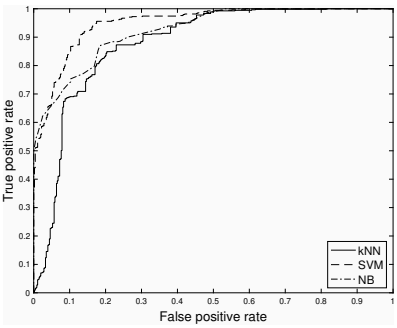
(c) ROC curve using reduced signal and HOS features.



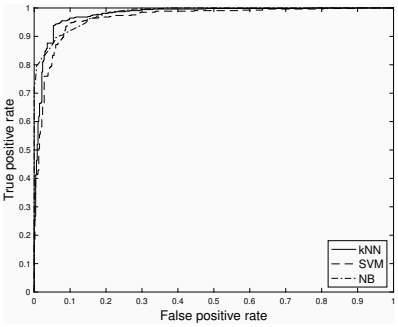
(d) ROC curve using all signal and HOS features.



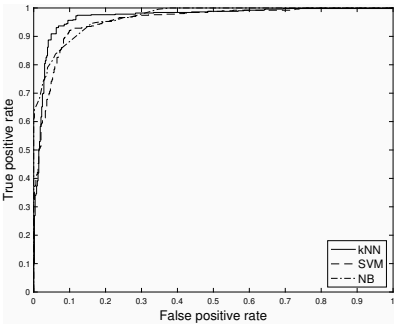
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

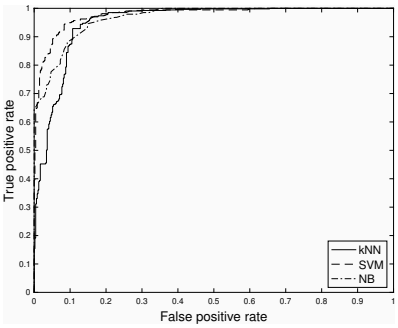


(g) ROC curve using reduced signal and SF.

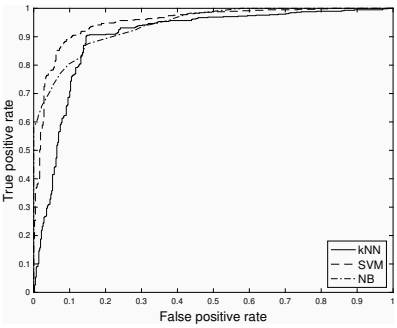


(h) ROC curve using all signal and SF.

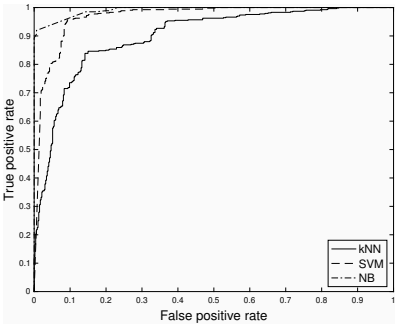
Figure A.5: Results using DEAP dataset and 34 features for classification of Arousal.



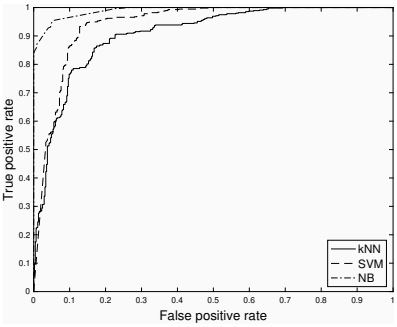
(a) ROC curve using reduced signal and HOC features.



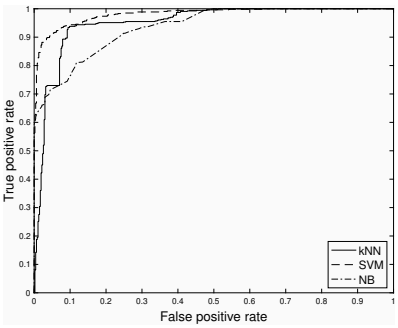
(b) ROC curve using all signal and HOC features.



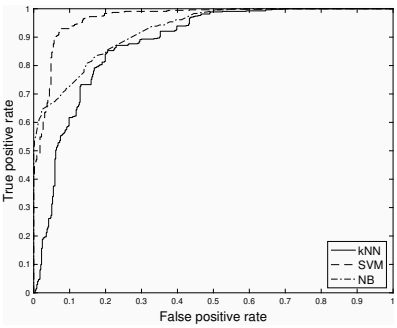
(c) ROC curve using reduced signal and HOS features.



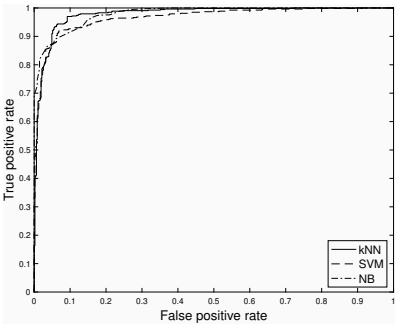
(d) ROC curve using all signal and HOS features.



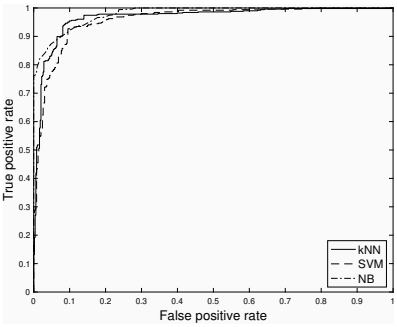
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

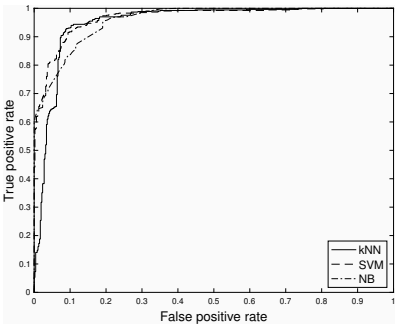


(g) ROC curve using reduced signal and SF.

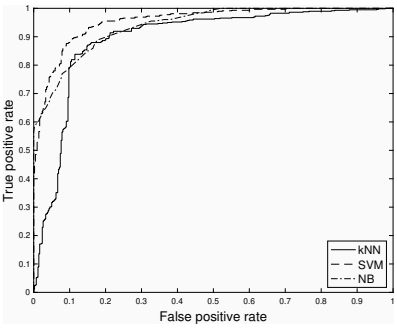


(h) ROC curve using all signal and SF.

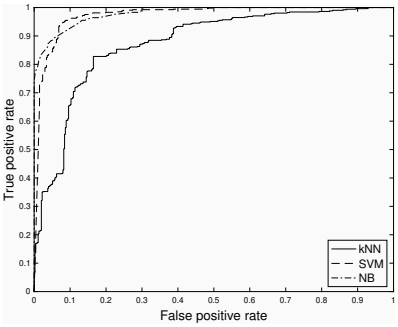
Figure A.6: Results using DEAP dataset and 35 features for classification of Arousal.



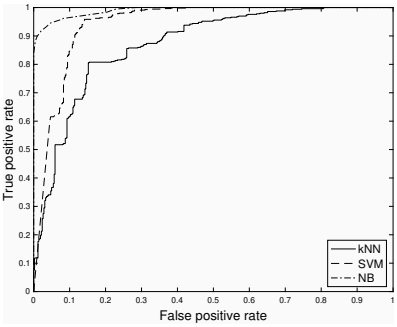
(a) ROC curve using reduced signal and HOC features.



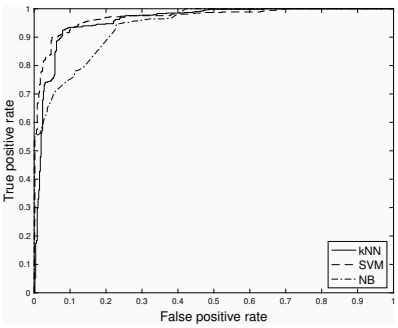
(b) ROC curve using all signal and HOC features.



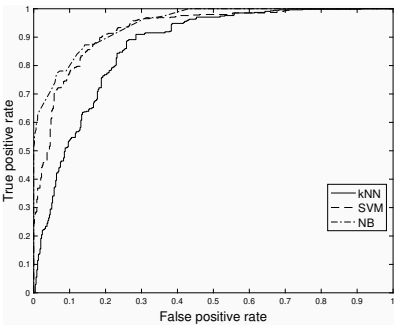
(c) ROC curve using reduced signal and HOS features.



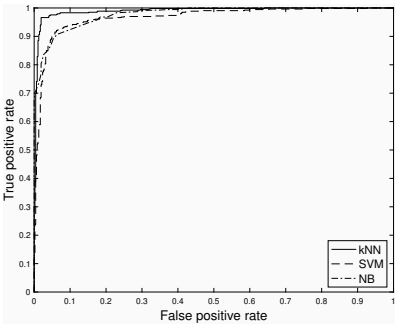
(d) ROC curve using all signal and HOS features.



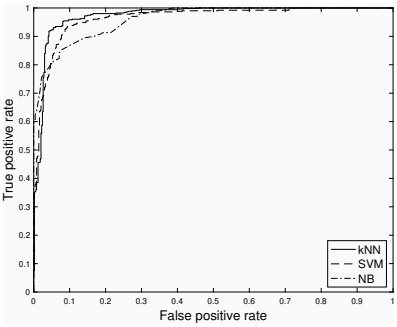
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

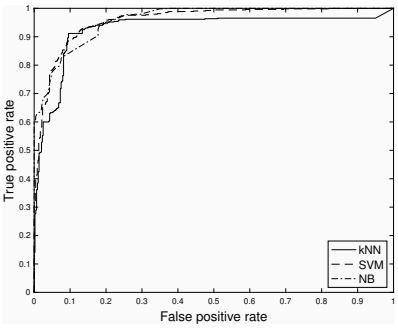


(g) ROC curve using reduced signal and SF.

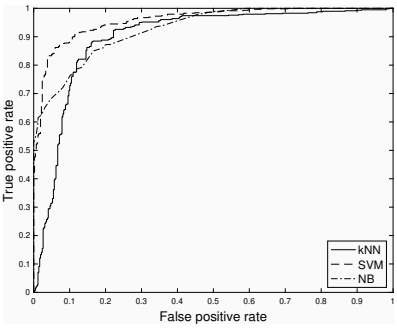


(h) ROC curve using all signal and SF.

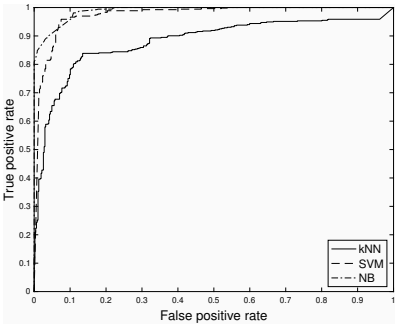
Figure A.7: Results using DEAP dataset and 36 features for classification of Arousal.



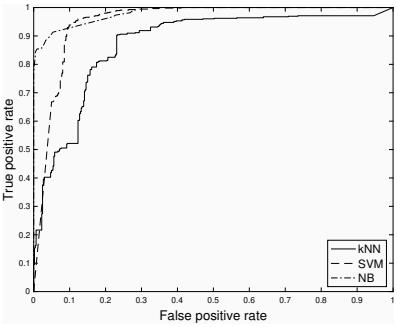
(a) ROC curve using reduced signal and HOC features.



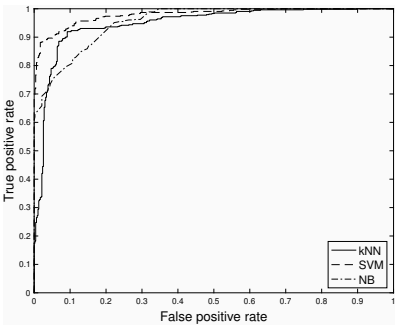
(b) ROC curve using all signal and HOC features.



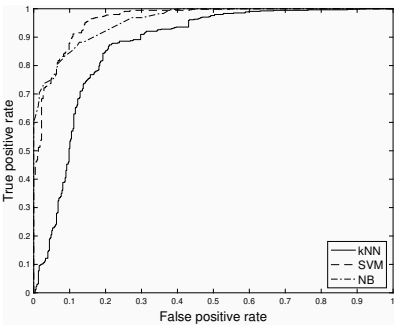
(c) ROC curve using reduced signal and HOS features.



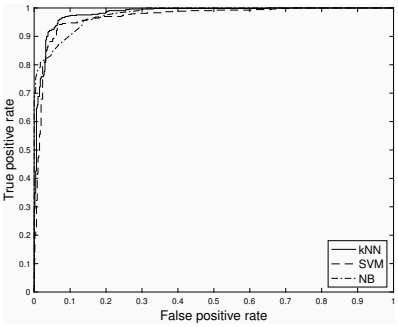
(d) ROC curve using all signal and HOS features.



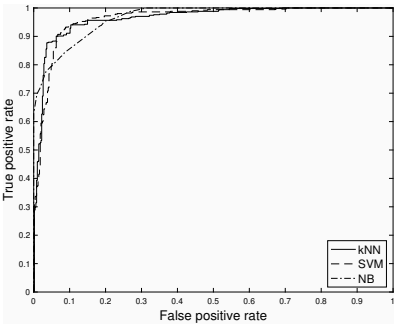
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

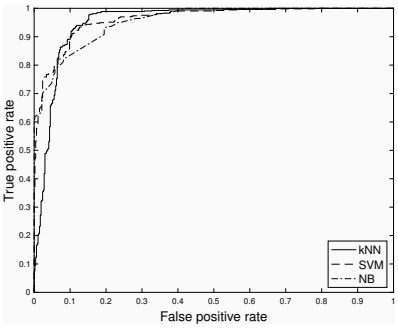


(g) ROC curve using reduced signal and SF.

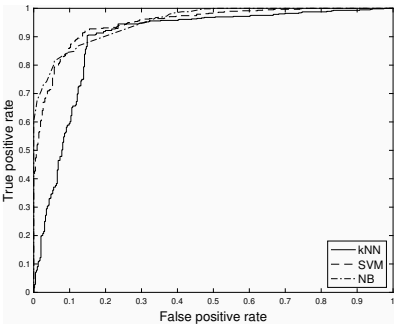


(h) ROC curve using all signal and SF.

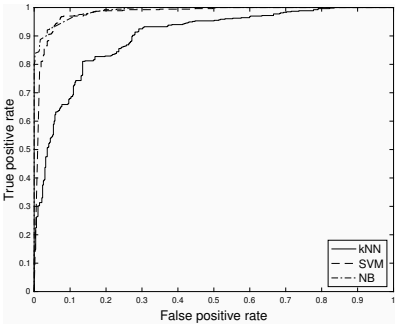
Figure A.8: Results using DEAP dataset and 37 features for classification of Arousal.



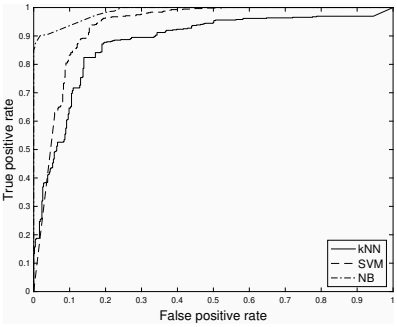
(a) ROC curve using reduced signal and HOC features.



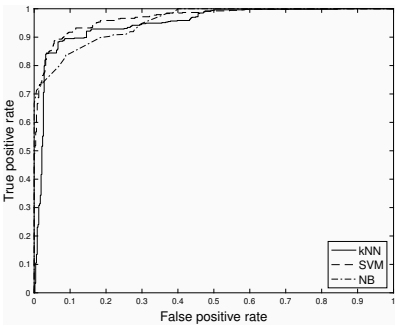
(b) ROC curve using all signal and HOC features.



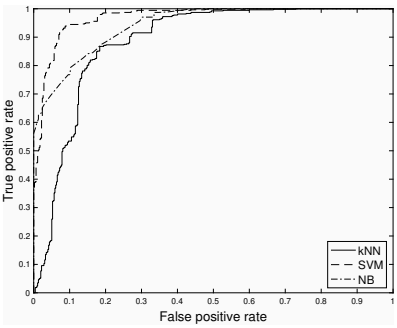
(c) ROC curve using reduced signal and HOS features.



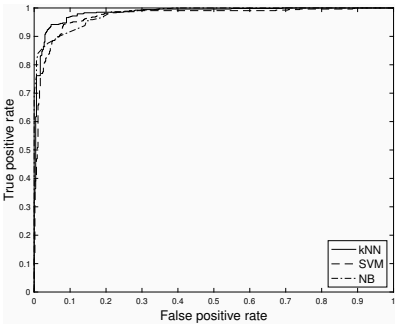
(d) ROC curve using all signal and HOS features.



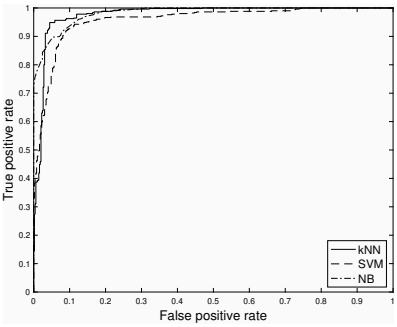
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

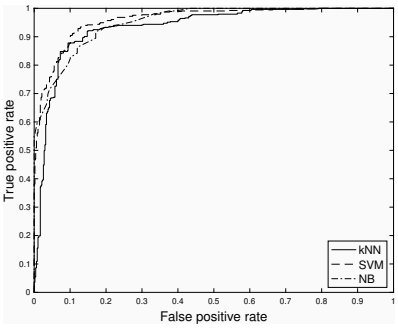


(g) ROC curve using reduced signal and SF.

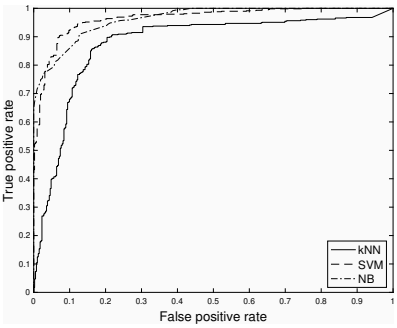


(h) ROC curve using all signal and SF.

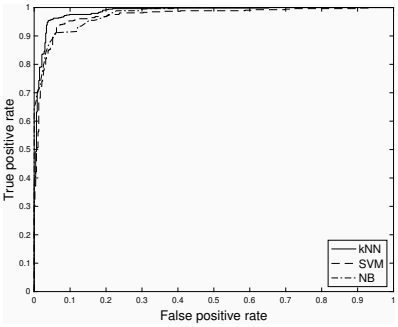
Figure A.9: Results using DEAP dataset and 38 features for classification of Arousal.



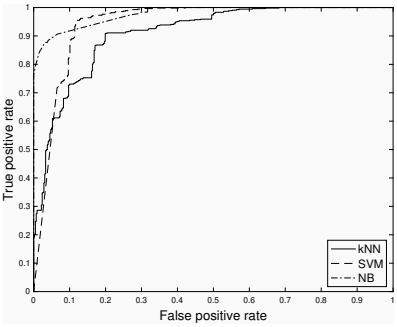
(a) ROC curve using reduced signal and HOC features.



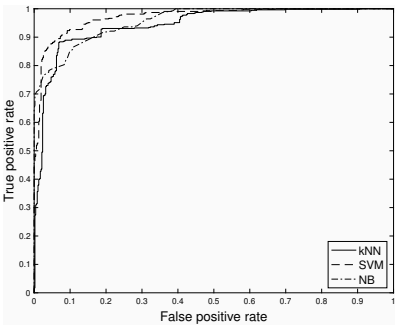
(b) ROC curve using all signal and HOC features.



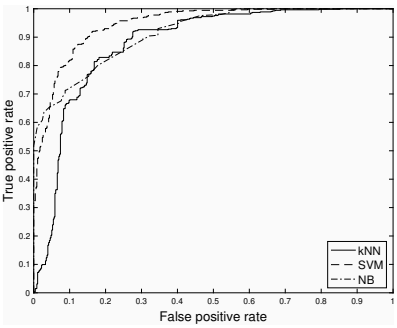
(c) ROC curve using reduced signal and HOS features.



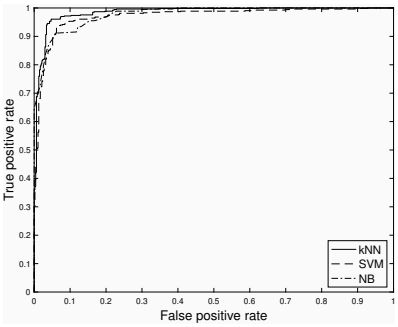
(d) ROC curve using all signal and HOS features.



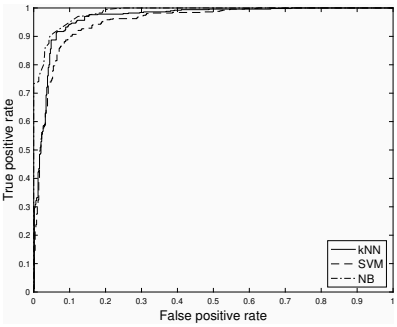
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

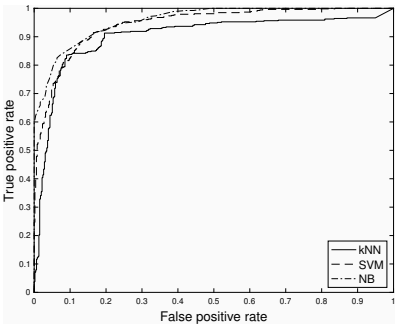


(g) ROC curve using reduced signal and SF.

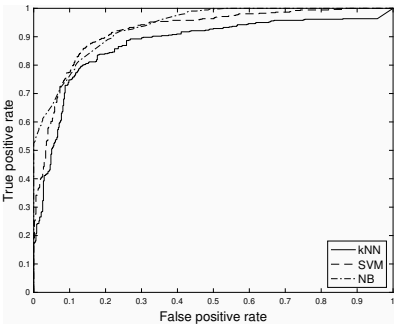


(h) ROC curve using all signal and SF.

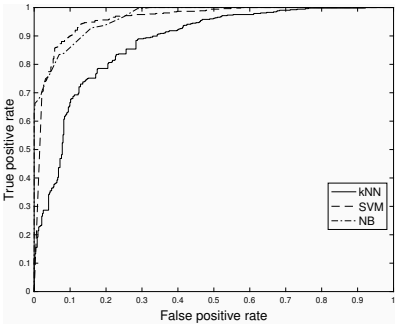
Figure A.10: Results using DEAP dataset and 39 features for classification of Arousal.



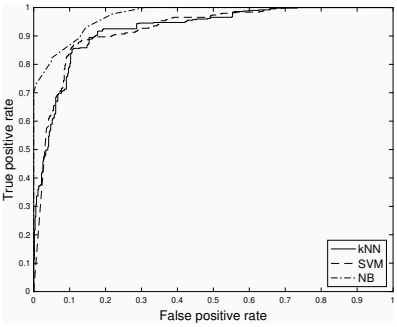
(a) ROC curve using reduced signal and HOC features.



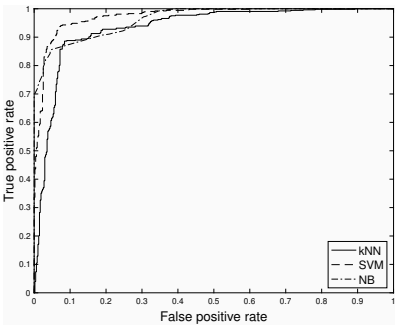
(b) ROC curve using all signal and HOC features.



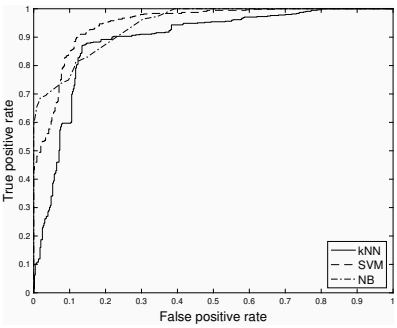
(c) ROC curve using reduced signal and HOS features.



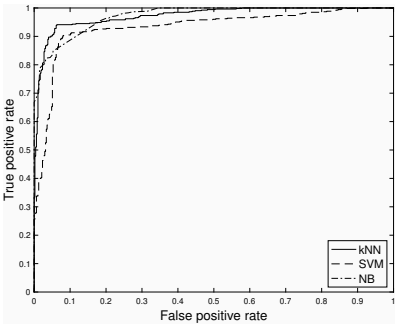
(d) ROC curve using all signal and HOS features.



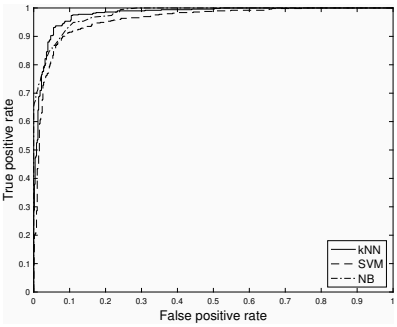
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

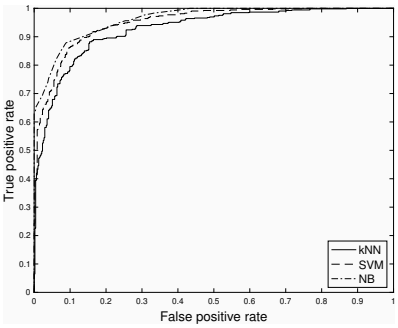


(g) ROC curve using reduced signal and SF.

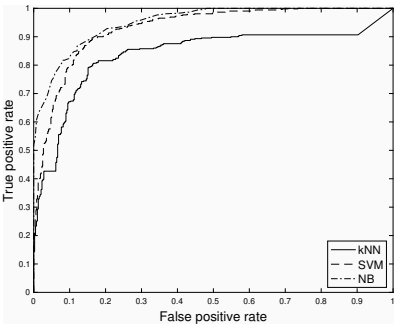


(h) ROC curve using all signal and SF.

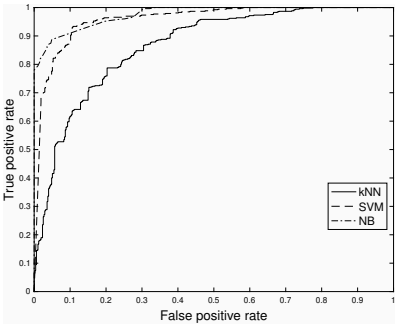
Figure A.11: Results using DEAP dataset and 30 features for classification of valence.



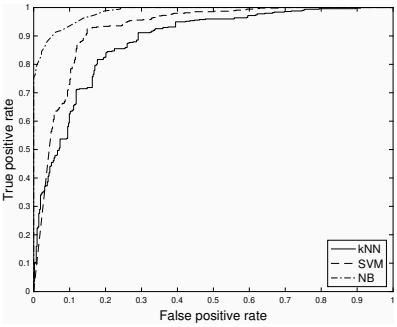
(a) ROC curve using reduced signal and HOC features.



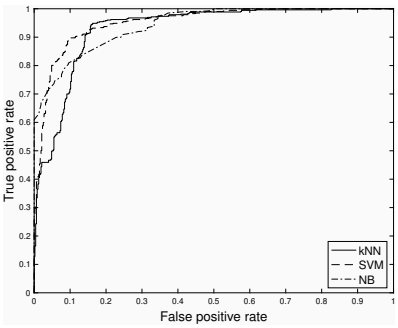
(b) ROC curve using all signal and HOC features.



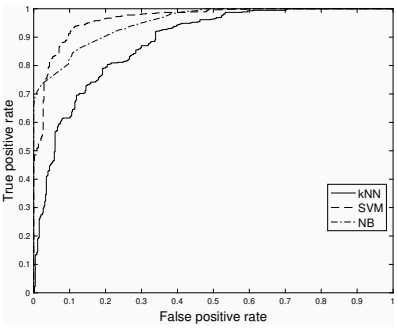
(c) ROC curve using reduced signal and HOS features.



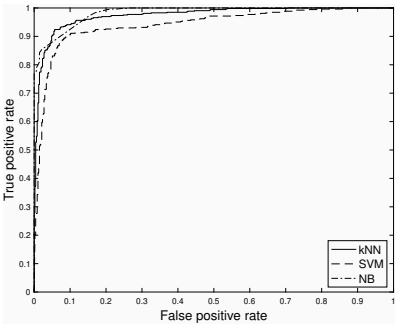
(d) ROC curve using all signal and HOS features.



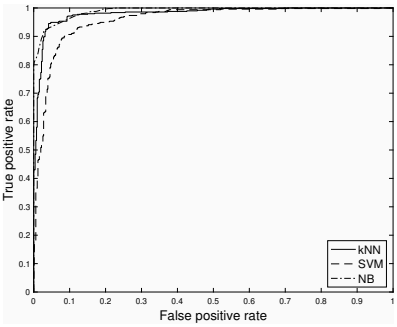
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

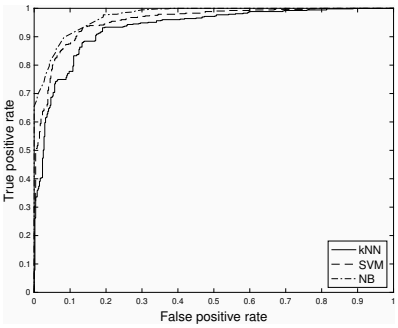


(g) ROC curve using reduced signal and SF.

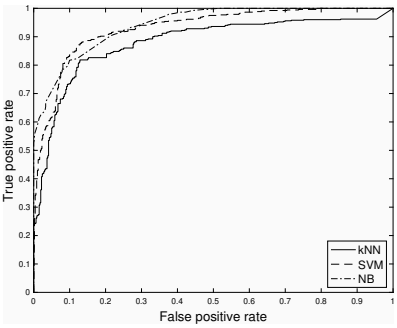


(h) ROC curve using all signal and SF.

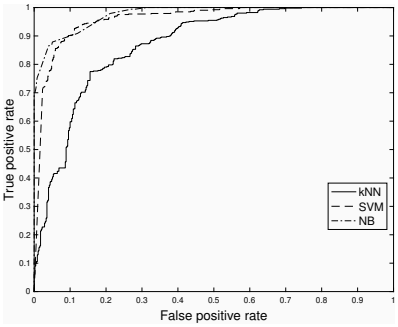
Figure A.12: Results using DEAP dataset and 31 features for classification of Arousal.



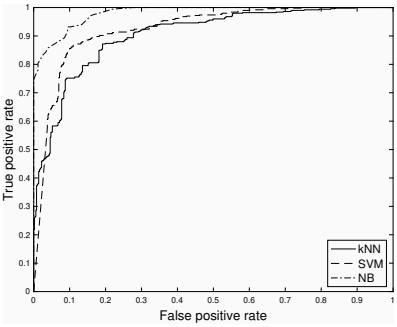
(a) ROC curve using reduced signal and HOC features.



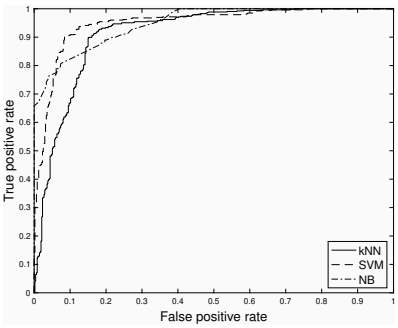
(b) ROC curve using all signal and HOC features.



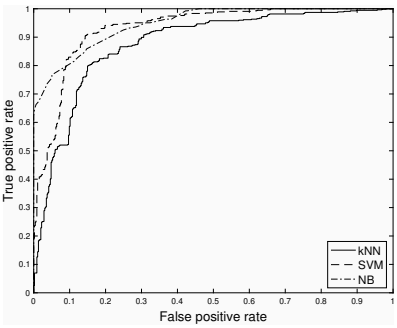
(c) ROC curve using reduced signal and HOS features.



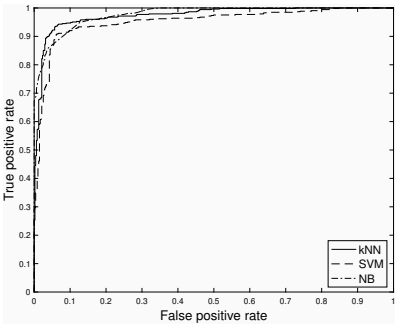
(d) ROC curve using all signal and HOS features.



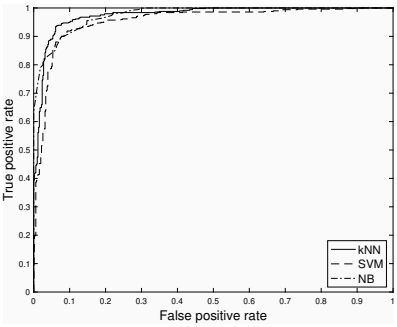
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

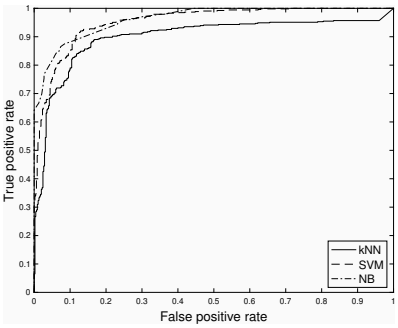


(g) ROC curve using reduced signal and SF.

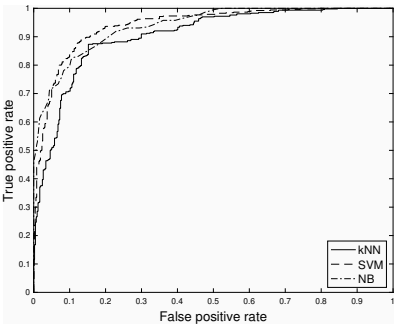


(h) ROC curve using all signal and SF.

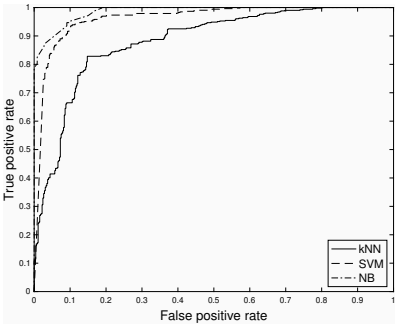
Figure A.13: Results using DEAP dataset and 32 features for classification of valence.



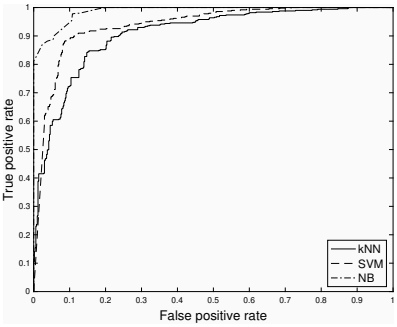
(a) ROC curve using reduced signal and HOC features.



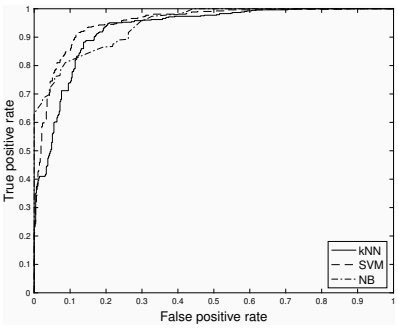
(b) ROC curve using all signal and HOC features.



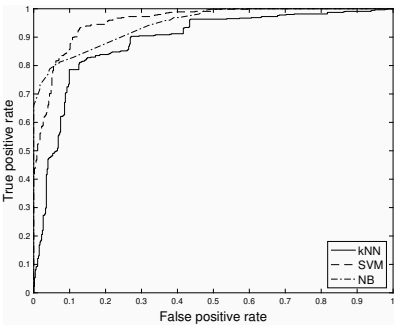
(c) ROC curve using reduced signal and HOS features.



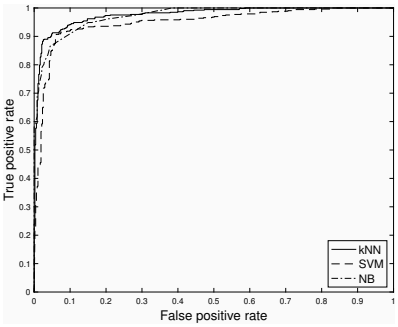
(d) ROC curve using all signal and HOS features.



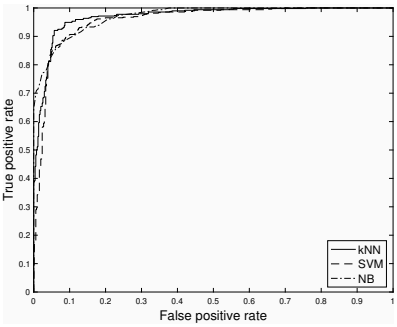
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

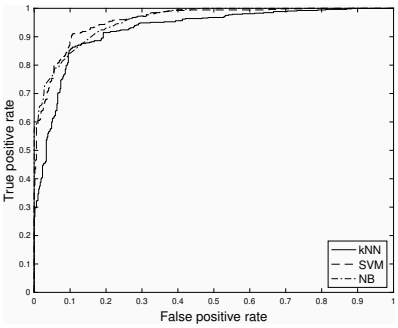


(g) ROC curve using reduced signal and SF.

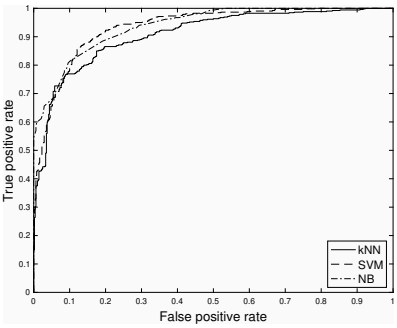


(h) ROC curve using all signal and SF.

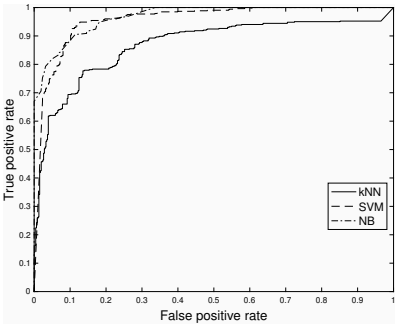
Figure A.14: Results using DEAP dataset and 33 features for classification of valence.



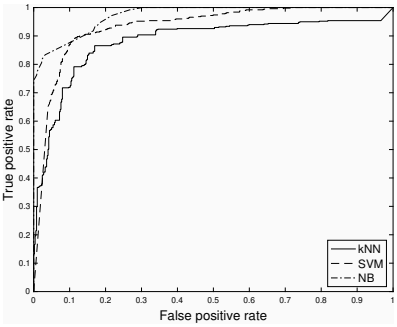
(a) ROC curve using reduced signal and HOC features.



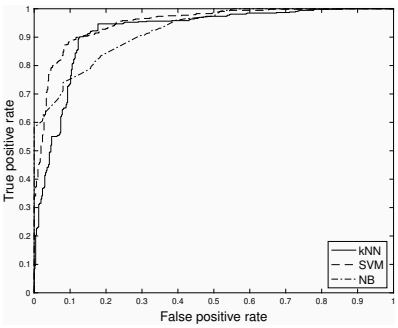
(b) ROC curve using all signal and HOC features.



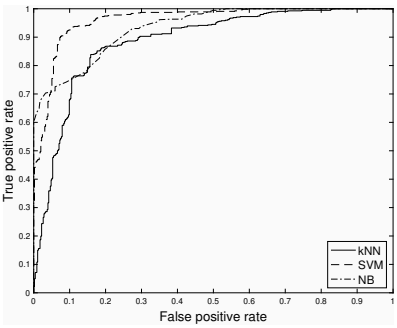
(c) ROC curve using reduced signal and HOS features.



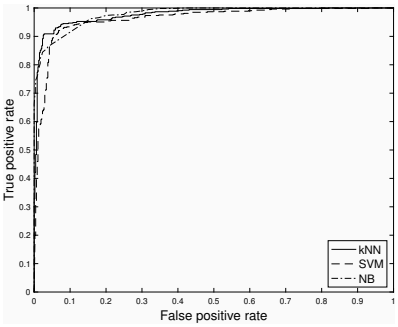
(d) ROC curve using all signal and HOS features.



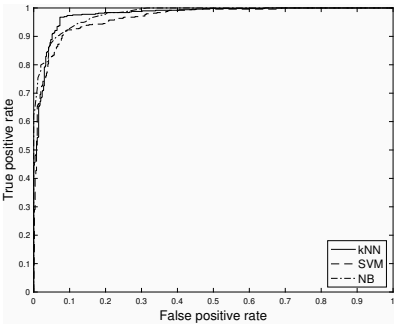
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

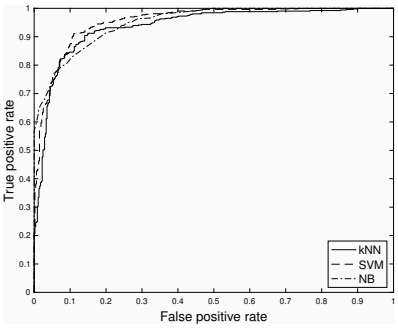


(g) ROC curve using reduced signal and SF.

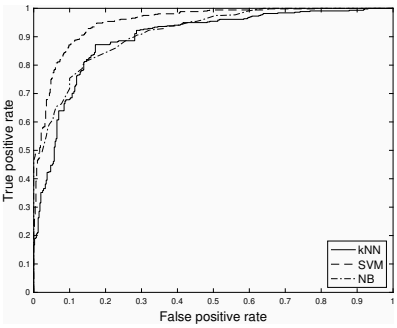


(h) ROC curve using all signal and SF.

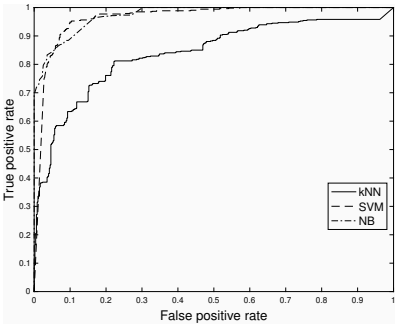
Figure A.15: Results using DEAP dataset and 34 features for classification of valence.



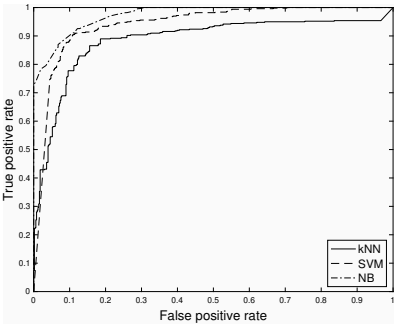
(a) ROC curve using reduced signal and HOC features.



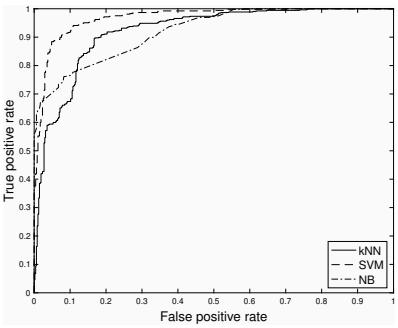
(b) ROC curve using all signal and HOC features.



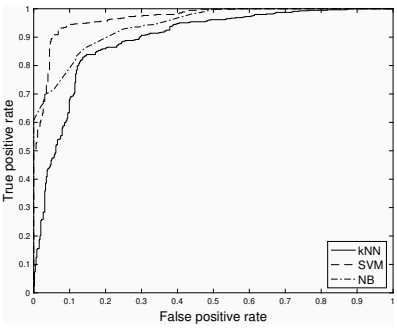
(c) ROC curve using reduced signal and HOS features.



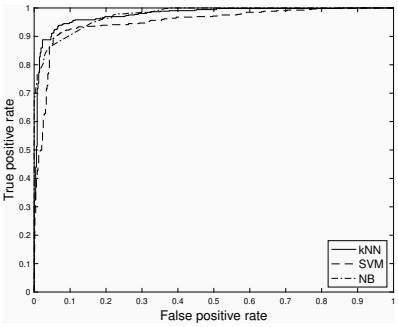
(d) ROC curve using all signal and HOS features.



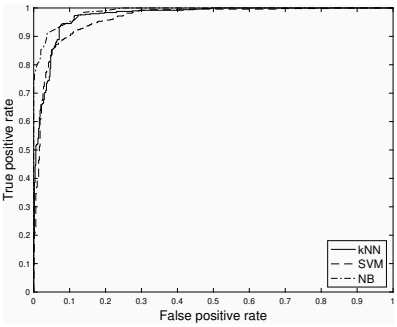
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

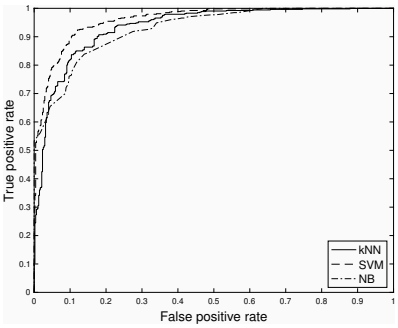


(g) ROC curve using reduced signal and SF.

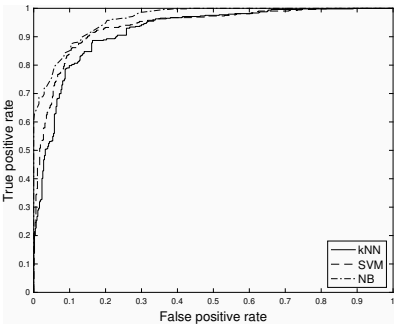


(h) ROC curve using all signal and SF.

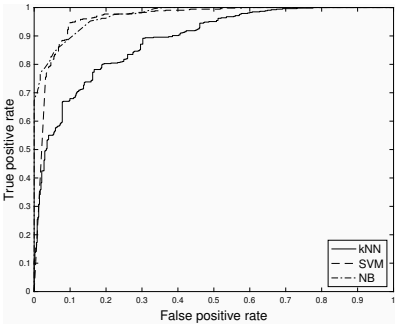
Figure A.16: Results using DEAP dataset and 35 features for classification of valence.



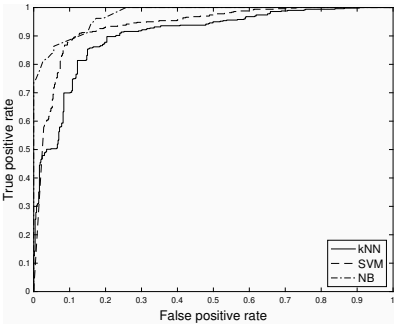
(a) ROC curve using reduced signal and HOC features.



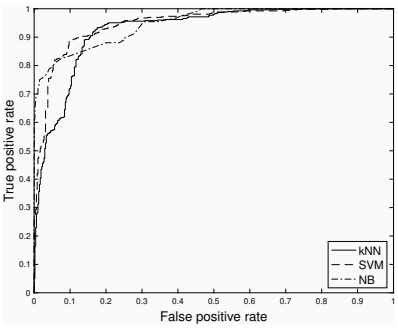
(b) ROC curve using all signal and HOC features.



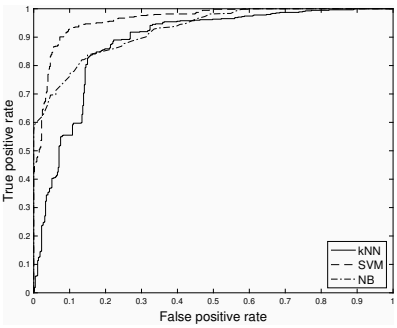
(c) ROC curve using reduced signal and HOS features.



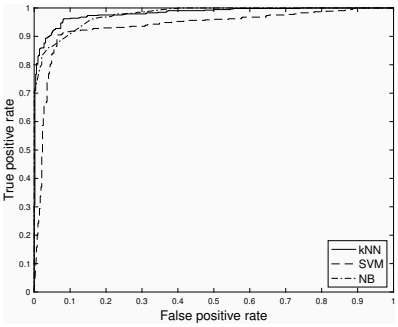
(d) ROC curve using all signal and HOS features.



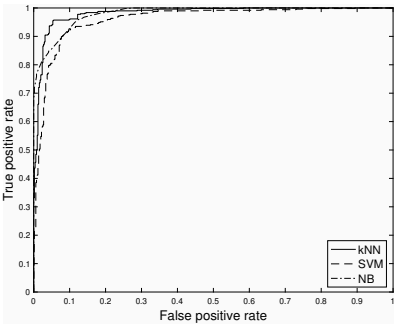
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

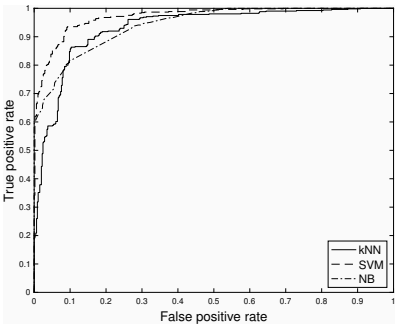


(g) ROC curve using reduced signal and SF.

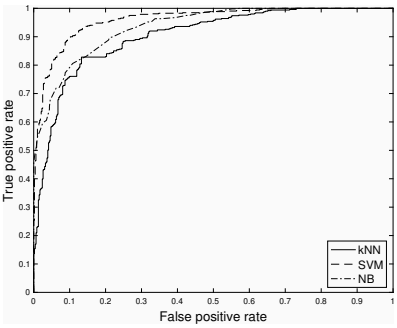


(h) ROC curve using all signal and SF.

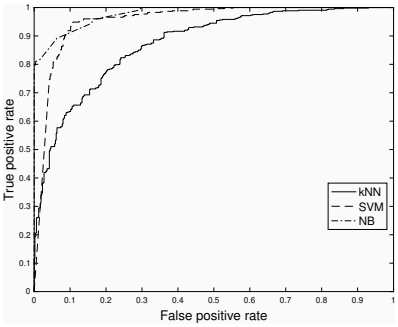
Figure A.17: Results using DEAP dataset and 36 features for classification of valence.



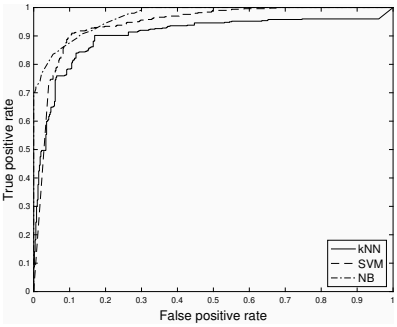
(a) ROC curve using reduced signal and HOC features.



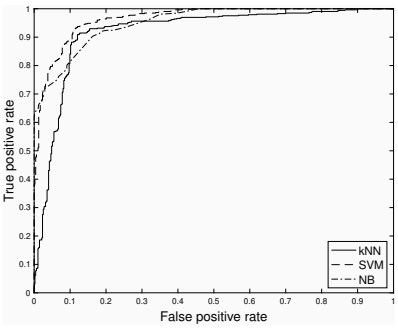
(b) ROC curve using all signal and HOC features.



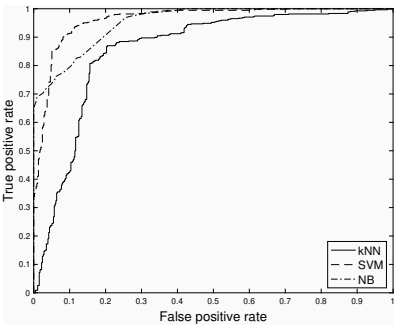
(c) ROC curve using reduced signal and HOS features.



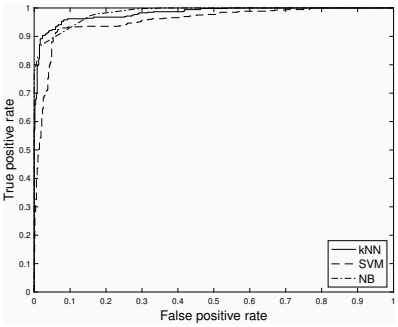
(d) ROC curve using all signal and HOS features.



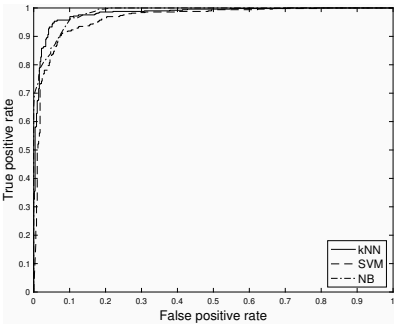
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

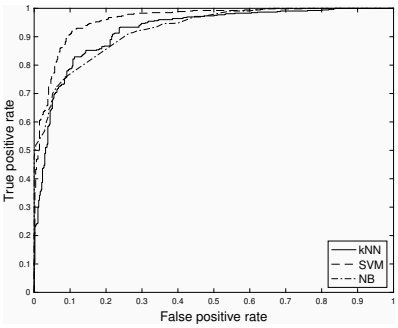


(g) ROC curve using reduced signal and SF.

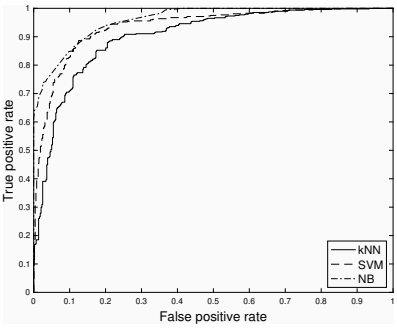


(h) ROC curve using all signal and SF.

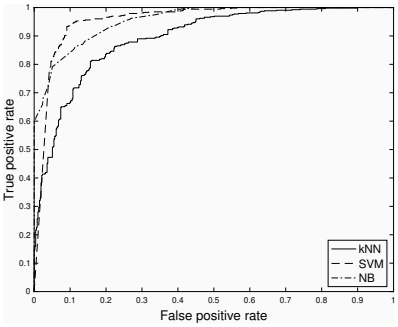
Figure A.18: Results using DEAP dataset and 37 features for classification of valence.



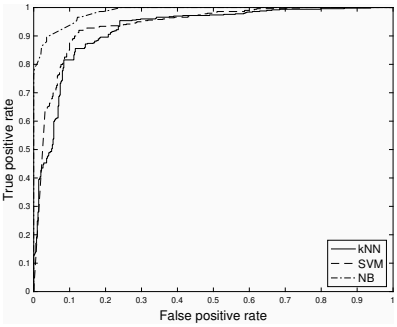
(a) ROC curve using reduced signal and HOC features.



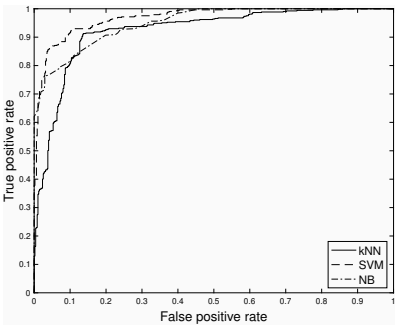
(b) ROC curve using all signal and HOC features.



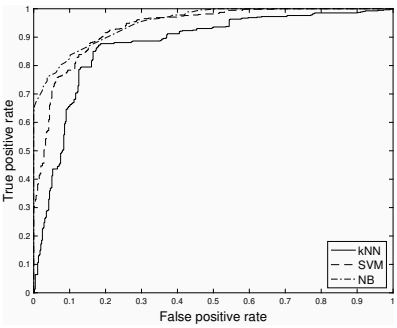
(c) ROC curve using reduced signal and HOS features.



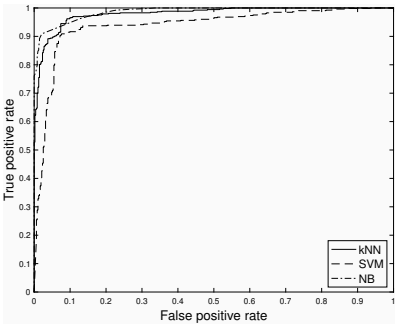
(d) ROC curve using all signal and HOS features.



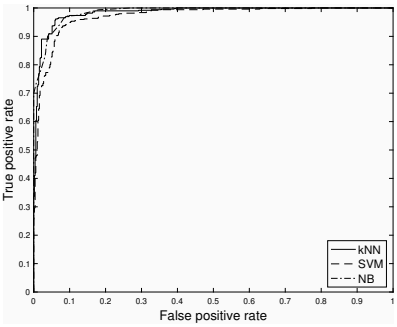
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

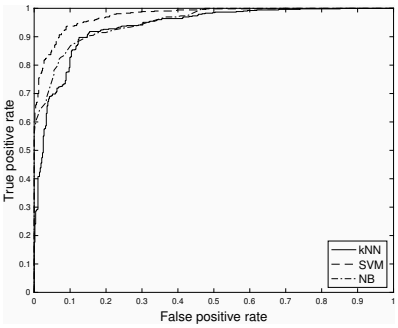


(g) ROC curve using reduced signal and SF.

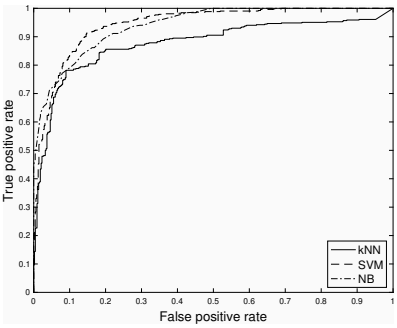


(h) ROC curve using all signal and SF.

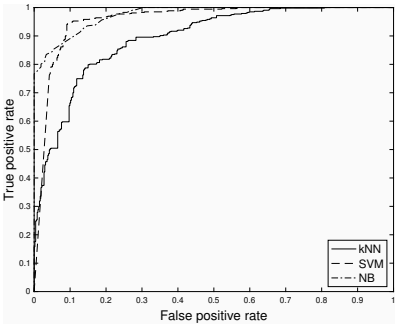
Figure A.19: Results using DEAP dataset and 38 features for classification of valence.



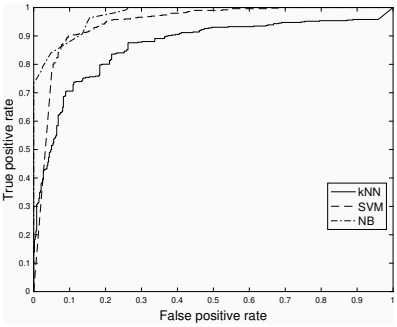
(a) ROC curve using reduced signal and HOC features.



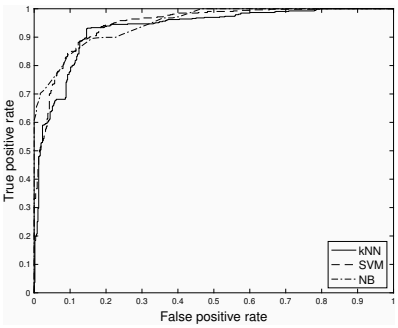
(b) ROC curve using all signal and HOC features.



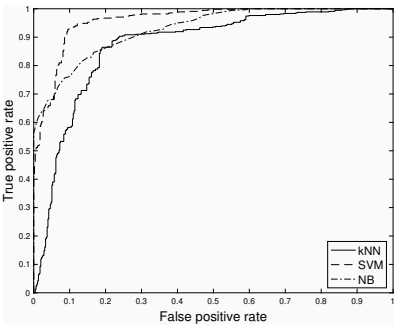
(c) ROC curve using reduced signal and HOS features.



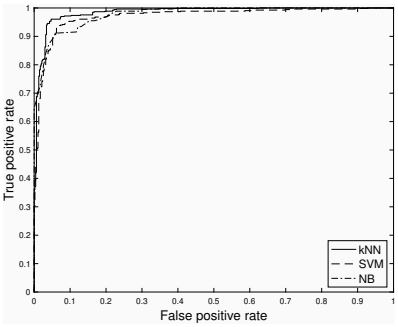
(d) ROC curve using all signal and HOS features.



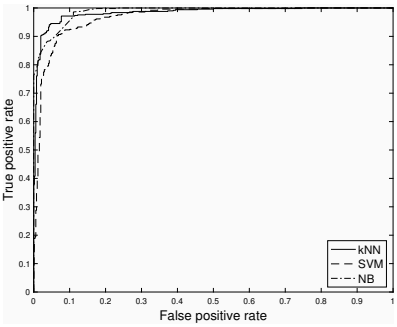
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.



(g) ROC curve using reduced signal and SF.

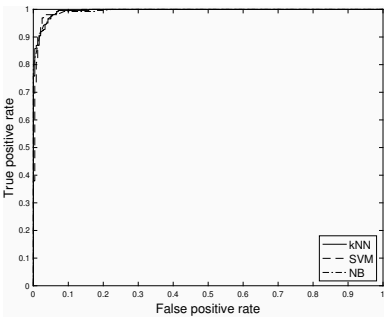


(h) ROC curve using all signal and SF.

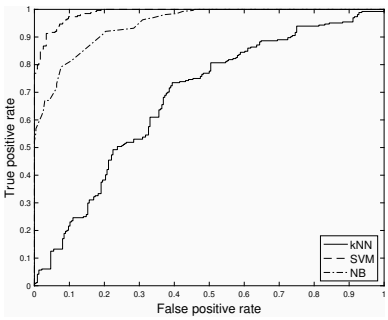
Figure A.20: Results using DEAP dataset and 39 features for classification of valence.

Appendix B

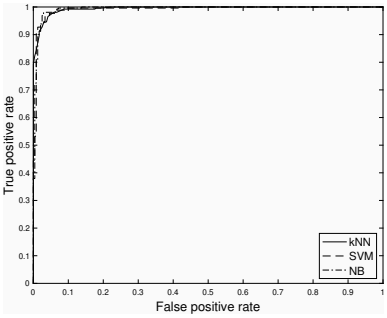
Subject-Dependent Emotion Recognition ROC Curves using MAHNOB dataset



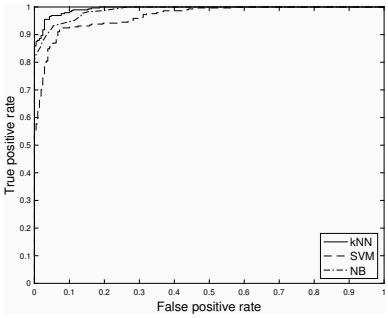
(a) ROC curve using reduced signal and HOC features.



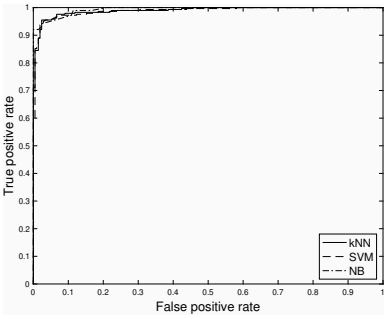
(b) ROC curve using all signal and HOC features.



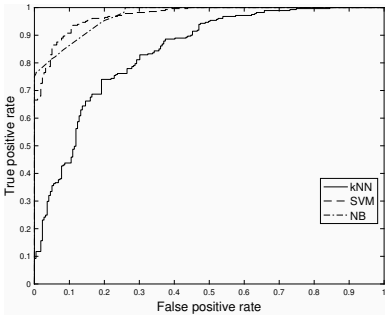
(c) ROC curve using reduced signal and HOS features.



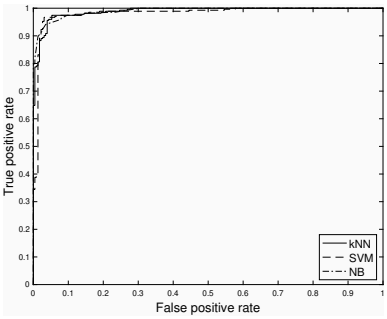
(d) ROC curve using all signal and HOS features.



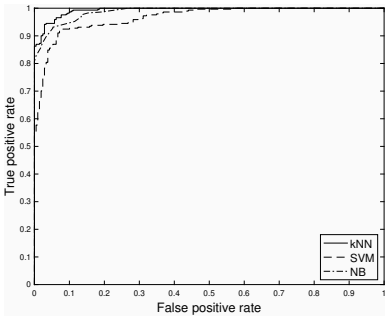
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

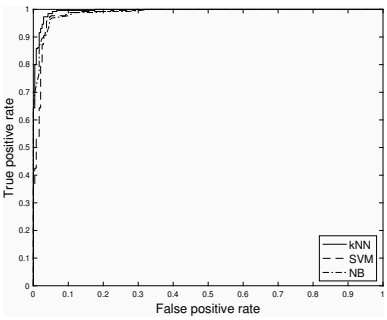


(g) ROC curve using reduced signal and SF.

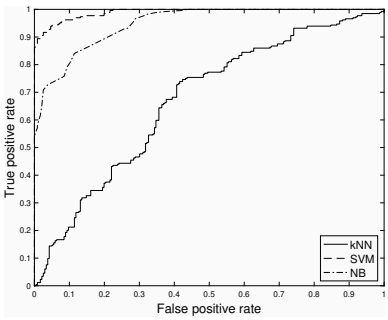


(h) ROC curve using all signal and SF.

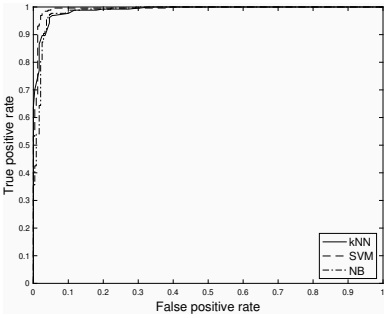
Figure B.1: Results using MAHNOB dataset and 10 features for classification of Arousal.



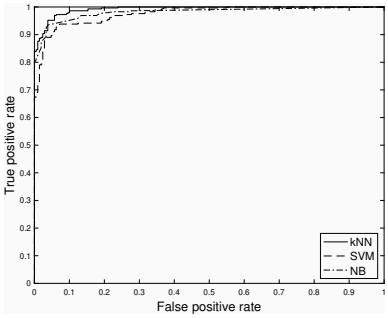
(a) ROC curve using reduced signal and HOC features.



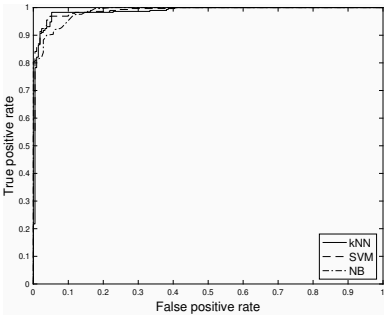
(b) ROC curve using all signal and HOC features.



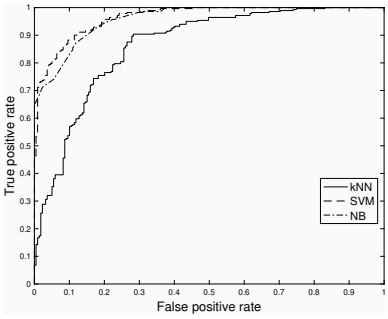
(c) ROC curve using reduced signal and HOS features.



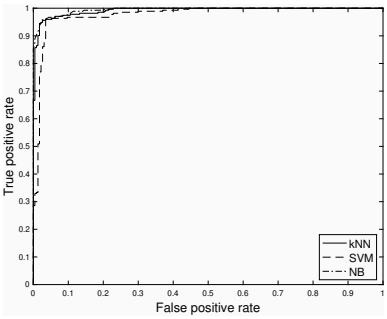
(d) ROC curve using all signal and HOS features.



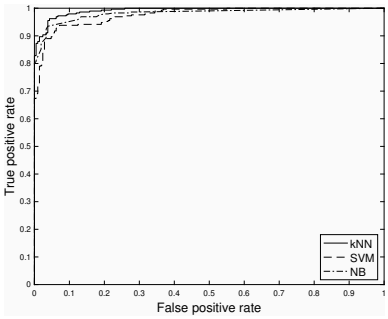
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

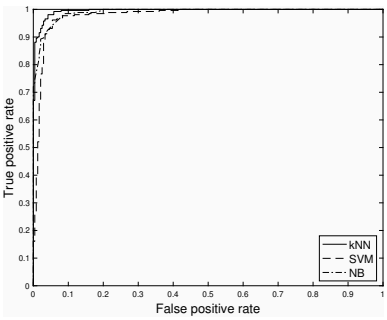


(g) ROC curve using reduced signal and SF.

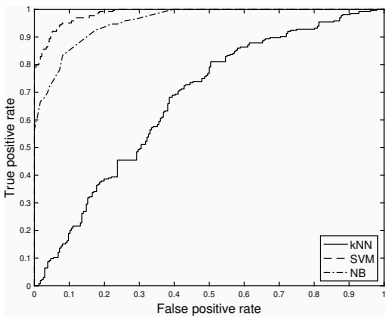


(h) ROC curve using all signal and SF.

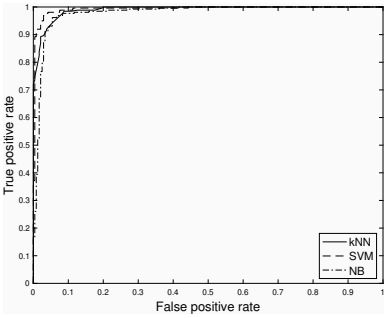
Figure B.2: Results using MAHNOB dataset and 11 features for classification of Arousal.



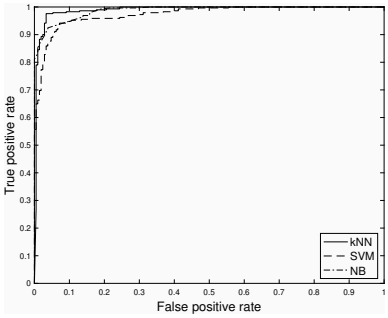
(a) ROC curve using reduced signal and HOC features.



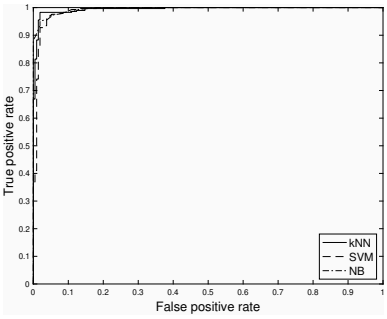
(b) ROC curve using all signal and HOC features.



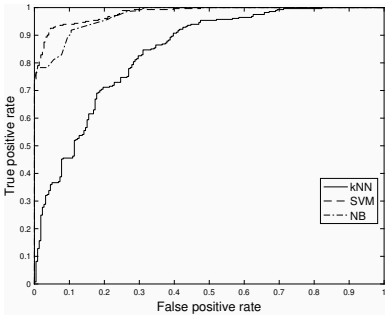
(c) ROC curve using reduced signal and HOS features.



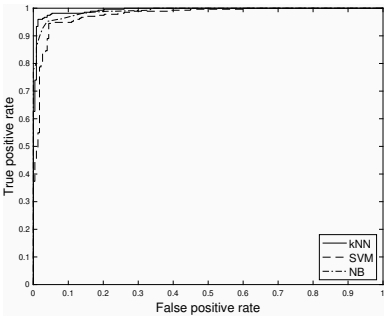
(d) ROC curve using all signal and HOS features.



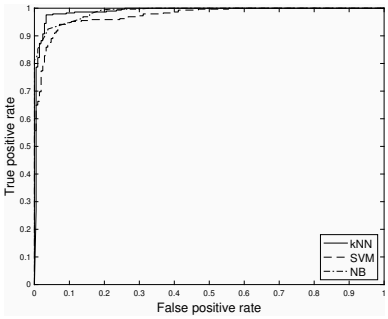
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

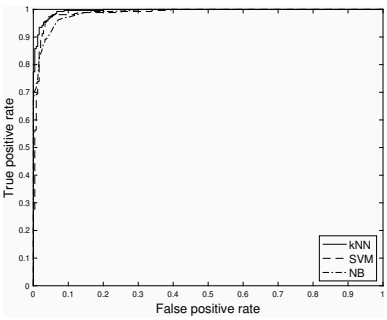


(g) ROC curve using reduced signal and SF.

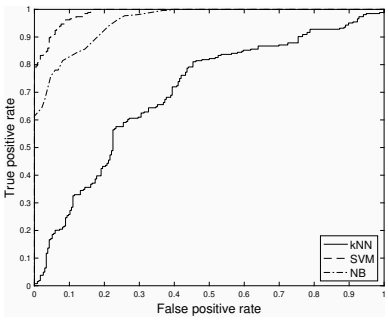


(h) ROC curve using all signal and SF.

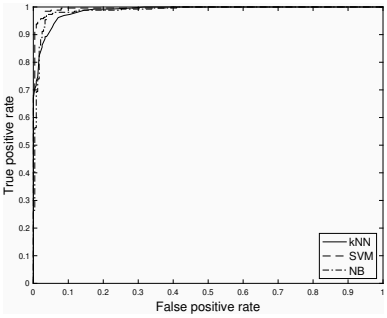
Figure B.3: Results using MAHNOB dataset and 12 features for classification of Arousal.



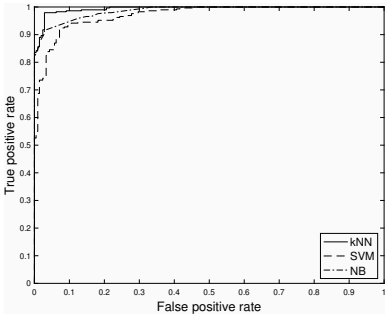
(a) ROC curve using reduced signal and HOC features.



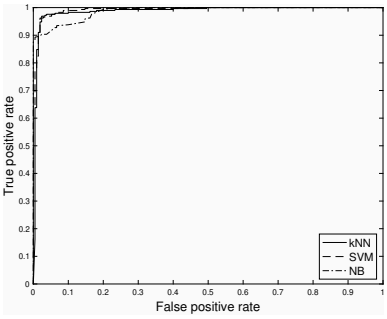
(b) ROC curve using all signal and HOC features.



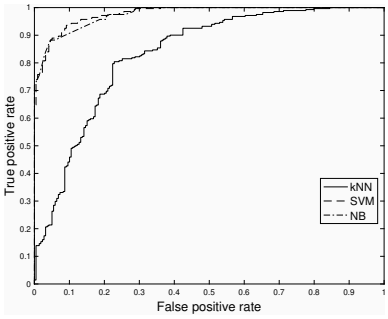
(c) ROC curve using reduced signal and HOS features.



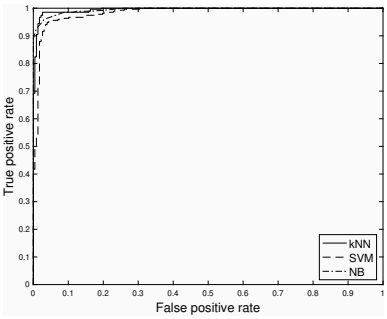
(d) ROC curve using all signal and HOS features.



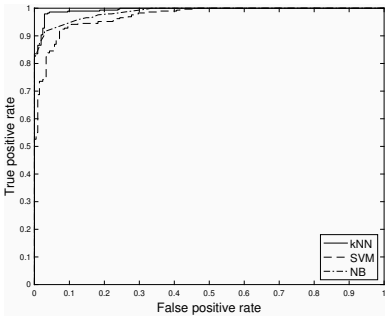
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

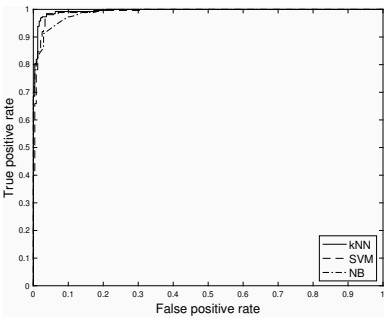


(g) ROC curve using reduced signal and SF.

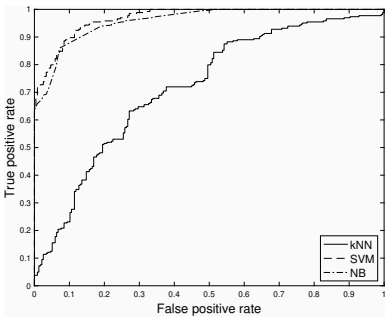


(h) ROC curve using all signal and SF.

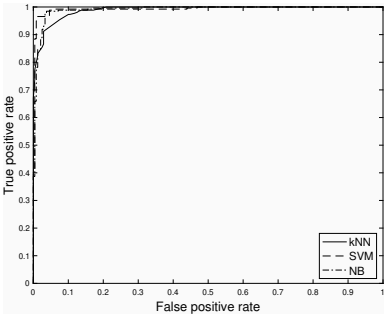
Figure B.4: Results using MAHNOB dataset and 13 features for classification of Arousal.



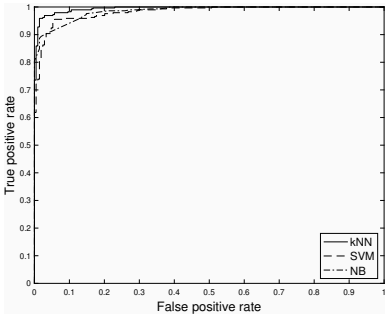
(a) ROC curve using reduced signal and HOC features.



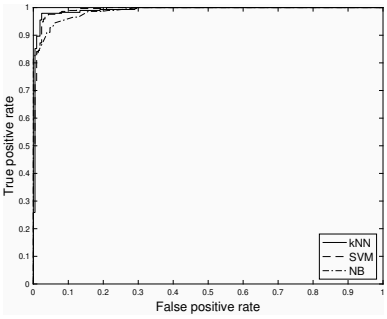
(b) ROC curve using all signal and HOC features.



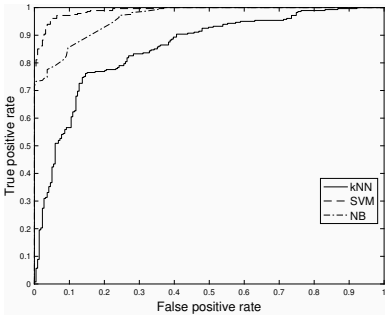
(c) ROC curve using reduced signal and HOS features.



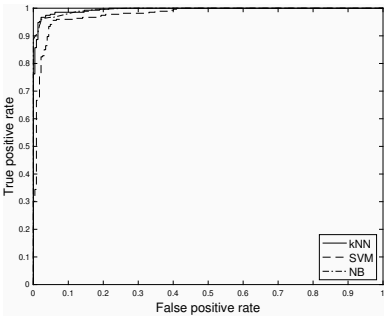
(d) ROC curve using all signal and HOS features.



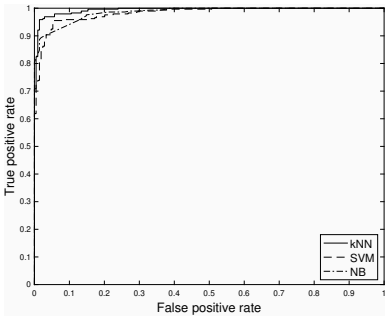
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

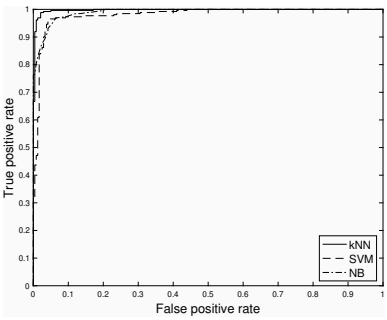


(g) ROC curve using reduced signal and SF.

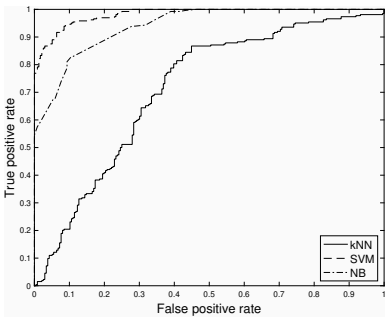


(h) ROC curve using all signal and SF.

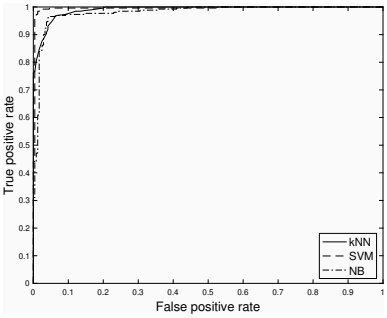
Figure B.5: Results using MAHNOB dataset and 14 features for classification of Arousal.



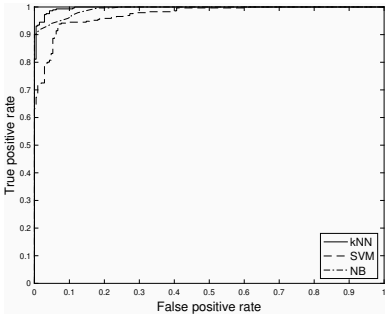
(a) ROC curve using reduced signal and HOC features.



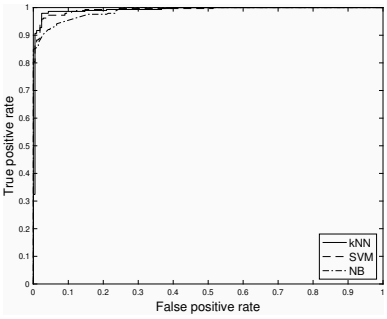
(b) ROC curve using all signal and HOC features.



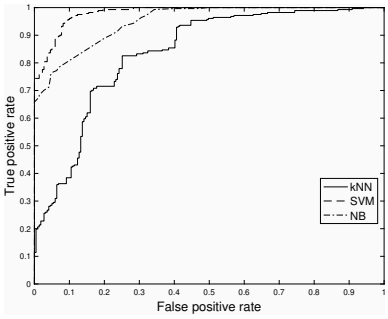
(c) ROC curve using reduced signal and HOS features.



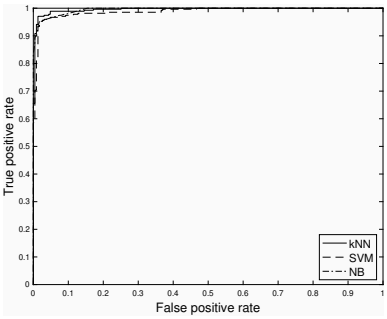
(d) ROC curve using all signal and HOS features.



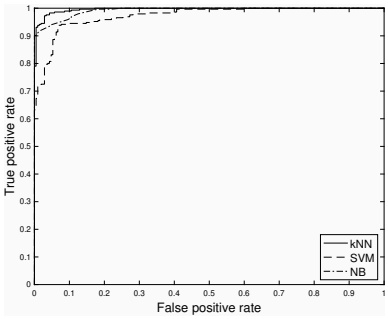
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

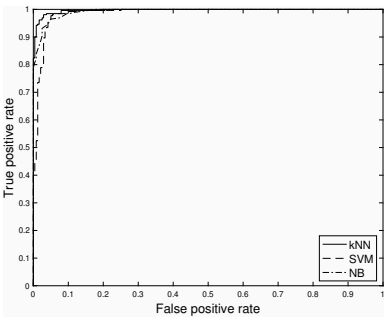


(g) ROC curve using reduced signal and SF.

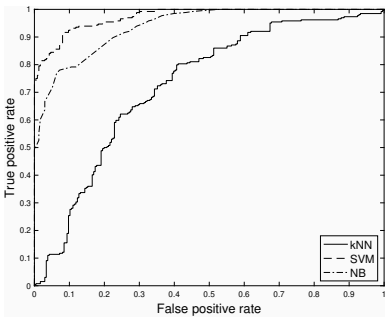


(h) ROC curve using all signal and SF.

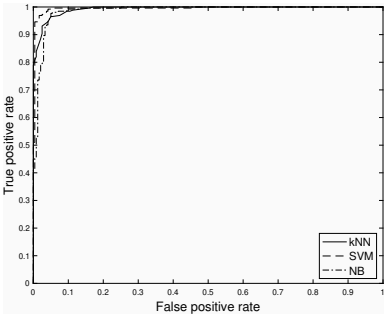
Figure B.6: Results using MAHNOB dataset and 15 features for classification of Arousal.



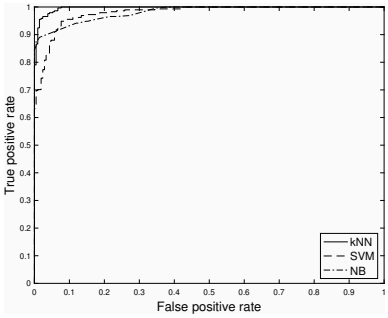
(a) ROC curve using reduced signal and HOC features.



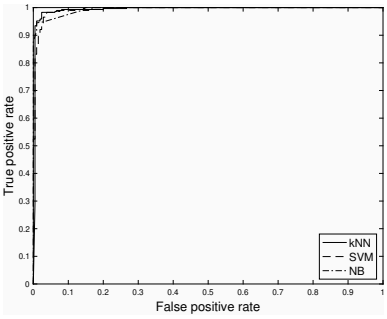
(b) ROC curve using all signal and HOC features.



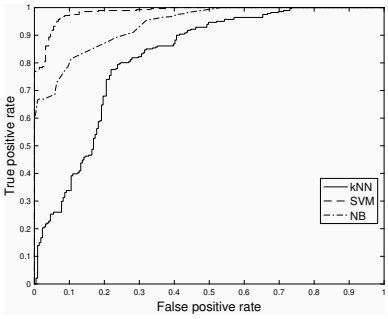
(c) ROC curve using reduced signal and HOS features.



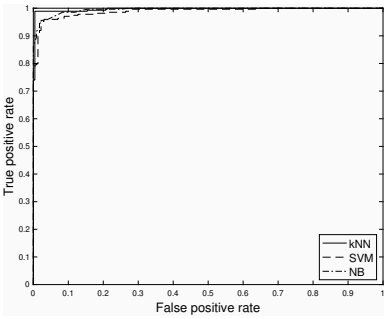
(d) ROC curve using all signal and HOS features.



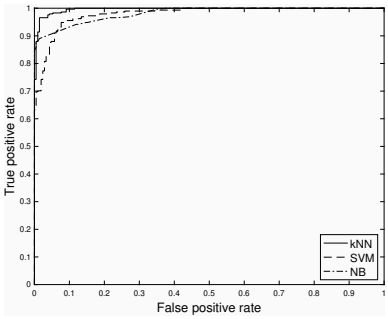
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

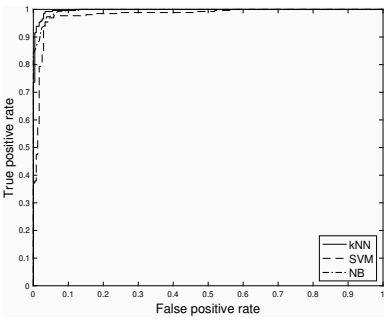


(g) ROC curve using reduced signal and SF.

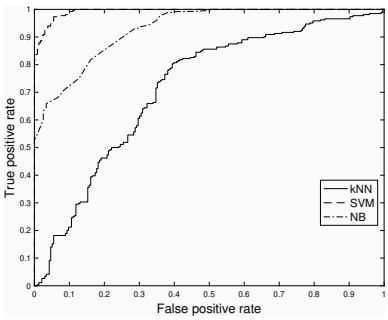


(h) ROC curve using all signal and SF.

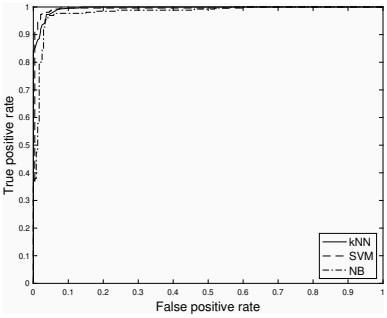
Figure B.7: Results using MAHNOB dataset and 16 features for classification of Arousal.



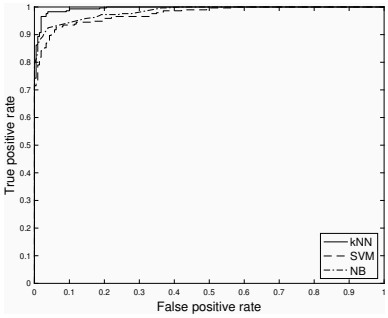
(a) ROC curve using reduced signal and HOC features.



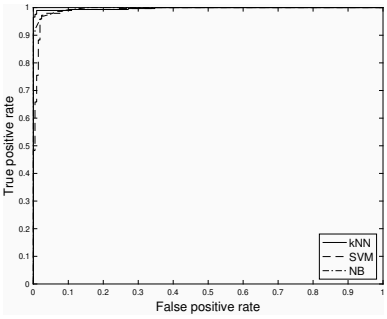
(b) ROC curve using all signal and HOC features.



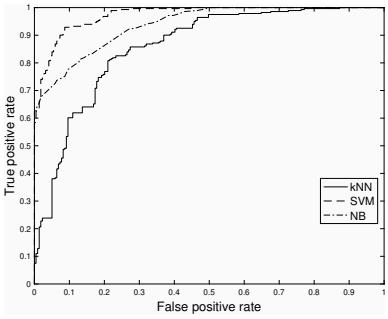
(c) ROC curve using reduced signal and HOS features.



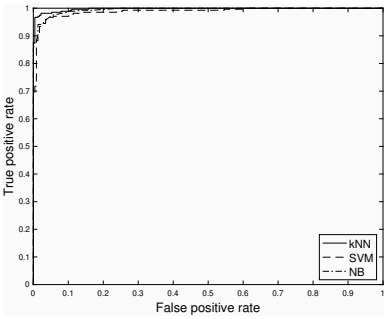
(d) ROC curve using all signal and HOS features.



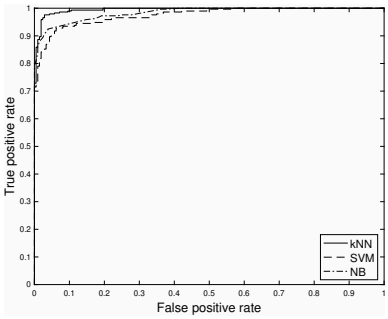
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

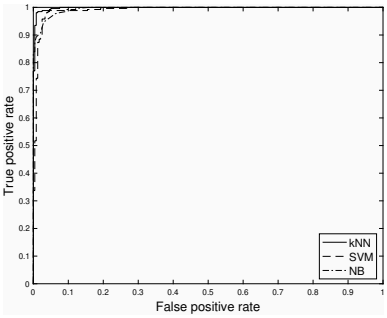


(g) ROC curve using reduced signal and SF.

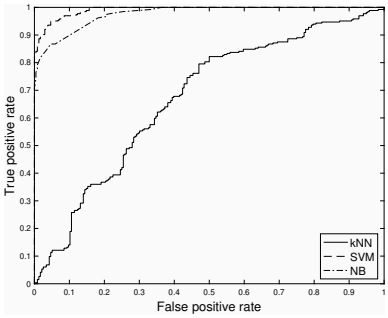


(h) ROC curve using all signal and SF.

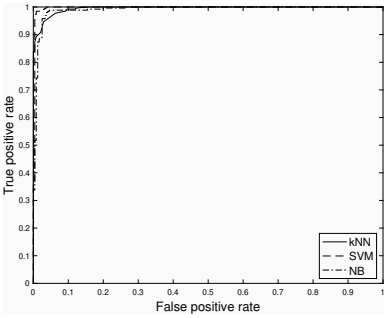
Figure B.8: Results using MAHNOB dataset and 17 features for classification of Arousal.



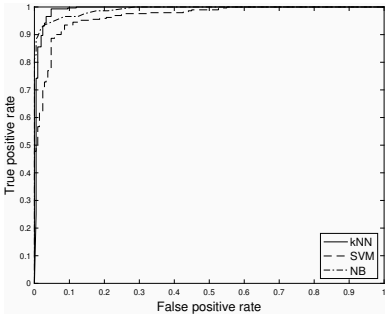
(a) ROC curve using reduced signal and HOC features.



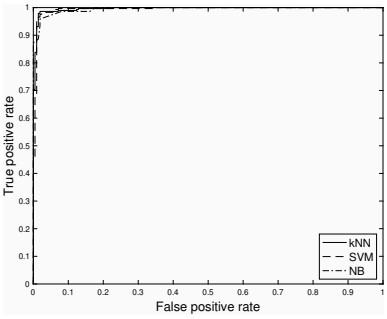
(b) ROC curve using all signal and HOC features.



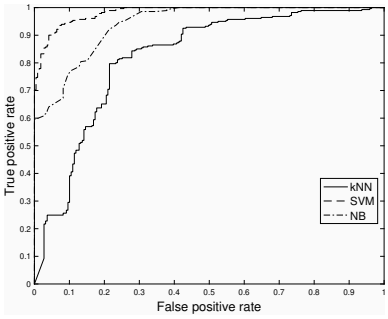
(c) ROC curve using reduced signal and HOS features.



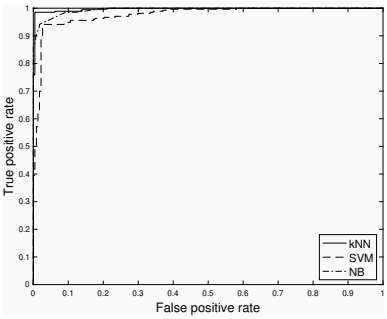
(d) ROC curve using all signal and HOS features.



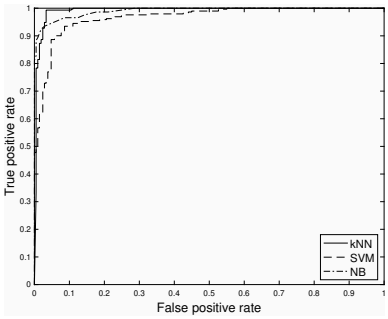
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

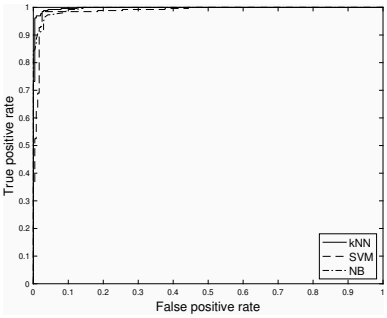


(g) ROC curve using reduced signal and SF.

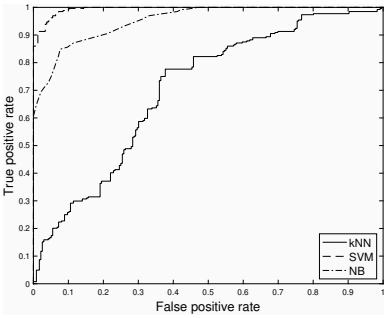


(h) ROC curve using all signal and SF.

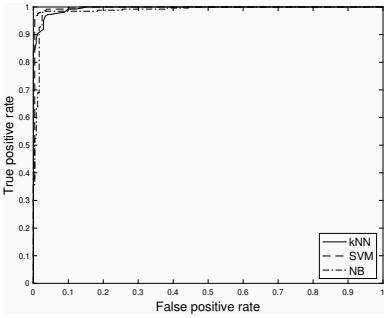
Figure B.9: Results using MAHNOB dataset and 18 features for classification of Arousal.



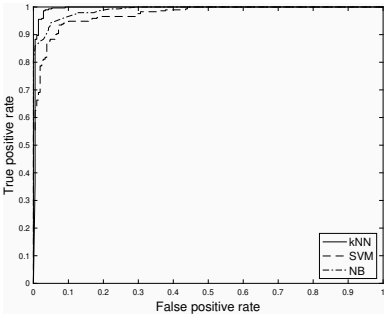
(a) ROC curve using reduced signal and HOC features.



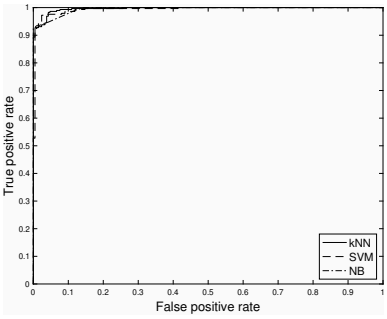
(b) ROC curve using all signal and HOC features.



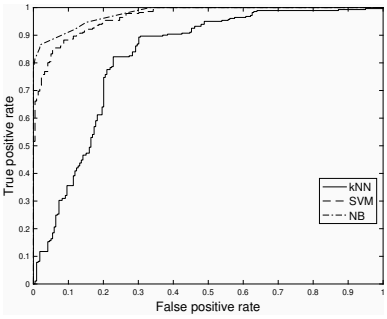
(c) ROC curve using reduced signal and HOS features.



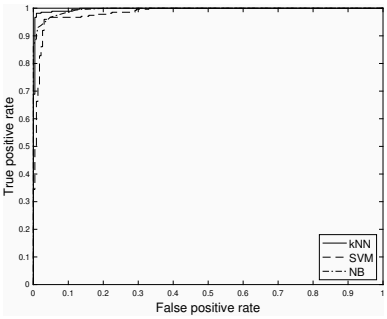
(d) ROC curve using all signal and HOS features.



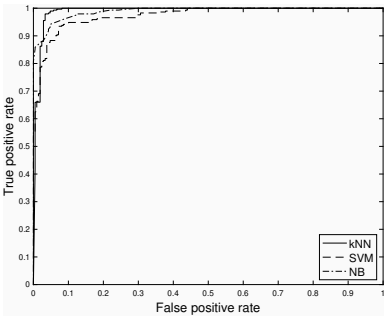
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.



(g) ROC curve using reduced signal and SF.

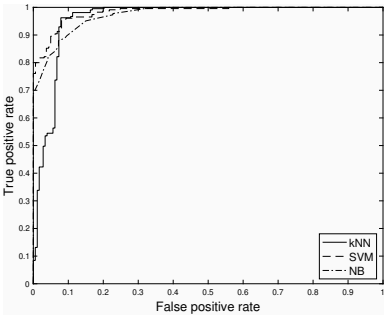


(h) ROC curve using all signal and SF.

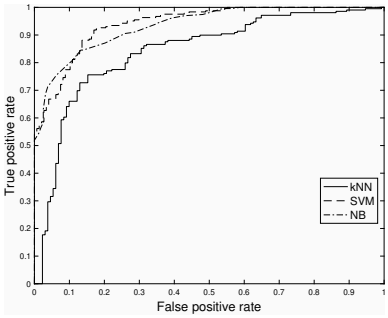
Figure B.10: Results using MAHNOB dataset and 19 features for classification of Arousal.

Appendix C

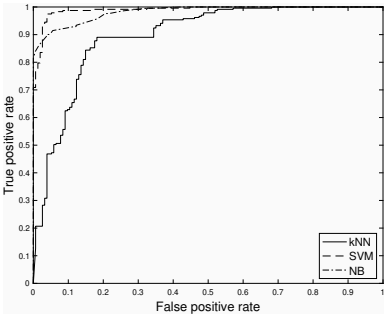
Subject-Dependent Emotion Recognition ROC Curves using DREAMER dataset



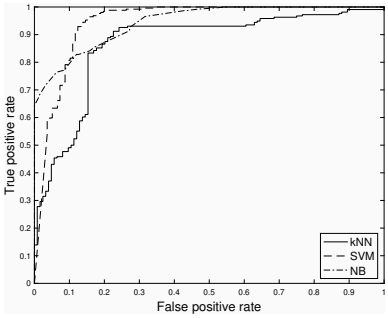
(a) ROC curve using reduced signal and HOC features.



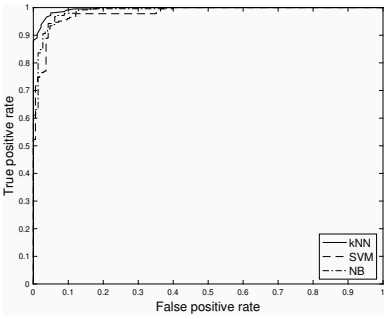
(b) ROC curve using all signal and HOC features.



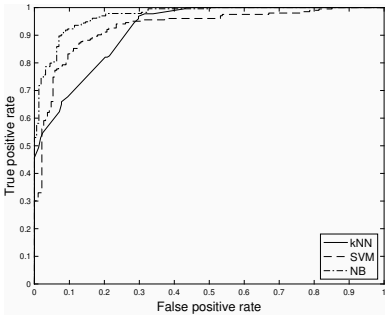
(c) ROC curve using reduced signal and HOS features.



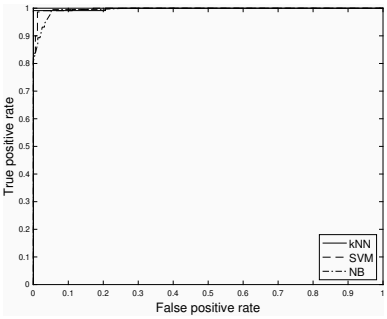
(d) ROC curve using all signal and HOS features.



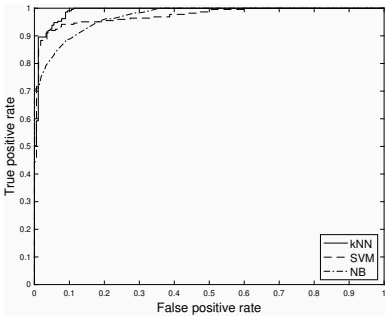
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

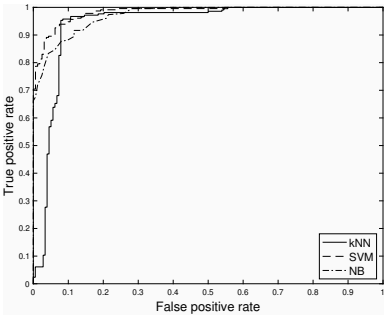


(g) ROC curve using reduced signal and SF.

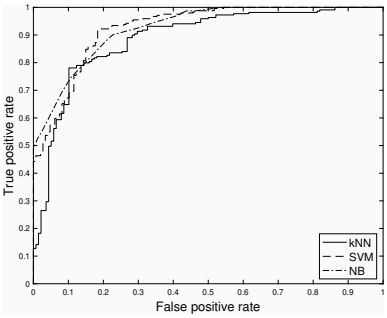


(h) ROC curve using all signal and SF.

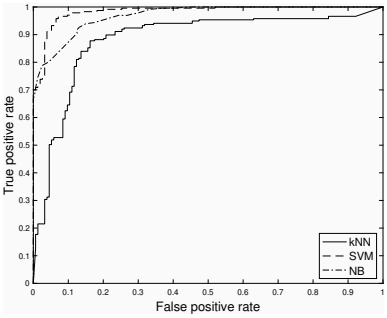
Figure C.1: Results using DREAMER dataset and 14 features for classification of Arousal.



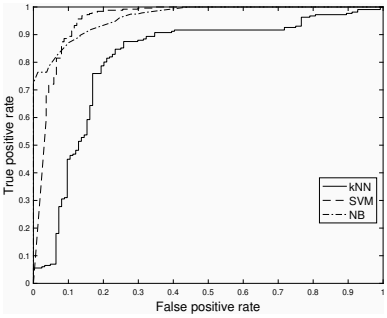
(a) ROC curve using reduced signal and HOC features.



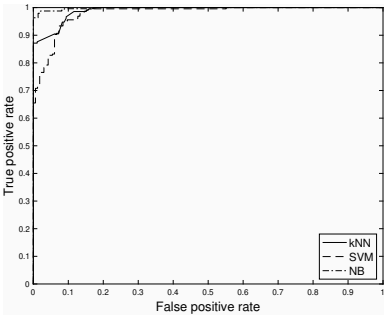
(b) ROC curve using all signal and HOC features.



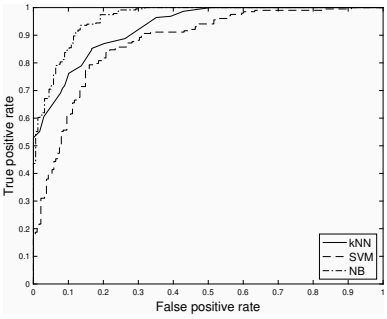
(c) ROC curve using reduced signal and HOS features.



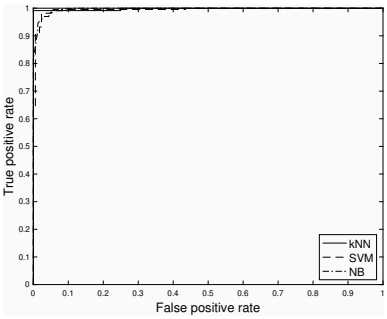
(d) ROC curve using all signal and HOS features.



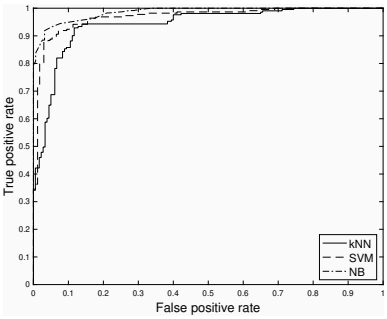
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

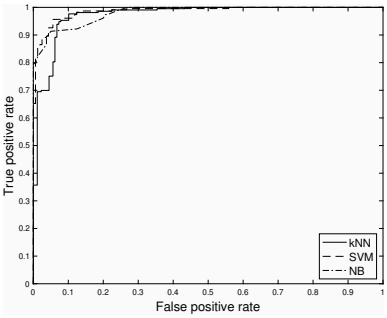


(g) ROC curve using reduced signal and SF.

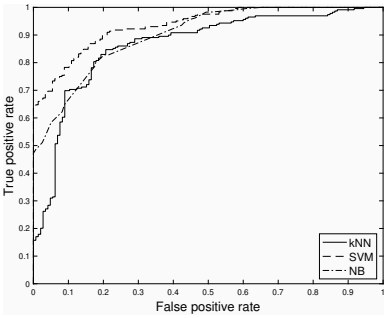


(h) ROC curve using all signal and SF.

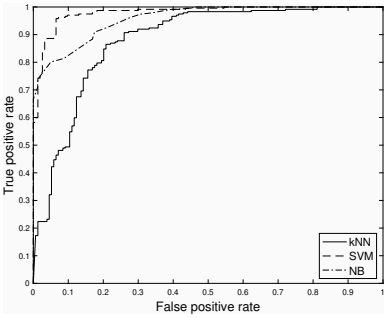
Figure C.2: Results using DREAMER dataset and 15 features for classification of Arousal.



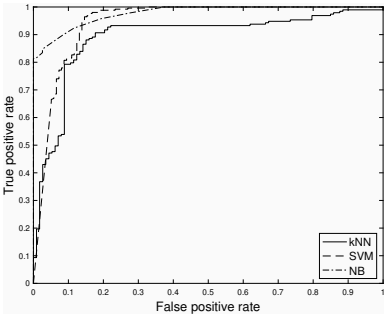
(a) ROC curve using reduced signal and HOC features.



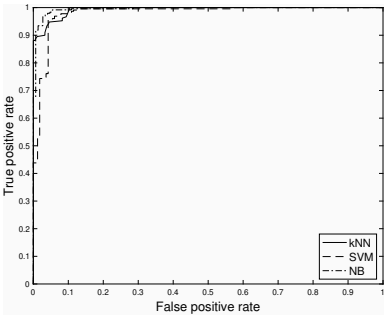
(b) ROC curve using all signal and HOC features.



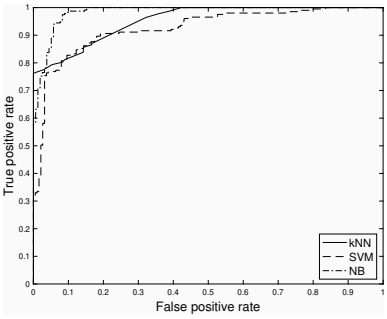
(c) ROC curve using reduced signal and HOS features.



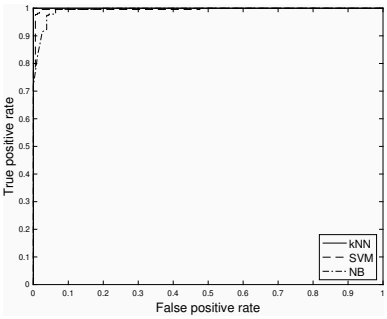
(d) ROC curve using all signal and HOS features.



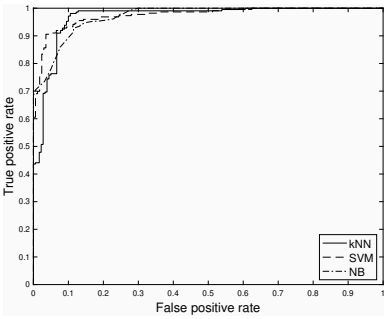
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

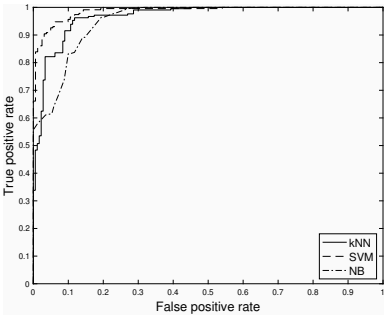


(g) ROC curve using reduced signal and SF.

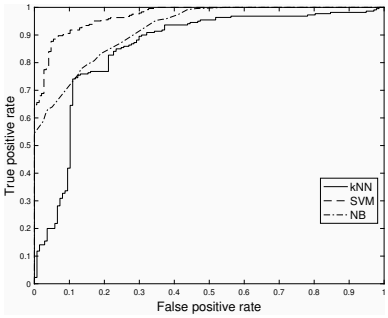


(h) ROC curve using all signal and SF.

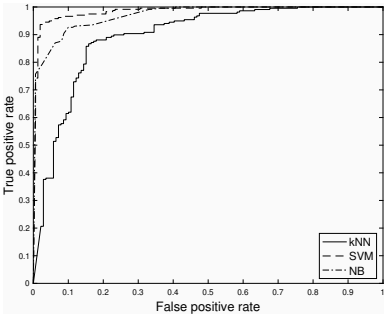
Figure C.3: Results using DREAMER dataset and 16 features for classification of Arousal.



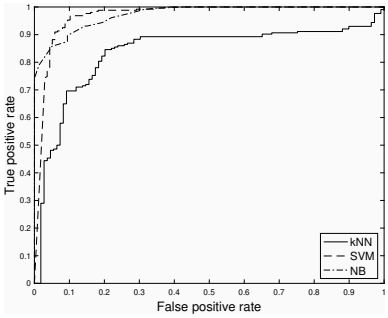
(a) ROC curve using reduced signal and HOC features.



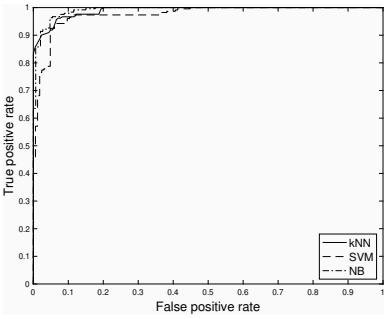
(b) ROC curve using all signal and HOC features.



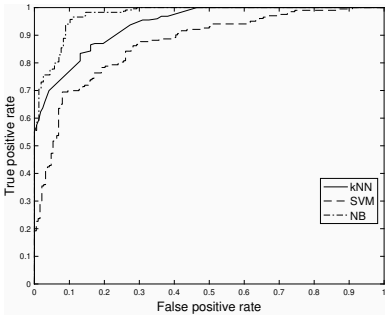
(c) ROC curve using reduced signal and HOS features.



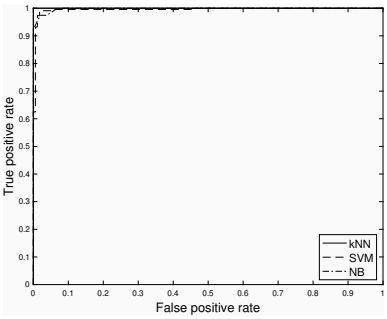
(d) ROC curve using all signal and HOS features.



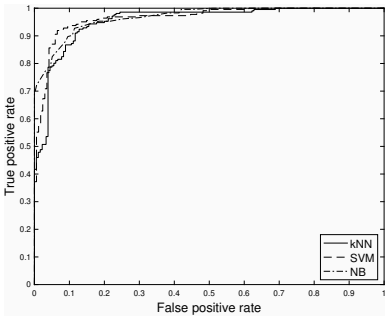
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

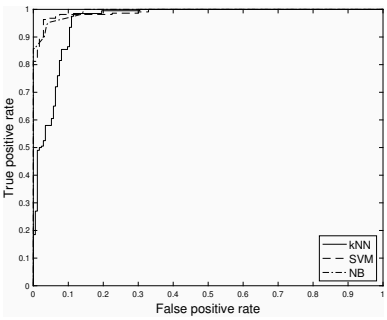


(g) ROC curve using reduced signal and SF.

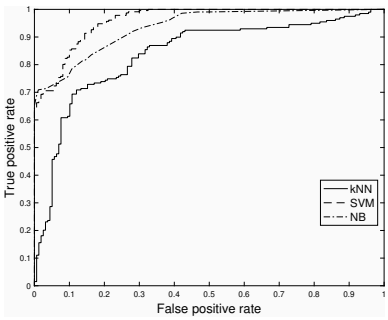


(h) ROC curve using all signal and SF.

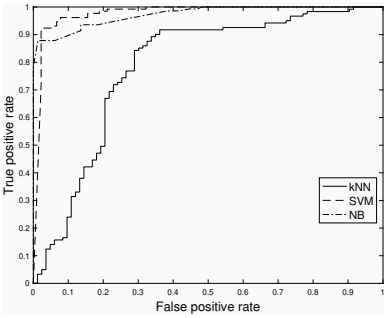
Figure C.4: Results using DREAMER dataset and 17 features for classification of Arousal.



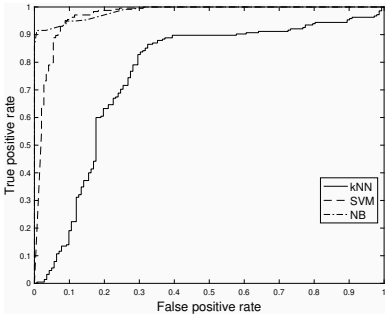
(a) ROC curve using reduced signal and HOC features.



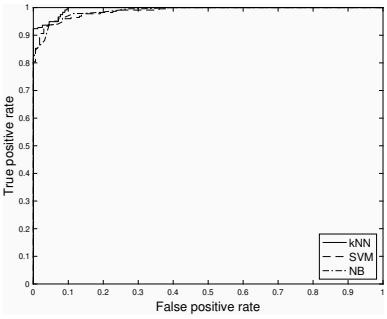
(b) ROC curve using all signal and HOC features.



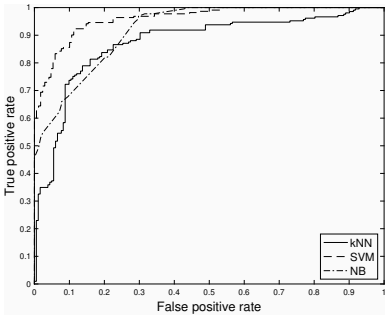
(c) ROC curve using reduced signal and HOS features.



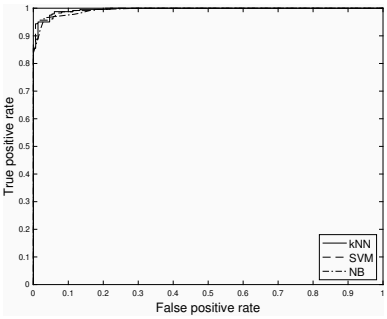
(d) ROC curve using all signal and HOS features.



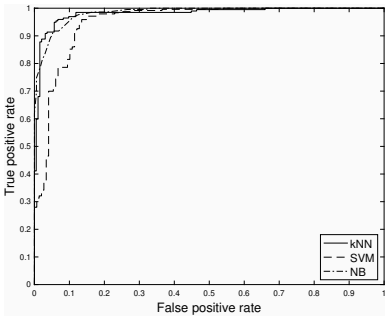
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

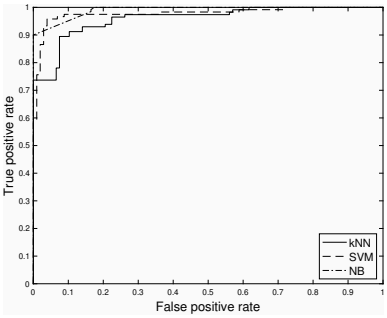


(g) ROC curve using reduced signal and SF.

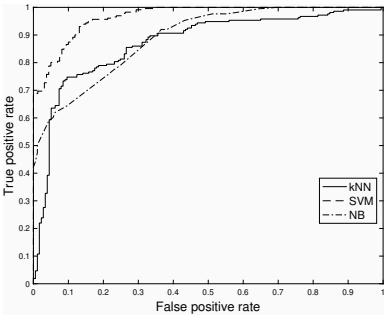


(h) ROC curve using all signal and SF.

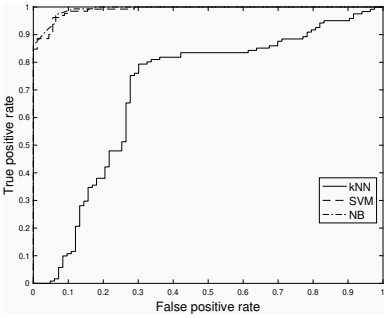
Figure C.5: Results using DREAMER dataset and 14 features for classification of valence.



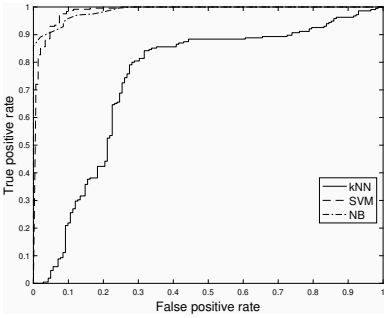
(a) ROC curve using reduced signal and HOC features.



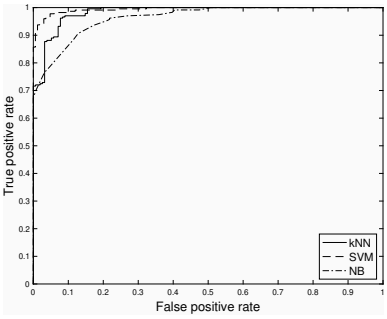
(b) ROC curve using all signal and HOC features.



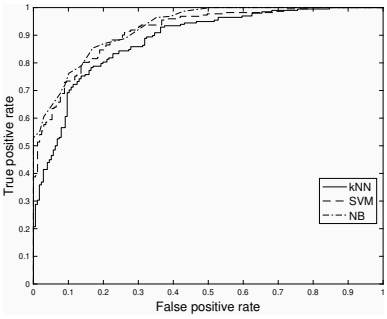
(c) ROC curve using reduced signal and HOS features.



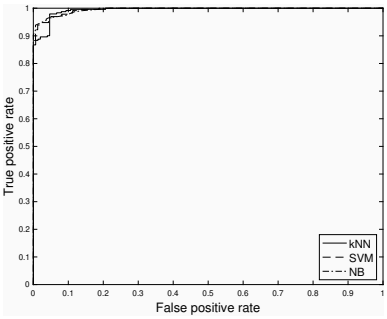
(d) ROC curve using all signal and HOS features.



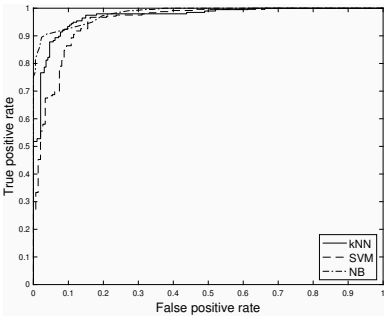
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

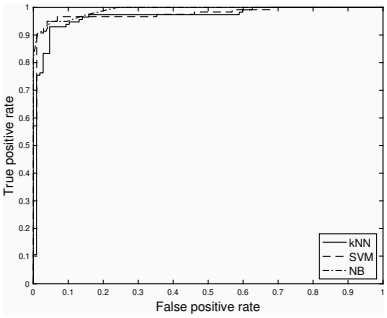


(g) ROC curve using reduced signal and SF.

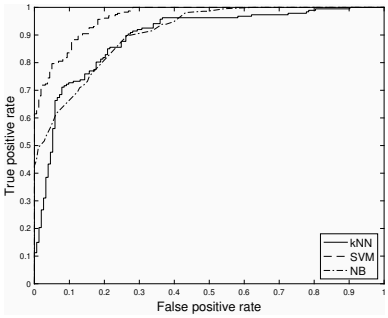


(h) ROC curve using all signal and SF.

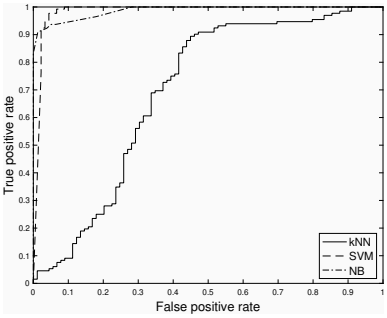
Figure C.6: Results using DREAMER dataset and 15 features for classification of Arousal.



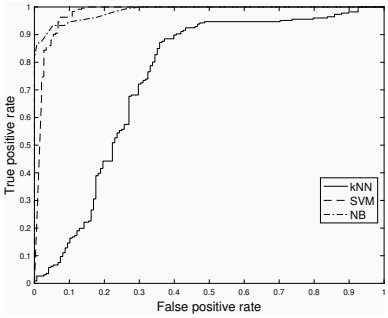
(a) ROC curve using reduced signal and HOC features.



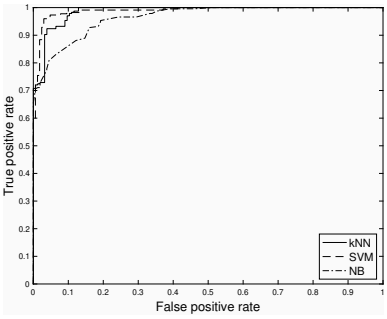
(b) ROC curve using all signal and HOC features.



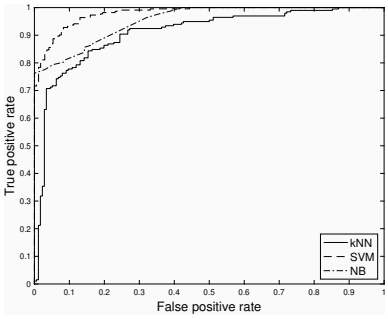
(c) ROC curve using reduced signal and HOS features.



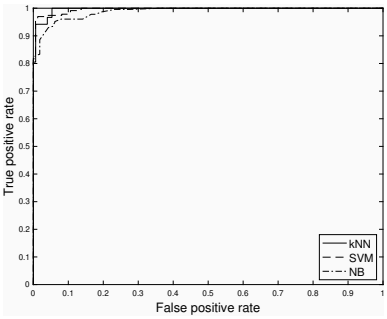
(d) ROC curve using all signal and HOS features.



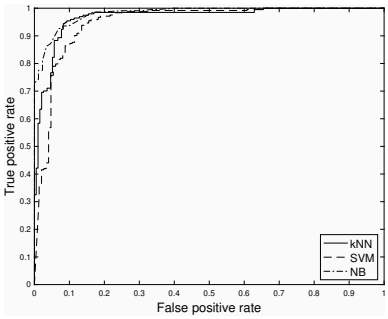
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.

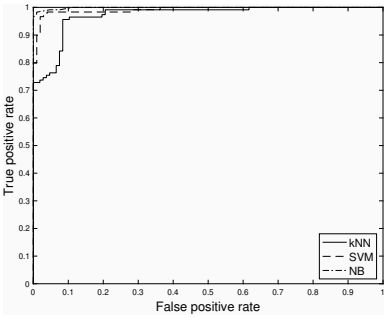


(g) ROC curve using reduced signal and SF.

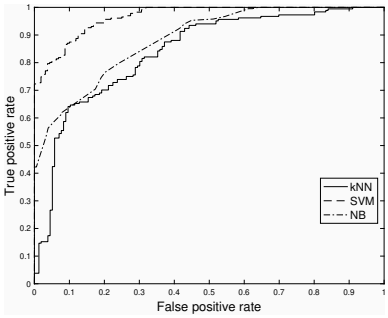


(h) ROC curve using all signal and SF.

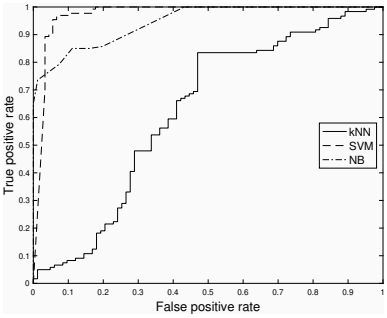
Figure C.7: Results using DREAMER dataset and 16 features for classification of valence.



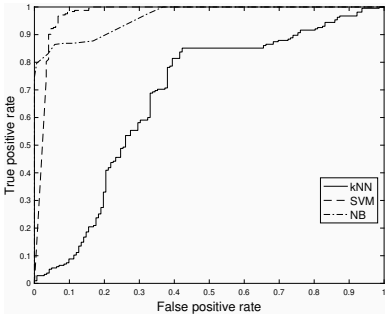
(a) ROC curve using reduced signal and HOC features.



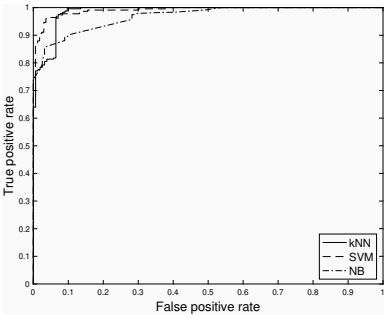
(b) ROC curve using all signal and HOC features.



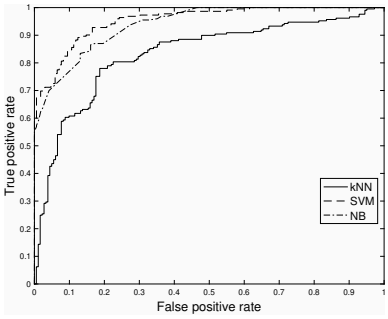
(c) ROC curve using reduced signal and HOS features.



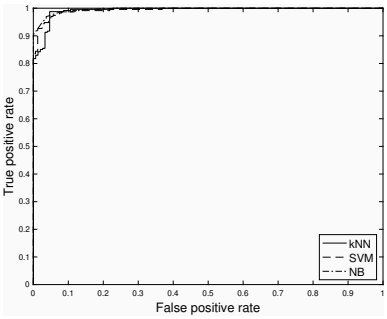
(d) ROC curve using all signal and HOS features.



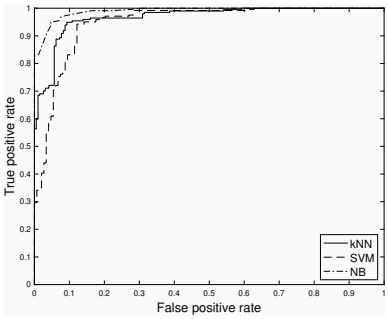
(e) ROC curve using reduced signal and PSE features.



(f) ROC curve using all signal and PSE features.



(g) ROC curve using reduced signal and SF.



(h) ROC curve using all signal and SF.

Figure C.8: Results using DREAMER dataset and 17 features for classification of valence.

Bibliography

[1] Mojtaba Khomami Abadi, Mostafa Kia, Ramanathan Subramanian, Paolo Avesani, and Nicu Sebe. Decoding affect in videos employing the meg brain signal. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[2] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *Affective Computing, IEEE Transactions on*, 6(3):209–222, 2015.

[3] Fadi Al Machot, Ali Elmachot, Mouhannad Ali, Elyan Al Machot, and Kyandoghere Kyamakya. A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors. *Sensors*, 19(7):1659, 2019.

[4] Hasimah Ali, Muthusamy Hariharan, Sazali Yaacob, and Abdul Hamid Adom. Facial emotion recognition using empirical mode decomposition. *Expert Systems with Applications*, 42(3):1261–1277, 2015.

[5] Theus H Aspiras and Vijayan K Asari. Log power representation of eeg spectral bands for the recognition of emotional states of mind. In *2011 8th International Conference on Information, Communications & Signal Processing*, pages 1–5. IEEE, 2011.

[6] John Atkinson and Daniel Campos. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers. *Expert Systems with Applications*, 47:35–41, 2016.

[7] Anirban Basu and Anisha Halder. Facial expression and eeg signal based classification of emotion. In *Electronics, Communication and Instrumentation (ICECI), 2014 International Conference on*, pages 1–4. IEEE, 2014.

[8] Adel Belouchrani, Karim Abed-Meraim, J-F Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 45(2):434–444, 1997.

[9] Danny Oude Bos et al. Eeg-based emotion recognition. *The Influence of Visual and Auditory Stimuli*, 56(3):1–17, 2006.

[10] Lachezar Bozhkov, Petia Georgieva, Isabel Santos, Ana Pereira, and Carlos Silva. Eeg-based subject independent affective computing models. *Procedia Computer Science*, 53:375–382, 2015.

[11] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[12] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual. technical report b-3. Technical report, University of Florida, Gainesville, Florida, 2007.

[13] Sandra Carvalho, Jorge Leite, Santiago Galdo-Álvarez, and Óscar F Gonçalves. The emotional movie database (emdb): A self-report and psychophysiological study. *Applied psychophysiology and biofeedback*, 37(4):279–294, 2012.

[14] Guillaume Chanel, Joep JM Kierkels, Mohammad Soleymani, and Thierry Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, 2009.

[15] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(6):1052–1063, 2011.

[16] Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion/expression recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 366–371. IEEE, 1998.

[17] Ingrid Daubechies. *Ten lectures on wavelets*, volume 61. Siam, 1992.

[18] Liyanage C De Silva and Pei Chi Ng. Bimodal emotion recognition. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 332–335. IEEE, 2000.

[19] Joseph Dien. Issues in the application of the average reference: Review, critiques, and recommendations. *Behavior Research Methods, Instruments, & Computers*, 30(1):34–43, 1998.

[20] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013.

[21] Richard O Duda, Peter E Hart, and David G Stork. Pattern classification and scene analysis 2nd ed. *ed: Wiley Interscience*, 1995.

[22] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[23] Richard A Fabes and Carol Lynn Martin. Gender and age stereotypes of emotionality. *Personality and social psychology bulletin*, 17(5):532–540, 1991.

[24] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[25] Hernán F García, Mauricio A Álvarez, and Álvaro A Orozco. Gaussian process dynamical models for multimodal affect recognition. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 850–853. IEEE, 2016.

[26] Walter Glannon. Neuromodulation, agency and autonomy. *Brain Topography*, 27(1):46–54, 2014.

[27] Germán Gómez-Herrero. Automatic artifact removal (aar) toolbox v1. 3 (release 09.12. 2007) for matlab. *Tampere University of Technology*, 2007.

[28] Germán Gómez-Herrero, Zbyněk Koldovskỳ, Petr Tichavskỳ, and Karen Egiazarian. A fast algorithm for blind separation of non-gaussian and time-correlated signals. In *2007 15th European Signal Processing Conference*, pages 1731–1735. IEEE, 2007.

[29] Bernhard Graimann, Brendan Allison, and Gert Pfurtscheller. Brain–computer interfaces: A gentle introduction. In *Brain-Computer Interfaces*, pages 1–27. Springer, 2009.

[30] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[31] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*, 2014.

[32] Pim Haselager, Rutger Vlek, Jeremy Hill, and Femke Nijboer. A note on ethical aspects of bci. *Neural Networks*, 22(9):1352–1357, 2009.

[33] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970.

[34] Seyyed Abed Hosseini, Mohammad Ali Khalilzadeh, Mohammad Bagher Naghibi-Sistani, and Vahid Niazmand. Higher order spectra analysis of eeg signals in emotional stress states. In *Information Technology and Computer Science (ITCS), 2010 Second International Conference on*, pages 60–63. IEEE, 2010.

[35] Christopher J James and Christian W Hesse. Independent component analysis for biomedical signals. *Physiological measurement*, 26(1):R15, 2004.

[36] Herbert Jasper. The ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol*, 10:371–375, 1958.

[37] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 2014.

[38] Ashish Kapoor and Rosalind W Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682. ACM, 2005.

[39] Stamos Katsigiannis and Naeem Ramzan. Dreamer: a database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2018.

[40] Jasjeet Kaur and Amanpreet Kaur. A review on analysis of eeg signals. In *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, pages 957–960. IEEE, 2015.

[41] Loic Kessous, Ginevra Castellano, and George Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture

and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1-2):33–48, 2010.

[42] Stephanie Khalfa, Daniele Schon, Jean-Luc Anton, and Catherine Liégeois-Chauvel. Brain regions involved in the recognition of happiness and sadness in music. *Neuroreport*, 16(18):1981–1984, 2005.

[43] Hyun-Chul Kim and Zoubin Ghahramani. Bayesian gaussian process classification with the em-ep algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1948–1959, 2006.

[44] Jonghwa Kim. Bimodal emotion recognition using speech and physiological changes. In *Robust speech recognition and understanding*. InTech, 2007.

[45] Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134, 1992.

[46] Sander Koelstra and Ioannis Patras. Fusion of facial expressions and eeg for implicit affective tagging. *Image and Vision Computing*, 31(2):164–174, 2013.

[47] 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*, 3(1):18–31, 2012.

[48] Zbynek Koldovsky, Petr Tichavsky, and Erkki Oja. Efficient variant of algorithm fastica for independent component analysis attaining the cramér-rao lower bound. *IEEE Transactions on neural networks*, 17(5):1265–1277, 2006.

[49] Igor Kononenko, Marko Robnik-Sikonja, and Uros Pompe. Relieff for estimation and discretization of attributes in classification, regression, and ilp problems. *Artificial intelligence: methodology, systems, applications*, pages 31–40, 1996.

[50] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[51] Shashidhar G Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K Sreenivasa Rao. Iitkgp-sesc: Speech database for emotion analysis. In *International conference on contemporary computing*, pages 485–492. Springer, 2009.

[52] Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. Introducing the open affective standardized image set (oasis). *Behavior research methods*, 49(2): 457–470, 2017.

[53] Malte Kuss and Carl E Rasmussen. Assessing approximations for gaussian process classification. In *Advances in Neural Information Processing Systems*, pages 699–706, 2006.

[54] Nojun Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. *IEEE transactions on pattern analysis and machine intelligence*, 24(12):1667–1671, 2002.

[55] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, pages 39–58, 1997.

[56] Mu Li and Bao-Liang Lu. Emotion classification based on gamma-band eeg. In *2009 Annual International Conference of the IEEE Engineering in medicine and biology society*, pages 1223–1226. IEEE, 2009.

[57] Xiang Li, Peng Zhang, Dawei Song, Guangliang Yu, Yuexian Hou, and Bin Hu. Eeg based emotion identification using unsupervised deep feature learning. In *SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research*, 2015.

[58] Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, and Bin Hu. Exploring eeg features in cross-subject emotion recognition. *Frontiers in neuroscience*, 12:162, 2018.

[59] Yisi Liu and Olga Sourina. Real-time subject-dependent eeg-based emotion recognition algorithm. In *Transactions on Computational Science XXIII*, pages 199–223. Springer, 2014.

[60] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based human emotion recognition and visualization. In *Cyberworlds (CW), 2010 International Conference on*, pages 262–269. IEEE, 2010.

[61] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based emotion recognition and its applications. In *Transactions on computational science XII*, pages 256–277. Springer, 2011.

[62] Federica Lucivero and Guglielmo Tamburrini. Ethical monitoring of brain-machine interfaces. *Ai & Society*, 22(3):449–460, 2008.

[63] Anima Majumder, Laxmidhar Behera, and Venkatesh K Subramanian. Emotion recognition from geometric facial features using self-organizing map. *Pattern Recognition*, 47(3):1282–1293, 2014.

[64] Dante Mantini, Mauro G Perrucci, Cosimo Del Gratta, Gian L Romani, and Maurizio Corbetta. Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104(32): 13170–13175, 2007.

[65] Paul McCullagh, Gaye Lightbody, Jaroslaw Zygierewicz, and W George Kernohan. Ethical challenges associated with the development and deployment of brain computer interface technology. *Neuroethics*, 7(2):109–122, 2014.

[66] Christophe Meilhac and Chahab Nastar. Relevance feedback and category search in image databases. In *proceedings IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 512–517. IEEE, 1999.

[67] A Milton and S Tamil Selvi. Class-specific multiple classifiers scheme to recognize emotions from speech signals. *Computer Speech & Language*, 28(3): 727–742, 2014.

[68] Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[69] Murugappan Murugappan, Nagarajan Ramachandran, and Yaacob Sazali. Classification of human emotion from eeg using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, 3(04):390, 2010.

[70] Muthusamy Murugappan. Human emotion classification using wavelet transform and knn. In *2011 International Conference on Pattern Analysis and Intelligence Robotics*, volume 1, pages 148–153. IEEE, 2011.

[71] Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct): 2035–2078, 2008.

[72] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. Eeg-based emotion recognition during watching movies. In *2011 5th International IEEE/EMBS Conference on Neural Engineering*, pages 667–670. IEEE, 2011.

[73] Ernst Niedermeyer et al. The normal eeg of the waking adult. *Electroencephalography: Basic principles, clinical applications, and related fields*, 167: 155–164, 2005.

[74] Julie A Onton and Scott Makeig. High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in human neuroscience*, 3:61, 2009.

[75] Pallavi Pandey and KR Seeja. Subject-independent emotion detection from eeg signals using deep neural network. In *International Conference on Innovative Computing and Communications*, pages 41–46. Springer, 2019.

[76] Isidoros Perikos and Ioannis Hatzilygeroudis. Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, 51:191–201, 2016.

[77] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):186–197, 2010.

[78] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis. A novel emotion elicitation index using frontal brain asymmetry for enhanced eeg-based emotion recognition. *IEEE Transactions on information technology in biomedicine*, 15 (5):737–746, 2011.

[79] Rosalind Wright Picard. Affective computing. Technical report, Massachusetts Institute of Technology, 1995.

[80] Laura Piho and Tardi Tjahjadi. A mutual information based adaptive windowing of informative eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 2018.

[81] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian process for machine learning*. MIT press, 2006.

[82] Guillermo Recio, Annekathrin Schacht, and Werner Sommer. Recognizing dynamic facial expressions of emotion: Specificity and intensity effects in event-related brain potentials. *Biological psychology*, 96:111–125, 2014.

[83] G Repovš. Dealing with noise in eeg recording and data analysis. *Informatica Medica Slovenica*, 15(1):18–25, 2010.

[84] Carson Reynolds and Rosalind Picard. Affective sensors, privacy, and ethical contracts. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, pages 1103–1106. ACM, 2004.

[85] Giorgio Roffo, Simone Melzi, and Marco Cristani. Infinite feature selection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4202–4210, 2015.

[86] Viktor Rozgić, Shiv N Vitaladevuni, and Rohit Prasad. Robust eeg emotion classification using segment level decision fusion. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 1286–1290. IEEE, 2013.

[87] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980. ISSN 0022-3514.

[88] Vangelis Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with eeg/meg. *Computers in biology and medicine*, 41 (12):1110–1117, 2011.

[89] Saeid Sanei and Jonathon A Chambers. Eeg signal processing. 2007.

[90] Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 383–392. IEEE, 2014.

[91] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[92] Dahlia Sharon, Matti S Hämäläinen, Roger BH Tootell, Eric Halgren, and John W Belliveau. The advantage of combining meg and eeg: comparison to fmri in focally stimulated visual cortex. *Neuroimage*, 36(4):1225–1235, 2007.

[93] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2012.

[94] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[95] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 2018.

[96] Steffen Steinert and Orsolya Friedrich. Wired emotions: Ethical issues of affective brain–computer interfaces. *Science and engineering ethics*, pages 1–17, 2019.

[97] Petre Stoica, Randolph L Moses, et al. *Spectral analysis of signals*, volume 1. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[98] Kazuhiko Takahashi et al. Remarks on emotion recognition from bio-potential signals. In *2nd International conference on Autonomous Robots and Agents*, volume 3, pages 1148–1153, 2004.

[99] Desney Tan and Anton Nijholt. Brain-computer interfaces and human-computer interaction. In *Brain-Computer Interfaces*, pages 3–19. Springer, 2010.

[100] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[101] Petr Tichavsky, Arie Yeredor, GermÁn Gomez-Herrero, Eran Doron, et al. A hybrid technique for blind separation of non-gaussian and time-correlated sources using a multicomponent approach. *IEEE Transactions on Neural Networks*, 19(3):421–430, 2008.

[102] Petr Tichavsky, Arie Yeredor, and Jan Nielsen. A fast approximate joint diagonalization algorithm using a criterion with a block diagonal weight matrix. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3321–3324. IEEE, 2008.

[103] Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of machine learning research*, 3(Mar):1415–1438, 2003.

[104] Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshu Mittal, and Samit Bhattacharya. Using deep and convolutional neural networks for accurate emotion classification on deap dataset. In *AAAI*, pages 4746–4752, 2017.

[105] Vladimir Vapnik. *Statistical learning theory. 1998*. Wiley, New York, 1998.

[106] Ashwini Ann Varghese, Jacob P Cherian, and Jubilant J Kizhakkethottam. Overview on emotion recognition system. In *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, pages 1–5. IEEE, 2015.

[107] Rutger J Vlek, David Steines, Dyana Szibbo, Andrea Kübler, Mary-Jane Schneider, Pim Haselager, and Femke Nijboer. Ethical issues in brain–computer interface research, development, and dissemination. *Journal of neurologic physical therapy*, 36(2):94–99, 2012.

[108] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[109] Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

[110] Haiyan Xu and Konstantinos N Plataniotis. Subject independent affective states classification using eeg signals. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1312–1316. IEEE, 2015.

[111] Arie Yeredor. Blind separation of gaussian sources via second-order statistics with asymptotically optimal weighting. *IEEE Signal Processing Letters*, 7(7): 197–200, 2000.

[112] Rafael Yuste, Sara Goering, Guoqiang Bi, Jose M Carmena, Adrian Carter, Joseph J Fins, Phoebe Friesen, Jack Gallant, Jane E Huggins, Judy Illes, et al. Four ethical priorities for neurotechnologies and ai. *Nature News*, 551(7679): 159, 2017.

[113] Angela Zeiler, Rupert Faltermeier, Ingo R Keck, Ana Maria Tomé, Carlos García Puntonet, and Elmar Wolfgang Lang. Empirical mode decomposition-an introduction. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

[114] Aihua Zhang, Bin Yang, and Ling Huang. Feature extraction of eeg signals using power spectral entropy. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, volume 2, pages 435–439. IEEE, 2008.

[115] Jianhai Zhang, Ming Chen, Shaokai Zhao, Sanqing Hu, Zhiguo Shi, and Yu Cao. Relieff-based eeg sensor selection methods for emotion recognition. *Sensors*, 16 (10):1558, 2016.

[116] Wei-Long Zheng and Bao-Liang Lu. Personalizing eeg-based affective models with transfer learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2732–2738. AAAI Press, 2016.

[117] Yachen Zhu, Shangfei Wang, and Qiang Ji. Emotion recognition from users' eeg signals with the help of stimulus videos. In *2014 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2014.