A Machine Learning Approach to EEG-Based Emotion Recognition

A MACHINE LEARNING APPROACH TO EEG-BASED EMOTION RECOGNITION

BY

SARA JAMIL, B.Eng.

A THESIS

SUBMITTED TO THE SCHOOL OF COMPUTATIONAL SCIENCE & ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2018) McMaster University

(Computational Science & Engineering) Hamilton, Ontario, Canada

TITLE:	A Machine Learning Approach to EEG-Based Emotion Recognition
AUTHOR:	Sara Jamil B.Eng., (Electrical & Biomedical Engineering) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Ranil Sonnadara
CO-SUPERVISOR:	Dr. Suzanna Becker
NUMBER OF PAGES:	xiii, 95

To my parents

Abstract

In recent years, emotion classification using electroencephalography (EEG) has attracted much attention with the rapid development of machine learning techniques and various applications of brain-computer interfacing. In this study, a general model for emotion recognition was created using a large dataset of 116 participants' EEG responses to happy and fearful videos. We compared discrete and dimensional emotion models, assessed various popular feature extraction methods, evaluated the efficacy of feature selection algorithms, and examined the performance of 2 classification algorithms. An average test accuracy of 76% was obtained using higher-order spectral features with a support vector machine for discrete emotion classification. An accuracy of up 79% was achieved on the subset of classifiable participants. Finally, the stability of EEG patterns in emotion recognition was examined over time by evaluating consistency across sessions.

Acknowledgements

I would like to acknowledge my supervisors, Dr. Ranil Sonnadara and Dr. Sue Becker, for their contributions and guidance. Their support has helped immensely through the research process and allowed me to overcome many obstacles along the way. I would also like acknowledge my supervisory committee members, Dr. Jim Reilly and Dr. Dan Bosnyak, for their valuable advice and support.

I would like to thank Dr. Kiret Dhindsa and Saurabh Shaw for their invaluable guidance. I would also like to acknowledge the members of the Performance Science Lab, Neurotechnology and Neuroplasticity Lab, and the LIVELab for the enlightening discussions. Furthermore, I would like to thank the SHARCNET team at McMaster for their help with high performance computing.

Finally, I would like to express my gratitude to my friends and family for providing me with continuous encouragement throughout my years of study. I would especially like to thank my parents for all their love and support. This accomplishment would not have been possible without them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Emotion plays a significant role in human interaction and behaviour, and is often associated with mood, temperament, and motivation. Despite its key role in our daily lives, our scientific knowledge on the mechanisms of emotional function in the brain is limited. Research on emotion recognition from electrophysiological responses has attracted great interest from a vast number of interdisciplinary fields ranging from psychology to engineering. In this thesis, we applied state-of-the-art machine learning techniques to create a general model for emotion recognition using electroencephalography (EEG).

The study of emotions has baffled scientists for years and has been the center of much controversy. Some scientists claim it is impossible to empirically study our subjective experience of emotional states since we cannot directly measure the internal experiences of other humans or even animals (Panksepp, 1998). Research in psychology was dominated by behaviourists for some time, who believed that subjective inner states of mind should be treated as a black box, but with the rise of cognitive science EEG became a useful tool to examine the neural pathways inside the

black box to understand such things as perception, memory, and attention (LeDoux, 1996). Whether emotion also lies within the realm of cognition has been contested throughout the years, and the study of affective states has largely been neglected by scientists until the end of the twentieth century (LeDoux, 2000) because emotions have been found to be notoriously difficult to study.

Although there are no direct metrics that can quantify emotional states without any ambiguity, there are many indirect measures that can be used to infer such states. These measures can be derived from responses ranging from facial expressions to autonomic nervous system responses, and more recently, to brain responses using neuroimaging techniques like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). EEG was first developed in the early twentieth century, but it wasn't until recent decades that developments in technology and artificial intelligence (Michalski *et al.*, 1983) made EEG a viable option for regular clinical use, and research using EEG in affective computing started to thrive.

Affective computing is defined as "computing that relates to, arises from, or deliberately influences emotions" (Van Den Broek, 2012) and encompasses a wide range of applications in brain-computer interfaces (BCI). A BCI is a system that measures activity in the central nervous system (CNS) and converts it into an artificial output that can enhance, replace, or supplement natural CNS output (Wolpaw, 2013). Affective brain-computer interfaces (aBCI) are the extension of the field of BCI that attempts to create devices able to detect affective states from neurophysiological signals (Mühl *et al.*, 2014). One such application is a prosthetic device that tries to detect the emotional state of a person with a communication disorder (Cowie *et al.*, 2001). Another application is to support a user's ability to modify their mental state

through neurofeedback in therapy for mood disorders (Hammond, 2005). It is also used for entertainment purposes like interactive games that use a player's detected emotion to manipulate the gameplay (Bos *et al.*, 2010). In general, aBCI can be used for active communication of emotional states and for passive sensing to inform the machine on the affective state of its user. Whether EEG is used to study the cognitive science of emotion or for aBCI applications, the study of affective computing is an exciting research domain for many scientists and engineers.

Chapter 2

Models of Emotion

While the study of affect is sometimes considered a part of cognitive science as they both contain conscious and unconscious processes, there are differences between cognition and affect. Cognition generally refers to less subjective phenomena that do not necessarily involve emotional feeling (e.g., attention, perception, and language), however cognitive and affective processes are tightly intertwined (Barrett *et al.*, 2007). Affect is a term that is used to describe emotion and mood, where emotions refer to intense short-term states while moods refer to less intense long-term states that might not be associated with one particular event but several that accumulate over time. The importance of creating a precise definition of the complex phenomenon of emotion is now recognized, and one widely used working definition is that "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labelling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behaviour that

is often, but not always, expressive, goal-directed, and adaptive." (Kleinginna and Kleinginna, 1981) There are a few approaches to model affective responses, but the two main approaches that are used in emotion recognition applications are discrete and dimensional emotion models.

2.1 Discrete Emotion Models

Discrete emotion models cluster emotional responses by a small number of universal discrete emotions. They are described as universal because they can be found cross culturally as well as some are shared with primates and other mammals, thus are evolutionarily developed responses (Panksepp, 1998). The specific set of basic emotions is still in dispute. For a long time the predominant view was that there are six: happiness, sadness, fear, anger, surprise and disgust (Ekman, 1992). Still, there have been many other basic discrete emotions that have been proposed (see e.g. (Plutchik, 1980)) and there is a general lack of consensus on the number and type of basic emotions. These discrete models of emotion are quite often used in emotion recognition applications where the machine tries to classify between two or more emotional states, however, they do not seem to adequately reflect the complexity of emotions.

2.2 Dimensional Emotion Models

Dimensional emotion models attempt to address the complexity of emotions by suggesting that there are several dimensions that can define an emotional state. One of the most commonly used dimensional models is Russel's circumplex model of emotion which assumes that any emotional feeling can be mapped in terms of its valence

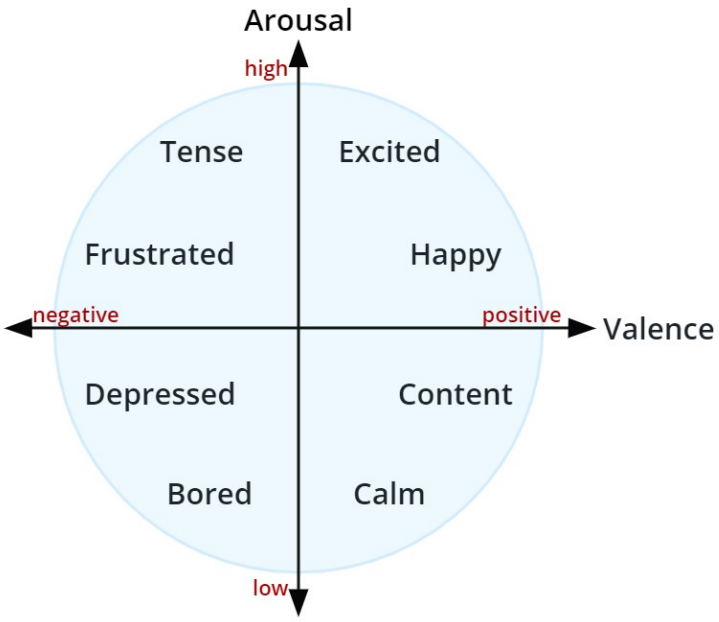

Figure 2.1: The 2-dimensional valence-arousal emotion model.

and arousal (Russell, 1980). Valence codes emotional states as positive or negative while arousal is a measure of the level of intensity of the emotion (see Figure 2.1). Although more complex emotions can be represented using a dimensional emotion model compared to a discrete emotion model, there is also debate on the number of dimensions that exist with other proposed dimensions like unpredictability and dominance (Russell and Barrett, 1999). Dimensional emotion models are used in emotion recognition and can be used in regression or classification problems.

These emotion models are closely related to the faculty approach to the neurophysiological basis of affect. The faculty approach proposes that different mental

faculties can be isolated from one another, whether it be cognitive or affective faculties. Discrete emotions like happiness or fear would be mapped to specific physiological responses or brain circuits according to this theory. However, as Hamann (2012) states, "Although neuroimaging studies have identified consistent neural correlates as associated with basic emotions or other emotion models, they have ruled out simple one-to-one mappings between emotions and brain regions, pointing to the need for more complex, network-based representations of emotion." Nevertheless, this does not disqualify these models from being adequate tools in differentiating between different affective states in various aBCI applications. Both emotion models are examined in this study, but the main focus will be on discrete models due to their relatively higher performance.

Chapter 3

Neurophysiology of Affect in EEG

In addition to categorizing emotion through models, we need an objective measure of emotional reaction. EEG is an electrophysiological recording method that measures the electrical activity via non-invasive electrodes placed at the scalp that measure voltage fluctuations resulting from the electric potentials of groups of neurons in the brain. The brain structures that are involved in emotions include parts of the orbital and medial prefrontal cortex, ventral parts of the basal ganglia, the mediodorsal nucleus of the thalamus, and the amygdala (Halterman, 2005) (see Figure 3.1). There is some debate on the efficacy of EEG measurement for deep brain structures, some of which are involved in emotional responses (Cohen, 2014). However, for the sake of EEG emotion recognition, we are mainly interested in accessible affect-related neural activity and whether it can be useful in classifying between different emotions via temporal or frequency based neural correlates. While time-domain correlates can be derived from the microvolt amplitude of the EEG signals at the instant during which they were sampled, frequency-domain correlates are derived from the frequency-domain representation in which the waveform is represented by amplitude and phase

values of each frequency component. Both of these representations can be useful in extracting features from the signal relevant to the problem of emotion recognition.

3.1 Time-Domain Correlates

Due to the stochastic and highly noisy nature of EEG signals, and the fact that the scalp potential represents the superposition of brain activity in many different regions, reflecting multiple sources of cognitive activity, a frequently used experimental protocol is the event-driven paradigm, where the same stimulus or class of stimuli is repeated many times, and the time domain correlates are analyzed by taking the average of these multiple responses to a particular stimulus, such as an image or a sound. These event-related potentials (ERPs) are stereotyped deflections of the electrophysiological response which range in duration from tens to hundreds of milliseconds post stimulus. Both early and late onset potentials are modulated by affective stimuli, where early potentials are attenuated for low-valence images and late potentials are enhanced for high-arousal images (Olofsson *et al.*, 2008). The averaged response has the advantage of reduced noise, as fluctuations in the signal that are not correlated to the stimulus are attenuated while maintaining a high temporal resolution, but requires many trials to be measured accurately because of this decrease in amplitude (Luck, 2014). However, for responses that are not time-locked to a stimulus, such as a continuous stimulus (e.g. a video), this event-averaging is not applicable. In this case, most (but not all) measures are calculated in the frequency domain.

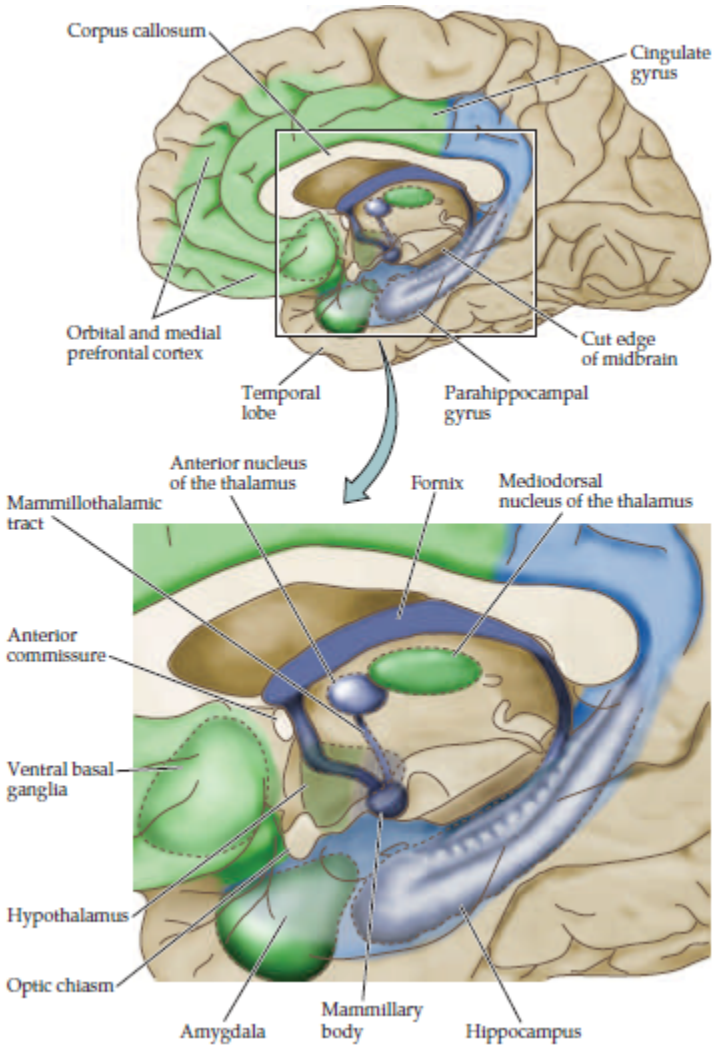

Figure 3.1: The main structures of the brain that are key in the experience and expression of emotion include the orbital and medial prefrontal cortex, amygdala, hypothalamus, and ventral striatum (coloured green).The mammilary bodies of the hypothalamus and the hippocampus (coloured blue), are parts of the limbic system that are no longer considered important to the processing of emotion. Reprinted from *Neuroscience, 4th Edition* (p. 741), by D. Purves, 2008, Massachusetts, USA: Sinauer Associates, Inc. Copyright [2008].

3.2 Frequency-Domain Correlates

Frequency domain correlates are found using different power extraction methods that measure the power of the signal at each frequency of interest. The conventionally used broad frequency bands (i.e. delta, theta, alpha, beta, and gamma) have been observed in association with affect and common patterns have emerged. Increases in delta band (0.5–4 Hz) power have been associated with high-arousal stimuli (Knyazev, 2012) and motivational states (eg. hunger, sexual arousal) (Balconi and Lucchiari, 2006). Theta band (4–8 Hz) power has also been associated with emotionally arousing stimuli in the frontal and parietal lobe (Aftanas *et al.*, 2002) as well as positive valence stimuli (Başar *et al.*, 2001). The alpha band (8–13 Hz) has one of the most prominent associations to affective states, and in particular, a widely used measure is frontal alpha asymmetry (FAA), which shows stronger rightward lateralization for positive emotions than negative emotions (see e.g. Harmon-Jones (2003). Beta band (13–30 Hz) and gamma band (>30 Hz) power have both been associated with positive valence stimuli (Onton, 2009); however, highly arousing stimuli have been shown to decrease beta power (Glauser and Scherer, 2008) while increasing gamma power (Balconi and Pozzoli, 2009). There is some ambiguity on whether these higher frequency bands are actually measuring EEG or if the signal is mostly corrupted by motion artifact (Goncharova *et al.*, 2003). Frequency-based analysis techniques have been proven to be powerful tools for emotion recognition and classification . While some of these time and frequency domain correlates have clinical significance, machine learning gives us the power to discover the most important features for emotion recognition.

Chapter 4

Process of EEG Classification

The process of EEG emotion recognition typically consists of several components that perform critical functions in the process cycle as shown in Figure 4.1. EEG classification can be performed online with closed-loop feedback, as in BCI studies, or it can be performed offline to study affective response in EEG. Table 4.1 includes a list of publications on such studies to provide an overview of EEG emotion recognition research. Mühl *et al.* (2014) provides an in-depth review on state-of-the-art affective brain-computer interface research. This section will review the relevant literature to highlight some of the most important aspects of EEG emotion recognition including stimuli and experimental setup, feature extraction, and classification algorithm.

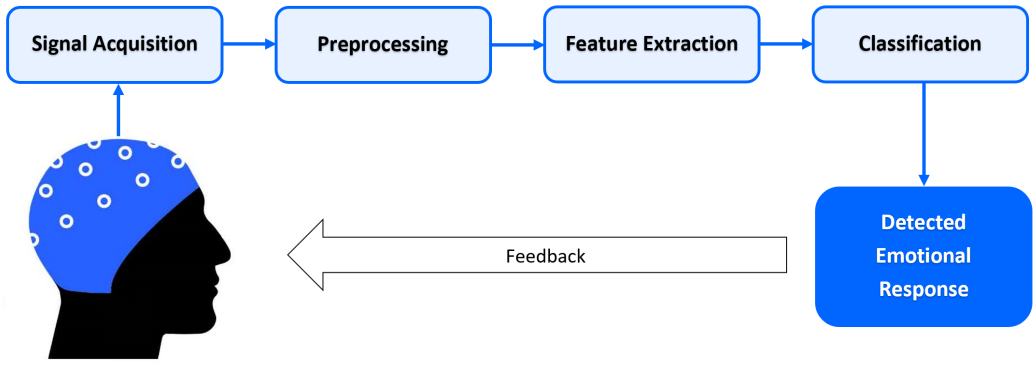

Figure 4.1: The main components of EEG emotion recognition. Real-time BCI applications may include some form of feedback to the user, however, this is not required in offline analysis of emotional responses.

Table 4.1: Literature review table

Reference	Emotion elicitation	Stimulus duration	Assessed emotions	Number of participants	Type of model	Number of channels	Noise preprocessing	Features extracted	Classification or regression algorithm	Performance
Kumar *et al.* (2016)	Videos	60 s	valence and arousal	32	User-independent	2	bandpass filter	Higher Order Spectra (HOS)	SVM and ANN	64.84% (arousal), 61.17% (valence)
Atkinson and Campos (2016)	Videos	60 s	valence and arousal	32	User-independent	14	bandpass filter, ocular artifacts removed, and down-sampling	Statistical features, Spectral features, Hjorth parameters (HP), and fractal dimension (FD)	SVM with feature selection	73.06% (2-way arousal), 73.14% (2-way valence), 60.7% (3-way arousal), 62.33% (3-way valence)
Jirayucharoensak *et al.* (2014)	Videos	60 s	valence	32	User-independent	32	bandpass filter	Spectral features	Deep learning network (DLN) using stacked autoecoder (SAE), principal component analysis (PCA)	53.43% (3-way valence), 52.05% (3-way arousal)
Soleymani *et al.* (2012)	Videos	35 - 117 s	valence and arousal	27	User-independent	14	band-pass filter and partici-pants were asked not to move	Spectral features and hemispheric asymmetry	SVM with feature selection	52.4% (3-way arousal), 57.0% (3-way valence)
Murugappan *et al.* (2010)	Videos	N/A	disgust, happiness, surprise, fear, and neutral	20	User-independent	64	bandpass filter, par-ticipants were asked not to move, and surface laplacian filtering	Wavelet transform using db4 and Absolute Logarithmic Recoursing Energy Efficiency (ALREE)	Linear discriminant analysis (LDA) and K-Nearest Neighbour (KNN)	83.26% (KNN), 75.21% (LDA)
Takahashi (2004)	Videos	N/A	joy, anger, sadness, fear, and relaxation	12	User-independent	3	notch filter	Statistics	SVM	42%

Table 4.1 continued from previous page

Reference	Emotion elicitation	Stimulus duration	Assessed emotions	Number of participants	Type of model	Number of channels	Noise preprocessing	Features extracted	Classification or regression algorithm	Performance
Nie *et al.* (2011)	Videos	∼4 min	valence	6	User-independent	62	bandpass filter and manual inspection	Spectral features and feature smoothing	SVM with feature selction	87.53% (all features), 89.22% (100 features), 84.94% (50 features)
Murugappan *et al.* (2009)	Videos	100 s	disguist, happiness, surprise, sadness, and anger	5	User-independent	64	bandpass filter and subtracting mean from channels	Wavelet transform using db4	ANN	56.66% (images), 66.67% (videos)
Frantzidis *et al.* (2010)	Images	2.5 s	valence and arousal	28	User-independent	3	bandpass filter and ocular artifacts removed using adaptive filter	Averaged ERP features and Wavelet transform	Mahalanobis Distance (MD), and SVM	79.5% (MD), 81.3% (SVM)
Petrantonakis (2010)	Images	5 s	happiness, surprise, anger, fear, disgust, and sadness	16	User-independent	3	bandpass filter and segments were averaged for each emotion	Higher Order Crossings (HOC)	Quadratic Discriminant Analysis (QDA), KNN, Mahalanobis Distance (MD), and SVM	83.33% (SVM)
Hosseini and Naghibi-Sistani (2011)	Images	12 s	calm-neutral and negatively excited	15	User-independent	5	bandpass filter	Fractal dimension (FD), Correlation dimension (D2), Approximante entropy (ApEn), Wavelet transform using db4, and Higher order spectra (HOS)	SVM and Elman neural network (ENN) with feature selection	84.6% (SVM), 83.1% (ENN)

Table 4.1 continued from previous page

Reference	Emotion elicitation	Stimulus duration	Assessed emotions	Number of participants	Type of model	Number of channels	Noise preprocessing	Features extracted	Classification or regression algorithm	Performance
Choppin (2000)	Images	6 - 10 s	happy, neutral, and unhappy	7	User-independent	19	ocular artifacts removed and muscle artifacts rejected	Spectral features and coherence	ANN	64%
Dhindsa and Becker (2017)	Videos	∼3.5 min	interest, amusement, happiness, sadness, fear, disgust, anger, hope, relief, surprise, and sympathy	40	User-dependent and user-independent	4	bandpass filtering using Filter-Bank Artifact Rejection	Spectral features, coherence, cross-frequency coupling, higher order spectra (HOS), and hemispheric asymmetry	SVM and LR with feature selection	75% (SVM), 70% (LR)
Zheng *et al.* (2016)	Videos	∼4 min	valence	15	User-dependent and user-independent	62	bandpass filter	Spectral features, differential entropy (DE), differential asymmetry (DASM), rational asymmetry (RASM), asymmetry (ASM), differential caudality (DCAU), and feature smoothing	KNN, LR, SVM, and GELM with feature selection	70.43% (KNN), 84.08% (LR), 78.21% (SVM - RBF), 83.26% (SVM - linear), 91.07% (GELM)
Hidalgo-Muñoz *et al.* (2013)	Images	3.5 s	valence	26	User-dependent and user-independent	21	artifact rejection and eye-movement correction	Spectral turbulance	SVM with feature selection	67% (user-dependent), 82% (user independent)

Table 4.1 continued from previous page

Reference	Emotion elicitation	Stimulus duration	Assessed emotions	Number of participants	Type of model	Number of channels	Noise preprocessing	Features extracted	Classification or regression algorithm	Performance
Petrantonakis and Hadjileontiadis (2011)	Images	5 s	valence and arousal	16	User-dependent and user-independent	3	band-pass filter	Hemispheric asymmetry	SVM	up to 62.58% (user-independent) and 94.40% (user-dependent)
Xu and Plataniotis (2016)	Videos	60 s	valence, arousal, and liking	32	User-dependent	32	bandpass filter	Narrow-band power spectral density features	SVM, stacked denoising autoencoder (SDAE), and deep belief networks (DBN)	86.67% (arousal), 86.60% (valence), 86.69% (liking)
Koelstra *et al.* (2012)	Videos	35 - 117 s	valence, arousal, and dominance	32	User-dependent	32	bandpass filter	Spectral features and hemispheric asymmetry	Gaussian Nave Bayes Bayesian ridge regression	66 - 71%
Zheng *et al.* (2014)	Videos	∼4 min	valence	6	User-dependent	62	band-pass filter and participants were asked not to move	Differential entropy (DE)	Deep belief network (DBN), Hidden markov model (HMM), Graph regularized Extreme Learning Machine (GELM), SVM, KNN	87.62% (DBN-HMM), 86.91% (DBN), 85.67% GELM, 84.08% (SVM), 69.66% (KNN)
Wang *et al.* (2011)	Videos	4 - 5 min	joy, relaxation, sadness, and fear	5	User-dependent	62	ocular artifacts removed	Statistics and spectral features	KNN, ANN, and SVM with feature selection	66.51% (4-way SVM)
Chanel *et al.* (2006)	Images	6 s	arousal	4	User-dependent	34	bandpass filter	Spectral features	Nave Bayes	60% (2-way), 46% (3-way)
Lin *et al.* (2010)	Sounds	30 s	joy, anger, sadness, and pleasure	26	User-dependent	32	bandpass filter and visual inspection	Spectral features and hemispheric asymmetry	SVM	82%

Table 4.1 continued from previous page

Reference	Emotion elicitation	Stimulus duration	Assessed emotions	Number of participants	Type of model	Number of channels	Noise preprocessing	Features extracted	Classification or regression algorithm	Performance
Lan *et al.* (2016)	Sounds	60 s	pleasant, happy, frightened, and angry	5	User-dependent	14	bandpass filter	Statistical features, spectral features, higher order crossings (HOC), fractal dimension (FD), intra-class correlation coefficient	SVM	71.75% (2-way), 49.63% (4-way), 73.10% (positive vs. negative emotion)
Duan *et al.* (2012)	Sounds	3 min	arousal	5	User-dependent	62	downsampling and manual artifact rejection	Spectral features, hemispheric asymmetry, and feature smoothing	SVM, KNN, and least-squares with feature selection	81.03%

4.1 Stimuli and Experimental Setup

4.1.1 Stimuli

One of the most important aspects of emotion recognition is determining what stimuli should be used to elicit the assessed emotions. Although some studies rely on mental imagery to elicit emotions (Chanel *et al.*, 2009), most studies use stimuli like images, music, and videos to passively elicit the desired emotions. Images are the most frequently used type of stimulus due to the availability of the International Affective Picture System (IAPS). This widely used dataset contains a large set of pictures with their own ratings for valence, arousal, and dominance which were averaged from a sample of 100 participants (J Lang *et al.*, 2008). This allows for easy a priori selection of images that can be used as a ground-truth for emotion assessment. The International Affective Digitized Sounds (IADS) (Bradley and Lang, 1999) is a database of music used for emotion elicitation, however, auditory stimuli are less commonly used despite being shown to be equally effective to visual stimulation in eliciting emotion (Zhou *et al.*, 2014). Videos taken from music or movie clips are the second most commonly used stimuli. Film clips have been shown to be very effective in inducing strong emotions (Westermann *et al.*, 1996). These methods have all been used in the literature to induce the assessed emotions.

4.1.2 Assessed Emotion

Another significant component of emotion recognition studies is to determine which emotions are to be assessed. Most of the studies in Table 1 used the valence-arousal space to define the emotional classes. This is usually done by splitting the 2D

space into several components or assessing high vs low valence or arousal separately, while few papers performed regression to directly predict the valence and arousal rating (Koelstra *et al.*, 2012). Few studies tried to analyze emotions on other dimensions like dominance (Reuderink *et al.*, 2013) despite studies like Fontaine *et al.* (2007) that illustrated that at least 4 dimensions are necessary to distinguish emotional language. However, the use of only two dimensions remains a popular method for reducing complexity of emotion analysis. Discrete emotions are also widely used in EEG emotion recognition which can consist of any number of different proposed basic emotions (i.e. fear, disgust, sadness, happiness, surprise, amusement, interest, etc.). Basic emotions seem to perform well in aBCI studies (Petrantonakis, 2010) suggesting that this emotion model can be adequately useful for various applications despite the lack of consensus on the theory of basic emotions.

4.1.3 Number of Participants

The number of participants in a study is another variable that can have an impact on the interpretation of the results especially with regards to participant-independent models. Participant independent classifiers try to create a general model of emotional response while participant dependent classifiers create a separate model for each individual. Participant independent models have the major advantages of not requiring a training phase for each user and requiring fewer trials conducted per participant. However, they also tend to perform worse compared to participant dependent models. The higher the number of participants, the more generalizable the results can be considered for both types of classifiers. The studies in Table 1 include 5 to 32 participants and some of the studies perform classification on the available

EEG emotion databases like the Database for Emotion Analysis using Physiological Signals (DEAP) (Koelstra *et al.*, 2012) or the MAHNOB-HCI database (Soleymani *et al.*, 2012). In our study, EEG analysis is conducted on 116 participants, which is the largest sample size of all the observed studies.

4.1.4 Number of Channels

The number and placement of EEG channels recorded is another important aspect of the experimental setup. Table 1 shows that the number of channels used in EEG emotion recognition varies between 3 to 64 electrodes. Having fewer electrodes produces less discomfort for the user while also being more affordable and overall reducing the complexity of the system. Several studies performed multimodal recordings in addition to EEG including electrocardiography (ECG), electromyography (EMG), electrooculography (EOG), facial expressions, eye gaze, pupillary response, galvanic skin response (GSR), speech, and more (Soleymani *et al.*, 2015). The advantage of EEG compared to other modalities is that emotional brain activity can be recognized a few milliseconds post stimuli (Grandjean and Scherer, 2008) and has superior temporal resolution compared to other cheap and non-invasive neuroimaging techniques like functional Near Infrared Spectroscopy (fNIRS). Although the use of additional modalities can improve classification accuracy, this study will focus mainly on emotion response of EEG alone using only 5 electrodes that could lead to a commercial or clinical application.

4.1.5 Noise Preprocessing

The final step that is usually taken prior to extracting relevant features is to perform preprocessing to reduce the amount of noise and artifacts. EEG is known to have low signal to noise ratio (SNR) meaning that the signal may contain a mixture of a large number of sources and may be corrupted by motion artifacts as well as noise from external sources. Artifacts may arise from muscle movement, such as eye or head movement or even heart rate, which can have higher amplitude than neural activity. Most of the studies examined in this review only performed basic filtering to remove drifts and power-line noise using a bandpass or notch filter. Sometimes it is noted that certain participants whose recordings are particularly noisy are removed from the study. Only a few studies performed artifact rejection such as removing eye blinks using Independent Component Analysis (ICA) or removing noisy portions of the signal. This may be due to the additional challenges introduced or assumptions that accompany these corrections. Although it is important to distinguish between the different sources of information that can contaminate the signal, some of these artifacts contain relevant emotional information and may be valuable in improving performance accuracy.

4.2 Feature Extraction

Feature extraction can be considered one of the most important steps in emotion classification. Computed features that exhibit the highest discernment between the assessed emotions can vastly improve the efficiency and performance of the pattern recognition stage. The most commonly used features are computed from the power

spectrum of the recorded EEG channels. Spectral analysis tends to focus on the power spectrum at several frequency bands using the discrete Fourier transform. However, some studies perform discrete Wavelet transforms using the Daubechies (db4) wavelet instead, which is chosen due to it's resemblance to the sum of neuronal action potentials (Glassman and Member, 2005). Another frequently-used class of features that is derived through spectral analysis is hemispheric asymmetry indexes that calculate the ratio of (or difference between) energies in a particular frequency band in the right and left lobes of the brain, such as the aforementioned frontal alpha asymmetry (FAA).

Although these analyses that utilize frequency domain correlates are most commonly used in feature extraction, several other methods exist to compute features relevant to emotional states. Higher-order spectral analysis is emerging as an effective feature and has shown to be successful for emotion recognition (Hosseini and Naghibi-Sistani, 2011). As the brain is a non-linear dynamic system, using non-linear higher-order spectral analysis techniques may be beneficial. Bispectral analysis is capable of tracking changes in signals arising from non-linear processes, and unlike power spectral analysis, it does not ignore potential interactions between components of the signal that are manifested as phase coupling (Sigl and Chamoun, 1994). Other features examined in emotion recognition studies include differential entropy, fractal dimension, and correlation dimension of the EEG signals (Zheng *et al.*, 2014) (Atkinson and Campos, 2016) (Lan *et al.*, 2016). Any combination of these features can be used as the feature space in classification of emotion.

With the ever-expanding possibilities for relevant features, it is easy to generate a large feature set that can increase the complexity and computational cost of the

system. Therefore, some studies opt to use feature selection methods in addition to feature extraction. Feature selection is the process of selecting a subset of features to optimize performance and reduce the dimensionality of the feature space (Hua *et al.*, 2005). One such method is the minimum Redundancy Maximum Relevance (mRMR) algorithm which uses mutual information to select a compact set of superior features at a low computational cost (Peng *et al.*, 2005). Other feature selection methods have been used such as ranking features with a one-way ANOVA (Xu and Plataniotis, 2016). Once an optimal feature set that can sufficiently discriminate between the assessed emotional states is chosen, a classification algorithm is used to create a model for emotion recognition.

4.3 Classification Algorithm

There are many pattern recognition algorithms that have been used in emotion recognition. The support vector machine (SVM) is among the most widely used machine learning algorithms. Support vector classification can separate classes in a high dimensional feature space using linear, polynomial, or Gaussian kernels. Although other classification algorithms have been used, like Linear Discriminant Analysis (LDA), logistic regression (LR), and Naïve Bayes classifiers, SVM is a popular classification method as it has been shown to perform well for EEG emotion recognition using non-linear kernels (Hosseini and Naghibi-Sistani, 2011).

In recent years, artificial neural networks (ANN) have been emerging as a powerful tool for many machine learning applications that deal with non-linear and complex systems. It has been applied to EEG analysis including a few emotion recognition studies that have employed supervised and unsupervised or semi-supervised learning

techniques (Xu and Plataniotis, 2016). The convolutional neural network (CNN) is an artificial neural network that can learn local patterns in data using convolution of a kernel/filter (LeCun *et al.*, 2015). CNNs have been implemented for the decoding of movement-related information from EEG (Schirrmeister *et al.*, 2017). In this study, the proposed architectures included a shallow and a deep CNN that take the raw EEG as input, alleviating the need for feature extraction. CNNs have also been used for P300 detection in non-affect related BCIs (Cecotti and Gräser, 2011). There is a vast number of different possible architectures and variable parameters involved in the use of artificial neural networks, which leaves much room for exploration in their use in affective BCIs and emotion recognition.

Due to the many available machine learning algorithms and the many parameters involved in each type of classifier, comparison between classification methods is difficult. In addition, the differences in experimental protocol and the large variety of possible features make comparisons of performance between studies highly problematic. Consequently, some studies opt to compare several pattern recognition methods to identify the optimal classifier (Zheng *et al.*, 2014). Although performance may not be directly comparable, these studies can provide a lot of insight into what may work best to meet the challenges of other EEG-based emotion recognition applications.

Chapter 5

Summary of previous studies

The ever-expanding body of work on EEG-based emotion recognition presents a large variety of possible methods that can be employed to collect and analyze the EEG data. Table 1 provides a non-exhaustive list of 24 emotion recognition studies to highlight the breadth of the methods used in the literature. Of these 24 emotion recognition studies, 14 used video stimuli for emotion elicitation, 16 performed user-independent analysis, and 10 used both video stimuli and user-independent modelling. Of the 10 studies, the number of participants ranged from 5 to 40 participants. This section attempts to highlight the studies that have presented an encouraging approach to EEG-based emotion recognition using similar datasets in order to compare the methods and the performance achieved in previous studies.

In order to compare methods used to our study, a brief summary of the methods are outlined here. The dataset examined here was collected in a previous study which assessed the affective state of participants as they watched 2–3 minute movie clips chosen to elicit fearful or happy responses. It includes 5-channel EEG recordings of over 100 participants over two separate sessions. The two basic emotions assessed

were fear and happiness as well as valence and arousal models. Some of the features extracted were spectral features, hemispheric asymmetry features, and higher-order spectral features. Two classification algorithms were also assessed, SVM and CNN, and feature selection was also tested.

The study that was found to be most similar to our study was conducted by Dhindsa and Becker (2017) which performed participant-independent modelling on 40 participants and was the largest dataset used from the literature review. Videos were also used to elicit emotions and data from only 4 electrodes were recorded using the Muse headband. The data were denoised using Filter-Bank artifact rejection. The features extracted include spectral features, higher-order spectral features, hemispheric asymmetry features, coherence, and cross-frequency coupling. The mRMR algorithm was also used for feature selection and the classification algorithms assessed were SVM and LR. However, some of the methods that differ to this study include the use of Filter-Bank Artifact Rejection and the assessment of 11 basic emotions (interest, amusement, happiness, sadness, fear, disgust, anger, hope, relief, surprise, and sympathy). On average, the performance achieved using Leave-One-Subject-Out (LOSO) cross validation was 75% using SVM and 70% using LR and the performance achieved using 10-fold cross validation was 70% using SVM and 71% using LR. They were able to achieve up to 88% performance for the classification of interest with 10-fold LR and the removal of non-classifiable subjects (i.e. subjects with a classification accuracy of less than 60%).

The next closest study was by Kumar *et al.* (2016) which also performed user-independent analysis on one-minute length video stimuli. They used 32 participants

(i.e. the DEAP dataset) which is the second largest dataset examined in the literature; however, they only used 2 EEG channels of the 32 that were recorded for the extraction of higher-order spectral features. They also only performed basic bandpass filtering to reduce noise and classified between high- and low- valence and arousal models using SVMs and ANNs. The performance they were able to achieve was 61.17% for valence and 64.84% for arousal.

Another similar study by Atkinson and Campos (2016) also employed the DEAP dataset. This study also used the mRMR algorithm for feature selection and SVM for classification. However, they used 14 EEG channels for feature extraction and the features extracted include spectral features as well as statistical features (e.g. mean, standard deviation) and fractal dimension (FD). Their model's performance for 2-way valence and arousal was 74.15% and 73.06%, respectively.

Other papers examined were found to have either much less participants or many more channels used in their analysis. For example, Soleymani *et al.* (2012) performed user-independent modelling on 27 participants who watched videos of a similar length. Although they used similar features like spectral features and hemispheric asymmetry and classified using SVM and feature selection, they used 32 channels for their feature extraction. They were able to achieve 52.4% and 57% accuracy for 3-way valence and arousal models, respectively.

Another study was able to achieve up to 89% for the classification of positive and negative emotions using similar feature extraction and classification methods (Nie *et al.*, 2011). However, they only had 6 participants in their study and employed 62 EEG channels in their analysis. Similar results were achieved in a paper by Zheng *et al.* (2016) which also used 62 channels with only 15 participants but assessed

a large range of features, several classification algorithms, feature smoothing, and feature selection. They were able to achieve an accuracy of 83.3% using SVM and 91.1% using graph regularized extreme learning machine (GELM). Another paper by Petrantonakis (2010) was also able to get 83.3% accuracy using SVM and on higher-order spectral features and only 3 electrodes, however, they had only 16 participants and used images of facial expressions for the emotion elicitation stimuli.

In terms of assessing the performance of participant-dependent modelling, there were 6 studies that used videos as their stimuli. Again, the most similar study that performed user-dependent modelling was able to achieve an average of 74% classification using SVM (Dhindsa and Becker, 2017). Most other papers either used more than 32 EEG channels or fewer than 32 participants in their analyses. One paper also performed participant-dependent between-session modelling, since their data was recorded over 3 sessions, and were able to achieve 79.3% accuracy using differential entropy features and GELM (Zheng *et al.*, 2016).

Chapter 6

Identifying challenges in emotion recognition

This section identifies some of the unique challenges involved in this emotion recognition study as well as stating the research problems tackled. Using few electrodes comes with its advantages (i.e. cost efficiency and usability) and disadvantages (i.e. signal processing limitations), as previously mentioned. However, the uniqueness of this dataset lies in the large number of participants involved, which surpasses all the examined literature. This dataset is nearly 4 times the size of the DEAP dataset and almost 3 times the size of the largest dataset assessed in the previous literature. The large number of participants not only comes with its own set of challenges, it can have interesting implications on the generalizability of emotion recognition models.

The main purpose of this research is to generate a participant independent model that can generalize well between participants. Due to the large number of participants, the results obtained can be considered more generalizable for the problem. A well performing general model can be very useful in creating an aBCI system that

does not require extensive training of a classifier to each individual, which can result in overall discomfort for the user and habituation to the stimuli. Since the dataset contains EEG responses to the same videos over two sessions, examining the performance of the second visit can provide insight on the affect of familiarization to the stimuli. The generalizability between sessions can also be examined by training on the first visit and testing on the second visit. Due to the large sample size, the performance of the classifier tends toward lower, more realistic estimates because they are associated with narrower error margins (Neuhaus and Popescu, 2018). Participant dependent models also tend to outperform participant independent models but require a larger number of recorded trials or longer duration stimuli.

Another challenge examined in this study is the time aspect of the stimuli. The video clips used to elicit emotional reactions were around 3 minutes in length, which lies on the longer end of the range as compared to the literature. The duration of the stimuli provides an additional challenge of trying to identify when the emotional response occurs or peaks, which remains uncertain unless participants can continuously record their emotional response throughout the entire video. EEG responses are typically segmented into smaller window sizes for long-duration stimuli prior to classification in order to increase the ratio of trials feature space. Thus, this study tries to determine the optimal length of time segments as input to a support vector classifier for emotion recognition. Since this method cannot take patterns between segmented time windows into account and because the response cannot be assumed to be continuous over the entire duration of the video, the use of convolutional neural nets may be a useful pattern recognition tool.

The input to the emotion classifier is regarded as one of the most crucial aspects of

emotion recognition. Thus, this study compares different feature extraction methods and examines the effectiveness of feature selection algorithms. Some of the computed features included in this study are the commonly used spectral analysis techniques and hemispheric asymmetries. Higher-order spectral analysis is also evaluated for the emotion recognition task. Feature selection of these extracted features is performed using the mRMR algorithms and is compared to manual selection of feature subsets. These feature sets calculated for different time durations can then be used by various machine learning algorithms.

This study also compares support vector machines and convolutional neural networks as the two main machine learning techniques used to classify between the assessed emotional states. Various parameters are varied for both methods to inspect their affect on performance. As mentioned above, convolutional neural networks can receive the entire duration of the video clips to find patterns while the support vector classification method does not. Thus, these two algorithms may be able to learn from different aspects of the EEG recordings.

Furthermore, this study also attempts to compare other important aspects of emotion recognition mentioned above. This includes assessing the performance of classification models using discrete versus dimensional emotion models. We also analyze the performance of models trained on a subset of classifiable participants and compare them to the subset of non-classifiable participants. Finally, we compare the performance of participant-independent models to participant-dependent models and attempt to create a model that can classify across sessions.

Chapter 7

Methodology

The major goal of the simulations reported here was to classify responses to emotional video clips in EEG data. This section describes the methodology used in the various tests performed throughout this study. The methods used for the EEG data acquisition are first detailed. Then the feature extraction methods, feature selection tools, and machine learning algorithms are described. Finally, the tests performed using all the outlined methods are explained.

7.1 Data Acquisition

The EEG dataset used was recorded from a previous study of responses to emotional videos, conducted in the LIVELab at McMaster University as part of a probiotic study with Lallemand Inc. EEG data of 116 participants were acquired from 7 scalp electrodes positioned according to the 10–20 system at CZ, FZ, F3, F4, PZ, A1, and A2 across two sessions. The recording was initially referenced to the CZ electrode with a ground electrode at the PZ position. The data were then re-referenced to the

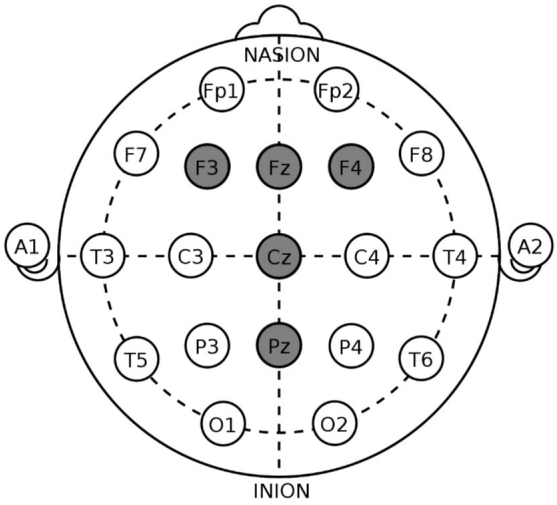

Figure 7.1: The EEG electrode locations used for emotion classification.

mean of the A1 and A2 electrodes offline after acquisition. Data from the remaining 5 electrodes, CZ, FZ, F3, F4 and PZ, were used for all subsequent analyses, as shown in Figure 7.1. The sampling rate of the EEG was 256 Hz and a bandpass filter was applied from 0.1–100 Hz with a notch filter at 60 Hz. The stimuli used to elicit emotions consisted of 8 video clips intended to elicit fearful and happy emotional responses (4 fearful and 4 happy) ranging from 2 to 3.5 minutes, presented in a random order for each session. The participants were asked to provide feedback on each of these videos on a valence and an arousal integer scale from 1 to 7 as well as provide written feedback. In addition, there were 2-minute pre- and post- baseline neutral videos as well as 7 30-second neutral videos between the emotional videos; however, only the responses to emotional videos were examined in this study. The experiment was conducted over 2 sessions where the participants were shown the same stimuli.

7.2 Feature Extraction

The features extracted for the task of classifying emotional EEG responses include spectral features, hemispheric asymmetries in power in particular frequency bands, cross-electrode coherence at particular frequencies, and higher order spectral features. Spectral features consist of the power spectral density and cross-power spectra density at individual electrodes, while higher order spectral features consist of bispectrum and bicoherence measures at individual electrodes.

7.2.1 Spectral Features

The power spectral density (PSD) was estimated for the delta (<4 Hz), theta (4–7 Hz), alpha (8–15 Hz), beta (16–31 Hz), and gamma (>32 Hz) frequency bands for each of the 5 electrodes. The PSD was according to the definition

$$P_{XX}(f) = \lim_{T\to\infty} \mathbf{E}[|\hat{X}_T(f)|^2] \tag{7.1}$$

where $\hat{X}_T(f)$ is the Fourier transform of a signal X. This resulted in a total of 25 features. The EEGLAB toolbox was used to perform bandpass filtering to obtain the spectral features (Delorme *et al.*, 2011).

The cross-spectral density (CSD) extends the definition of the PSD to the Fourier Transform of the product of two signals as defined in

$$P_{XY}(f) = \lim_{T\to\infty} \mathbf{E}[\hat{X}_T(f)\hat{Y}_T(f)] \tag{7.2}$$

where $\hat{X}_T(f)$ and $\hat{Y}_T(f)$ are the Fourier transforms of signals X and Y, respectively.

The CSD was calculated between each pair of electrodes with respect to the delta, theta, alpha, beta, and gamma bands yielding a total of 50 features.

7.2.2 Hemispheric Asymmetry Features

It has been found that asymmetry in alpha power between the frontal cortices is associated with emotional responses (Harmon-Jones *et al.*, 2010). Relatively increased left-frontal activity is linked to approach motivations (e.g. joy and anger) while relatively increased right-frontal activity is linked to withdrawal motivations (e.g. fear and disgust). The frontal alpha asymmetry (FAA) was calculated using the F3 and F4 electrodes using the formula

$$AA = \frac{L - R}{L + R} \tag{7.3}$$

where L and R are left and right alpha power, respectively. In addition, the frontal asymmetry of the other bands (i.e. delta, theta, beta, and gamma) were also calculated to examine their relation to the emotional responses.

7.2.3 Coherence Features

The coherence between all pairs of electrodes were computed. Coherence between two signals is defined as

$$C_{XY}(f) = \frac{|P_{XY}(f)|^2}{P_{XX}(f)P_{YY}(f)} \tag{7.4}$$

where P_{XY} is the cross spectral density between the two signals. The coherence was also calculated between all pairs of electrodes with respect to the delta, theta, alpha, beta, and gamma bands, resulting in 50 features.

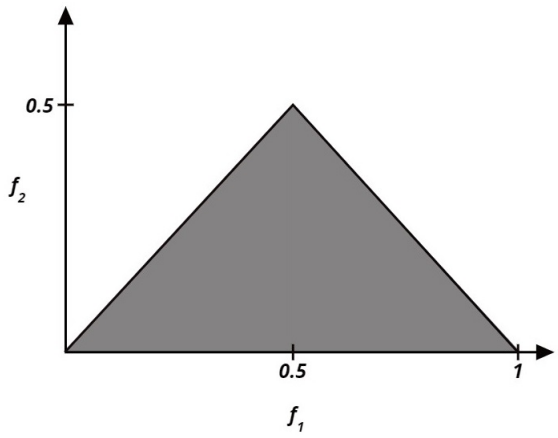

Figure 7.2: The non-redundant region used for computation of bispectrum.

7.2.4 Higher-Order Spectral Features

The bispectral feature used here is the 2D Fourier Transform of the third order cumulant generating function, hereafter referred to as bispectrum of a signal. The bispectrum and bicoherence measure quadratic non-linear interactions between pairs of frequencies and they are computed from the non-redundant region of the bispectrum defined as the triangle $f_2 \geq 0$, $f_1 \geq f_2$, and $f_1 + f_2 \leq \pi$, as shown in Figure 7.2 (Sigl and Chamoun, 1994).

The bispectrum is computed by the equation

$$B(f_1, f_2) = \mathcal{F}^*(f_1 + f_2)\mathcal{F}(f_1)\mathcal{F}(f_2) \tag{7.5}$$

where $\mathcal{F}$ is the Fourier Transform and $\mathcal{F}^*$ is its complex conjugate. The bicoherence is the normalized bispectrum for a signal as computed by the equation

$$B_c(f_1, f_2) = \frac{|\sum_n \mathcal{F}^*(f_1 + f_2)\mathcal{F}(f_1)\mathcal{F}(f_2)|}{\sum_n |\mathcal{F}^*(f_1 + f_2)\mathcal{F}(f_1)\mathcal{F}(f_2)|} \tag{7.6}$$

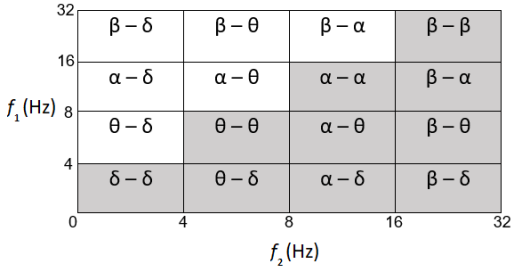

Figure 7.3: The different regions used for analysis in the bifrequency plane.

These bispectrum and bicoherence features were calculated with respect to each pair of frequency bands for each of the 5 channels, resulting in a total of 50 features, as shown in Figure 7.3. These higher order spectral features were computed using the HOSA (Higher-Order Spectral Analysis) toolbox for MATLAB (Swami *et al.*, 1993).

7.3 Feature Selection

Feature selection was performed using the minimum Redundancy Maximum Relevance (mRMR) feature selection method (Peng *et al.*, 2005). The mRMR feature selection method maximizes the mutual information between the features of the feature subset and the training labels while also minimizing the mutual information among the selected features within each class. Two mRMR algorithms were tested, one using mutual information difference (MID) and the other using mutual information quotient (MIQ). The quotient-combination has a greater penalty on the redundant features than the difference-combination of mRMR (Ding and Peng, 2005). The advantages of feature selection include dimension reduction to decrease computational cost and reduction of noise to improve classification accuracy. Another commonly

used technique to reduce the dimensionality of the feature space is Principal Component Analysis (PCA), however, this method transforms the original feature space and the PCA dimensions are not as readily interpretable as the original features.

7.4 Machine Learning Algorithm

The two machine learning algorithms used in this study for the purpose of classification between fearful and happy emotional responses are Support Vector Machines (SVM) and Convolutional Neural Networks (CNN).

7.4.1 Support Vector Machines

The aim of Support Vector classification is to devise a computationally efficient way of learning separating hyperplanes in a high dimensional feature space (Cristianini and Shawe-Taylor, 2000). A hyperplane is a subspace that is one dimension less than its input feature space which is meant to separate the two categories of the input data. A computationally efficient algorithm would be able to deal with very large sample sizes of 105 and up. The SVM algorithm is based on the assumption that the best hyperplane is one that represents the largest separation, or margin, between the two classes so that the distance to the nearest points from both classes is maximized. If test data follow the same probability distribution as the training data, then this should lead to good generalization performance. This linear classifier is known as a maximal margin classifier or large margin classifier. The hyperplane depends on the distance to the points that lie closest to the hyperplane, which are called support vectors, as illustrated in Figure 7.4.

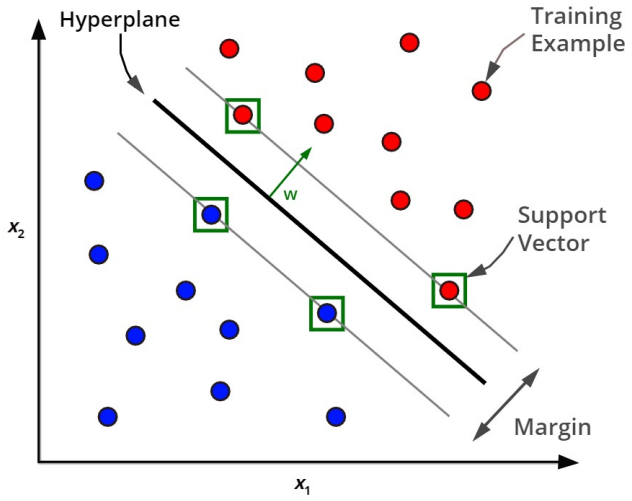

Figure 7.4: An illustration of the large margin classifier hyperplane of the SVM where x_1 and x_2 represent two features. The support vectors are the chosen training examples that lie closest to the hyperplane.

SVMs are also able to perform non-linear classification using the kernel trick. The x input feature space is transformed by a function g(x), which is referred to as the kernel. The kernel function nonlinearly transforms the data into a space in which the classes are linearly separable, allowing the algorithm to fit the maximal margin hyperplane in the transformed feature space. This means that this kernel trick allows the creation of a non-linear separating hyperplane in the original input space. One commonly used kernel in EEG classification is the radial basis function (RBF) or Gaussian kernel.

Although there aren't any SVM parameters that directly control the number of support vectors chosen by the algorithm, the gamma and C parameters can affect this number. The parameter which controls the radius of the radial basis function kernel used by the SVM is usually referred to as gamma (or its inverse, sigma) or

the KernelScale in the MATLAB SVM toolbox. The penalty parameter of the error term is usually referred to as C or BoxConstraint in the SVM toolbox and controls the trade-off between a smooth decision boundary and classifying the training points correctly. Both these parameters are tested to examine their effect on the accuracy.

7.4.2 Convolutional Neural Networks

Convolutional neural networks are artificial neural networks that can learn local patterns in input data through the use of learned convolutional kernels (also referred to as a filter or a feature map). In essence, CNNs are neural networks that use convolution in place of general matrix multiplication in at least one of their layers (Goodfellow *et al.*, 2016). In recent years, CNNs have been increasingly successful in various applications, such as computer vision and speech recognition (LeCun *et al.*, 2015). CNNs tend to perform best on signals which have an inherent hierarchical structure by representing high level features as compositions of low level features through the use of multiple layers and learning local non-linear features. While some studies have examined the use of CNNs in EEG classification tasks (Schirrmeister *et al.*, 2017), they have not been widely examined in the use of EEG emotion recognition.

Convolution is a commonly used operator in signal processing defined as the sum of the product of two functions, typically a signal and a kernel function, after the kernel is reversed and shifted in time. This is similar to the cross-correlation of a signal with a kernel function except that the kernel is not reversed; however, some implementations of the CNN use cross-correlation rather than convolution. The CNN roughly performs convolution by applying the weighted sum of the product of a kernel function with an image or signal at multiple spatial or temporal positions. A convolutional layer

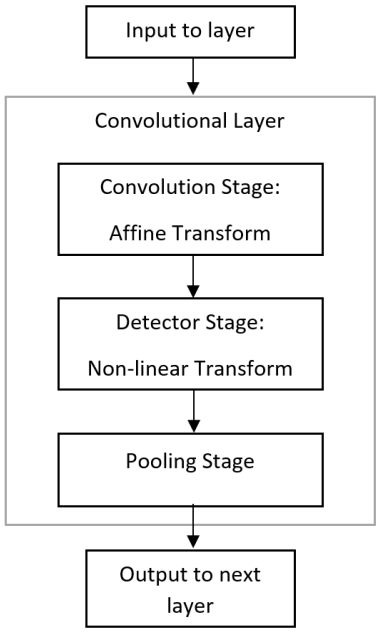

Figure 7.5: The different stages that are typically included in each convolutional layer of a CNN.

typically consists of a convolution stage where the kernel is convolved with its input through a learned affine transformation, a detector stage where the resulting signal is transformed through a pre-chosen nonlinear activation function (such as a sigmoid or a rectifier), and a pooling stage that computes the maximum or average of a group of units from the detector stage output in order to reduce its size (see Figure 7.5).

One or more of these convolutional layers can be used in a CNN. The changeable parameters involved in the convolutional layers include the number of kernels, the size of the kernels, the stride of the kernel, the activation function and the size and stride of the pooling layer. The values of the units in the kernel themselves are learned through the CNN and are not a changeable parameter in the network. These layers are then followed by a fully connected dense layer which also contains a nonlinear activation function. Although any number of dense layers can be used, typically only

one layer with a variable number of neurons is used with multiple convolutional layers.

With EEG as the input to the CNN, a 1-dimensional CNN is typically used although sometimes channels are treated as the second dimension in the input. The raw EEG data can be used in the input layer; however, this study mainly focuses on the use of computed features in the CNN rather than using the CNN as a feature extraction method. The parameters examined in this study are the number of convolutional layers (i.e. 1 or 2 layers), the number of filters used in the CNN (i.e. n = 4, 8, 16, 32, or 64 filters), the size of the kernel used in the filter (i.e. k = 2, 3, or 4 units), and the number of neurons in the dense layer (i.e. 10, 50, 100, 500, 1000, or 1500 neurons). A rectifier (also referred to as rectified linear unit or ReLU) was used as the activation function in the detector stage of the CNN while a sigmoid activation function was used in the fully connected dense layer of the neural network. An Adam optimizer, an extension to stochastic gradient descent (Kingma and Ba, 2014), was used to optimize the learning rate during training. The architectures of the CNN models created are depicted in Figure 7.6. All CNN models were created using the Keras neural network API with TensorFlow.

7.5 Test Cases

Many tests were performed on this EEG dataset using the methods described above; however, this section highlights some of the main tests conducted in this study. The first main test case consisted of an investigation on the ideal size of time segments for use in classification. In order to prepare the features as input to the SVM, the EEG data were segmented into shorter time windows in order to increase the amount of training data. This assumes that the emotional response is relatively uniform

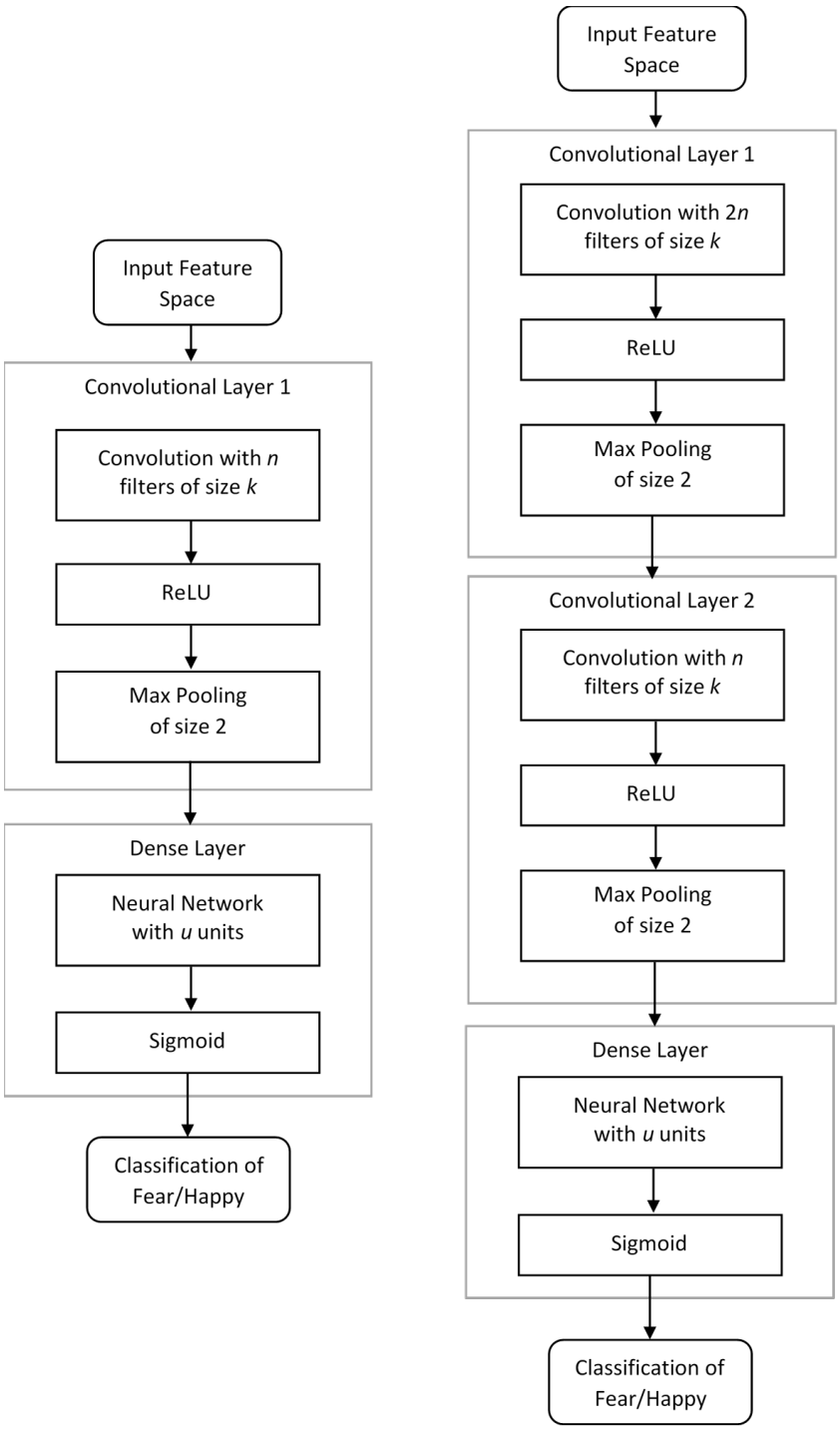

Figure 7.6: The architecture of the 1-layer and 2-layer CNN models examined in this study.

throughout the entire length of the video and treats each time segment as an independent trial. We attempted to determine the best size for a time window over which to calculate the features and segment the 2 to 3-minute EEG data. These time segments were also tested using the CNN method; however, unlike SVM, the CNN approach can identify patterns between the time segments as well and thus they are not assumed to be independent.

Another test case involves the comparison of the various feature extraction methods and the evaluation of the feature selection method. The classification accuracy of each feature extraction method was evaluated separately. Various combinations of these feature sets were also chosen manually and compared to the use of the algorithmic feature selection method to evaluate its classification performance. This allows us to determine which feature extraction method is most relevant to the problem or which set of individual features provide the best accuracy.

The two classification algorithms and cross-validation methods were also examined and compared for this dataset. Both the SVM and CNN approach contain various parameters that affect the performance of the model. For the SVM algorithm, this includes parameters that affect the number of support vectors chosen as well as the type of kernel used. Several models of the CNN approach were also tested by varying the parameters mentioned above.

Another test case performed was the the analysis of dimensional emotion models on our dataset. Although regression models can be created for valence and arousal using Support Vector Regression, this paper focuses on classification using 2-way or 3-way classification of valence and arousal due to their improved performance. Thus, we were able to compare the performance of classification models based on discrete

emotion models to classification models based on dimensional emotion models.

In order to determine whether we can achieve improved performance on our best model by reducing the number of participants, we created another model using the best performing participants. We identified the participants for which the model obtained an accuracy of 60% or higher and created a classifiable participant model. We also created another model using the subset of non-classifiable participants, whose test accuracy was below 60%, to observe whether improved performance can be achieved on this group.

Although the majority of this study is focused on participant-independent modelling, participant-dependent models were also tested on the best performing feature set and time segment using SVM. This comparison allows us to determine whether a general model is relatively feasible since participant-dependent models are generally expected to work better for the individual.

Finally, the difficult task of performing between-session classification was tested on the best performing case. The model was trained using the first session and tested on the participants in the second session. This allows us to determine the ability of the model to generalize over time throughout different sessions.

Chapter 8

Results

The results for this study focus on three main aspects of emotion recognition: feature extraction and feature selection, time segment analysis, and analysis of classification algorithms. The first section deals with comparing the various feature extraction methods based on test set accuracy and the effectiveness of feature selection algorithms and determines the best feature set. The second section addresses issues with length of time segments over which to extract the features for the two classification algorithms using the best feature set to determine the best time segment. The third section uses the best feature set and time segment to determine the best classification parameters for the respective classifiers for achieving the best accuracy. Sections 4 to 7 summarize findings from analyses including the use of valence vs. arousal models, classifiable participant modelling, participant-dependent modelling, and between-session modelling, respectively.

Some of the main findings from the results are briefly summarized here for brevity. In sections 1 and 2, the best performing model obtained a test accuracy of 75.86%

with SVM using 20 second time segment input with the bicoherence feature set. Section 3 found that the CNN models were only able to perform up to 65.40% after the parameter search while SVM model performance remained the same. Section 4 shows that models based on the classification of valence and arousal underperformed the classification of discrete emotions of fearful and happy responses, with valence outperforming arousal at 70.21%. Section 5 found that the model trained on classifiable participants only was able to outperform our best model and get 78.57%. Section 6 was only able to achieve up to 72.71% accuracy for participant-dependent modelling. Finally, section 7 showed some reliability across sessions by performing at 67.13% for participant-independent between-subject classification.

8.1 Feature Extraction and Feature Selection Analysis

A major component of emotion recognition is determining the best features to represent the dataset for classification, meaning that the features should be as separable as possible. This presents a major challenge in the application of machine learning; in this case the classification between fearful and happy affective states. Some commonly used features in EEG emotion recognition are tested including spectral analysis and hemispheric asymmetry, as previously mentioned. These were assessed through the power spectral density (PSD) and frontal alpha asymmetry (FAA) features described in the methodology section.

Other features that were tested include higher-order spectral features like the bispectrum and bicoherence features as well as coherence and cross spectral density.

However, in this study it was found that the models trained with the coherence, cross spectral density, and bispectral features were not able to perform over chance accuracy (i.e. 50% test accuracy) and were therefore eliminated from the study. These features were found to reduce the test accuracy when combined with other feature sets, and the accuracy did not improve with the use of feature selection algorithms.

Thus, the main features examined in this study that were able to provide above chance accuracies were the PSD, FAA, and bicoherence (BIC) features. In addition to these features, the RMS feature was used as a feature to measure the root-mean-squared value of all the channels for a given input. This provides information on the overall amplitude of the signal, as a large RMS value corresponds to high amplitude signals. Since artifacts like motion artifact tend to have a higher amplitude than the EEG signals, the RMS feature is used in order to get an understanding of how much the artifacts contribute to classification accuracy alone or whether this information can actually be useful to determine between the fearful and happy emotional states. All computed features were standardized prior to classification, as this has been found to improve classification accuracies.

Table 8.1 shows the training and test accuracies of using the entire feature set for the separate features on visit 1 and 2 using 20 second time length data to compute the features. The machine learning algorithm used to classify between the two states was SVM and the cross-validation method used was a 10-fold cross validation where approximately 10% of the data is completely separated from the training data in order to obtain the test accuracy. This table shows that the best feature set examined was the bicoherence feature set. It was able to achieve 73.01% training accuracy on visit 1 using all 50 features. It also shows that the model trained using the RMS feature

Table 8.1: Training and test accuracy for each feature set

	Visit 1		Visit 2	
Feature Set	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
BIC	94.60	**73.01**	94.65	**65.13**
FAA	99.37	57.71	98.53	63.65
PSD	92.08	64.35	91.75	59.02
RMS	99.83	65.18	99.52	64.24

alone was able to predict emotion with above chance accuracy (65.18% on visit 1) and perform better than the models trained with some of the other computed features tested.

Furthermore, all possible combinations of these feature sets were merged to determine if any of these combinations can outperform the bicoherence features alone. Table 8.2 shows the training and test accuracy for these combined features for visit 1 and visit 2. The highest accuracy obtained was using the combination of the bicoherence and PSD features to provide an accuracy of 71.92% on visit 1 using 10-fold cross validation with the combination of bicoherence and RMS features coming in as a close second. None of the models trained on the combined feature sets were able to outperform the use of bicoherence features alone. The model trained on the combination of all the features obtained a test accuracy of 69.03% on visit 1.

The use of feature selection was then introduced using the mRMR feature selection algorithm to determine whether this can improve the accuracy of the model trained on the best performing features. Both the mutual information difference (MID) and mutual information quotient (MIQ) algorithms were tested on the bicoherence feature set. Figure 8.1 and Figure 8.2 show the training and test accuracy using the MID and MIQ feature selection methods, respectively. Shaded regions in all the following

Table 8.2: Training and test accuracy for combinations of feature sets

	Visit 1		Visit 2	
Feature Set	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
BIC, FAA	94.00	68.54	91.95	59.40
BIC, PSD	95.02	**71.92**	94.51	**64.96**
BIC, RMS	95.33	71.38	94.47	64.80
FAA, PSD	92.54	62.86	92.36	60.36
FAA, RMS	99.09	59.18	98.60	64.32
PSD, RMS	90.50	65.17	89.80	62.71
BIC, FAA, PSD	93.74	69.24	91.90	58.95
BIC, FAA, RMS	94.50	68.92	92.01	59.14
FAA, PSD, RMS	92.43	63.90	92.27	60.79
All Features	92.85	69.03	90.63	59.32

Figures represent the standard error of the accuracy. The number of features included are ordered from most relevant, least redundant to least relevant, most redundant features.

Table 8.3 summarizes the findings in Figures 8.1 and 8.2 by providing the best test accuracy obtained from both feature selection algorithms (i.e. 75.65% and 75.86% accuracy, respectively). They both obtained almost the same accuracy, although with a different number of features. Both models were able to outperform the use of all bicoherence features only by about 2%. Figures 8.1 and 8.2 show that the test accuracy does not vary dramatically with the number of features as it seems to be fairly consistent. However, this does show that only about 10 to 15 features are necessary to obtain high accuracy, which can reduce computation time in online applications. Unfortunately, the features selected using the MID and MIQ methods were not always consistent between models using 10-fold or Leave-one-subject-out (LOSO) cross validation. There was no particular frequency pairing per electrode

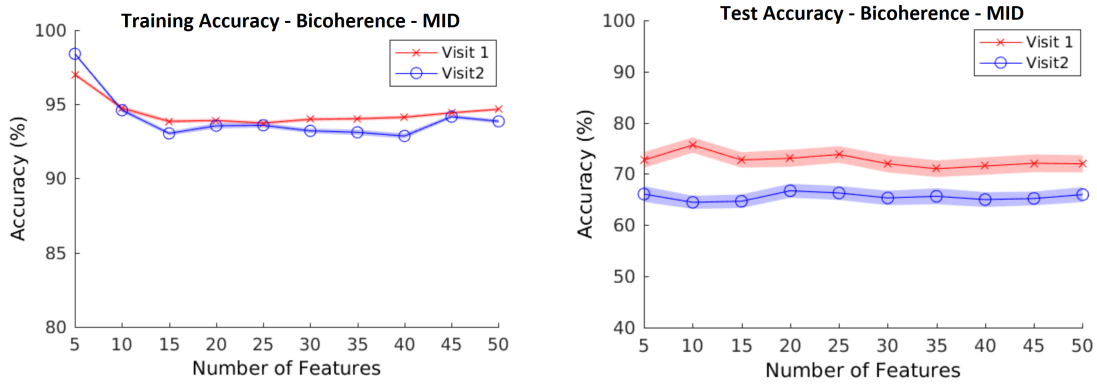

Figure 8.1: Training and test accuracy with respect to number of features selected using the MID feature selection algorithm on the bicoherence feature set.

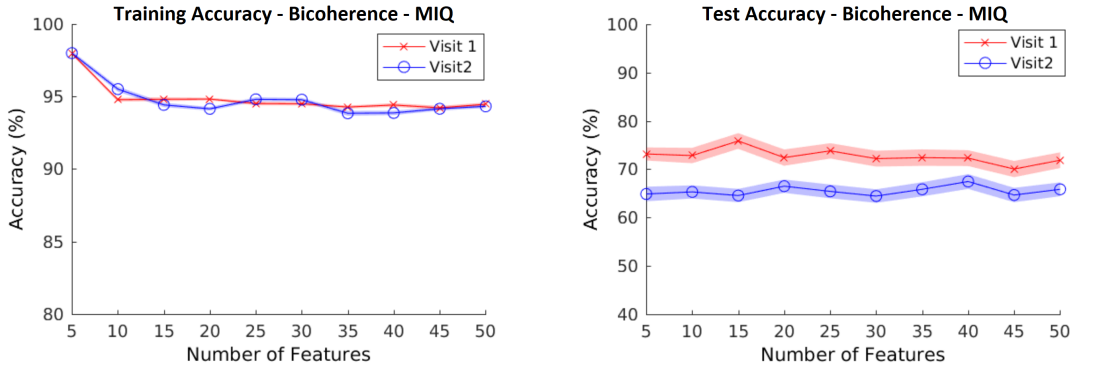

Figure 8.2: Training and test accuracy with respect to number of features selected using the MIQ feature selection algorithm on the bicoherence feature set.

Table 8.3: Best training and test accuracy of the bicoherence feature set with feature selection

	Visit 1		Visit 2	
mRMR Algorithm	Best Test Accuracy	Number of Features	Best Test Accuracy	Number of Features
MID	75.65	10	66.70	20
MIQ	**75.86**	15	67.46	40

that was consistently chosen to be in the top 5 features selected. The Figures also show that the training accuracy when using few features suggesting that the SVM may be overfitting when given only 5 features for classification.

Using the best performing test case of 15 selected features with MIQ, a confusion matrix was plotted in Figure 8.3 where 0 represents the Happy class and 1 represents the Fearful class. The rows correspond to the predicted class (i.e. output class) and the columns correspond to the true class (i.e. target class). The diagonal cells correspond to the observations that are correctly classified and off-diagonal cells correspond to incorrectly classified observations. Both the number of observations and the percentage of the total number of observations are shown in each cell. The column on the far right shows the percentages of all the examples predicted to belong to each class that are correctly and incorrectly classified, referred to as precision and false discovery rate, respectively. The row at the bottom shows the percentages of all the examples belonging to each class that are correctly and incorrectly classified, referred to as the recall (or true positive rate) and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy. The proportion of fearful to happy data was close to being the same although not exactly equal. This matrix shows that the rate of false positives and negatives are also roughly equal, suggesting

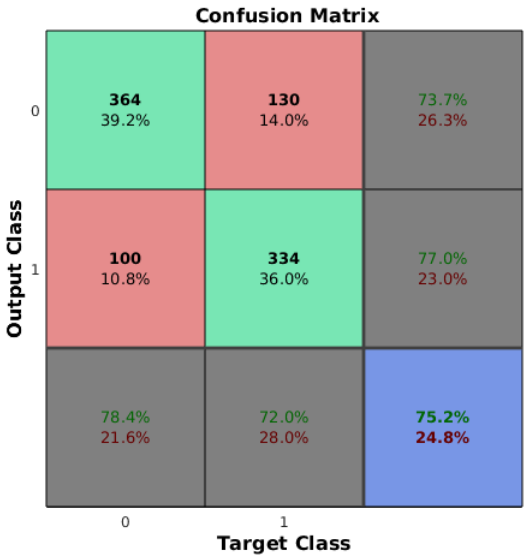

Figure 8.3: Confusion matrix of best performing feature set.

that the model created is not more likely to predict one class over the other.

In addition to performing feature selection on the bicoherence feature set, feature selection was also tested on the combination of all the features to see which features are chosen as most relevant and least redundant by the MID and MIQ algorithms. Figure 8.4 and Figure 8.5 show the training and test accuracy using the MID and MIQ algorithms, respectively. Table 8.4 summarizes the best obtained accuracy and the number of features used. The two algorithms also performed similarly in terms of best accuracy with MIQ being slightly higher (i.e. 69.42%) using 60 features. There seems to be little to no trend between test accuracy and number of features used and the addition of feature selection only improved the accuracy by about 0.4% for this case. Some bicoherence features were chosen as most relevant (i.e. in the top 5 features); however, a consistent trend was the frontal alpha asymmetry feature being

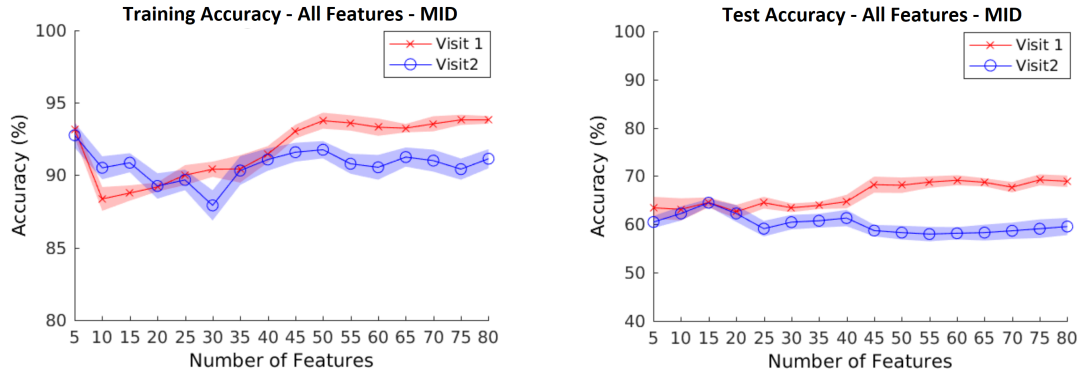

Figure 8.4: Training and test accuracy with respect to number of features selected using the MID feature selection algorithm on all extracted features.

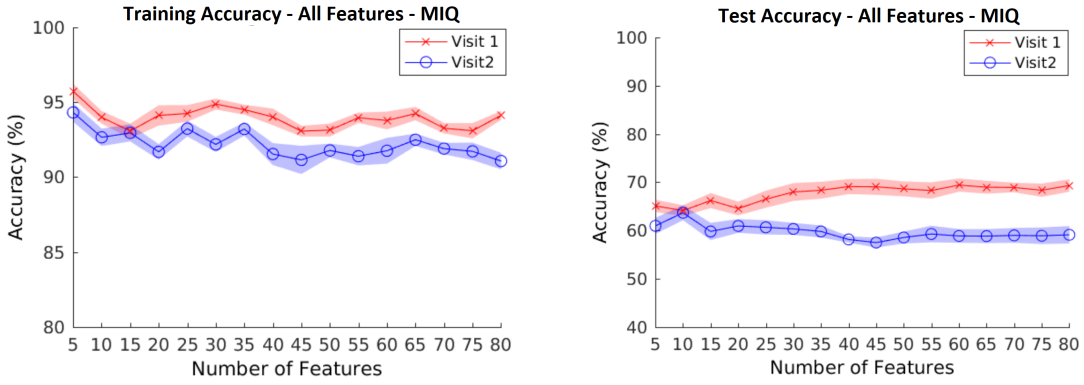

Figure 8.5: Training and test accuracy with respect to number of features selected using the MIQ feature selection algorithm on all extracted features.

in the top 5 features in all the models. Other asymmetry features, the beta and theta frontal asymmetry, also appeared in the top 5 chosen features. Although these features were found to be relevant to the problem, they only gave a test accuracy of about 65% for the top 5 features, as shown in the Figures.

Other combinations of features were also tested; however, none were able to provide better results than the bicoherence features with feature selection. One set that was chosen was the use of the best 15 features from the bicoherence MIQ test with

Table 8.4: Best training and test accuracy using feature selection on all extracted features

	Visit 1		Visit 2	
mRMR Algorithm	Best Test Accuracy	Number of Features	Best Test Accuracy	Number of Features
MID	69.21	75	64.47	15
MIQ	**69.42**	60	63.37	10

the addition of frontal alpha asymmetry and RMS features. These particular features were chosen because of their relevance to emotion recognition in previous studies and their performance in this study. This resulted in a test accuracy of 72.95% which is still less than using the best bicoherence features alone (i.e. 75.22%). Performance using LOSO and 10-fold cross validation was similar but slightly higher for the LOSO method.

8.2 Time Segment Analysis

Another challenge this study attempts to address is time aspect issues with classification. Since the stimuli for emotion recognition studies can span several minutes, the ideal raw EEG data window length over which to calculate the extracted features is unclear. This analysis attempts to find the best length for the time segment. It also investigates whether having overlapping time windows in the raw EEG can produce additional trials that improve the overall performance.

8.2.1 SVM Time Segment Analysis

In the case of using SVM as the classification algorithm, the splitting of data into several time segments increases the number of training samples, which is especially useful when dealing with a smaller dataset or a large number of extracted features. With SVM, each time segment is a separate sample, so it is assumed that the emotional response is constant throughout the duration of the video. Since the stimuli were at minimum 2 minutes in length, the time segments analyzed were between 10 seconds to 30 seconds. It was found that segments less than 10 seconds performed worse and required much longer computation time and segments longer than 30 seconds provided too few training samples with respect to the feature space. We also examined zero overlap and 50% overlap between the time windows of the raw EEG data as input to the feature extraction phase.

The results of the models trained on the best performing feature set are shown in Figures 8.6 to 8.9. These Figures show the best test accuracy obtained with feature selection with respect to length of the time segment. The cross-validation method used to obtain these results was the LOSO method. Figure 8.6 shows the best test accuracy for zero and 50% overlap, respectively, using the MID feature selection algorithm while Figures 8.7 shows the same results for the MIQ feature selection algorithm. Table 8.5 is provided as a summary of the results from the plots. These results show that using the bicoherence feature set, the best performing time segment length is 20 seconds for both feature selection algorithms.

Figures 8.8 and 8.9 show the same results for bicoherence features but with the 10-fold cross validation method. This method provided a best accuracy of nearly 75% for the 25 second time length. These results show comparable accuracy for the

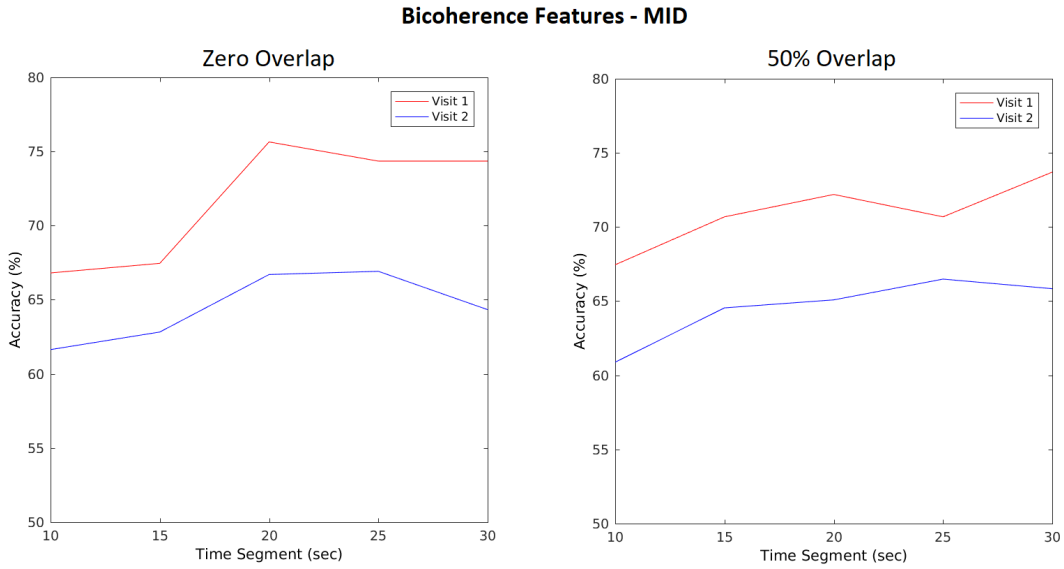

Figure 8.6: The test accuracy for zero overlap and 50% overlap time segments using the bicoherence feature set and MID feature selection algorithm. The test accuracy was calculated using the LOSO cross-validation method.

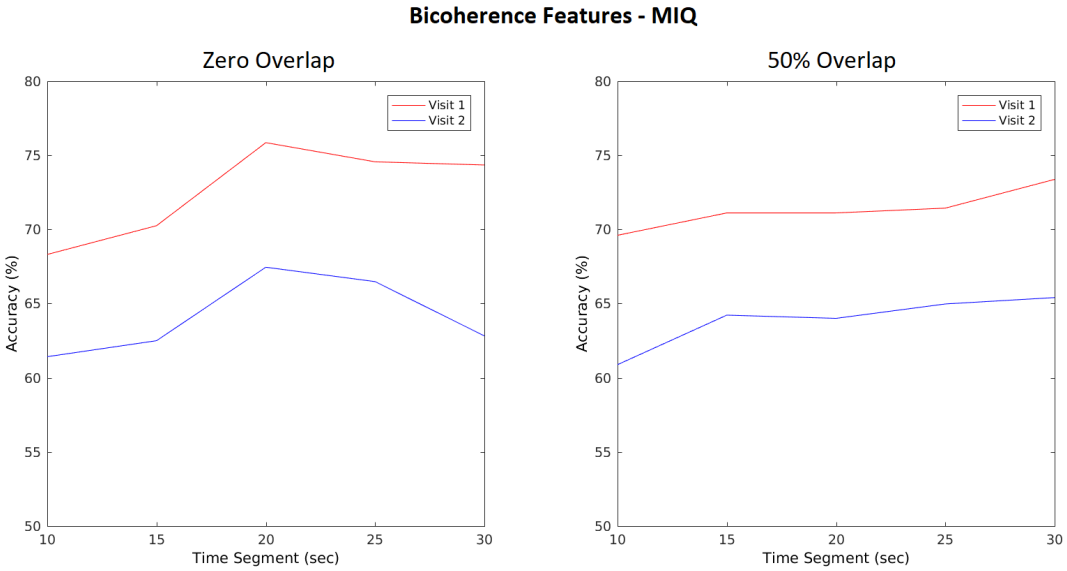

Figure 8.7: The test accuracy for zero overlap and 50% overlap time segments using the bicoherence feature set and MIQ feature selection algorithm. The test accuracy was calculated using the LOSO cross-validation method.

Table 8.5: Test accuracy and number of features for each time segment length and amount of overlap using both feature selection algorithms

	MID				MIQ			
	Zero Overlap		50% Overlap		Zero Overlap		50% Overlap	
Time Segment Length	Best Testing Accuracy	Number of Features	Best Testing Accuracy	Number of Features	Best Testing Accuracy	Number of Features	Best Testing Accuracy	Number of Features
10	66.81	50	67.46	15	68.32	35	69.61	25
15	67.46	15	70.69	15	70.26	30	71.12	45
20	**75.65**	**10**	72.20	20	**75.86**	**15**	71.12	30
25	74.35	15	70.69	15	74.57	30	71.44	40
30	74.35	15	73.71	5	74.35	25	73.38	25

25 and 30 second time windows. The bicoherence results in general seem to show a slight decrease in accuracy with the 50% overlapping windows suggesting that the extra training data does not provide a significant improvement in results. However, the chosen length of a time segment has an affect on accuracy and can improve the accuracy by about 10%.

8.2.2 CNN Time Segment Analysis

In the case of using CNN as the classification model, the implications of the time aspects differ. The input to the first layer of the CNN can receive the entire duration of the video stimuli. When using the feature space as input, the length of the time segment has an effect on the features extracted. However, unlike SVM the time segments are not taken as independent of one another. The CNN can determine patterns across time segments and may be better able to handle fluctuations in emotional response.

Using the best feature set determined with SVM (i.e. bicoherence), the length of the time segments was examined from 1 to 30 second windows with zero and 50% overlap. The parameters of the CNN were kept constant; however, a search for

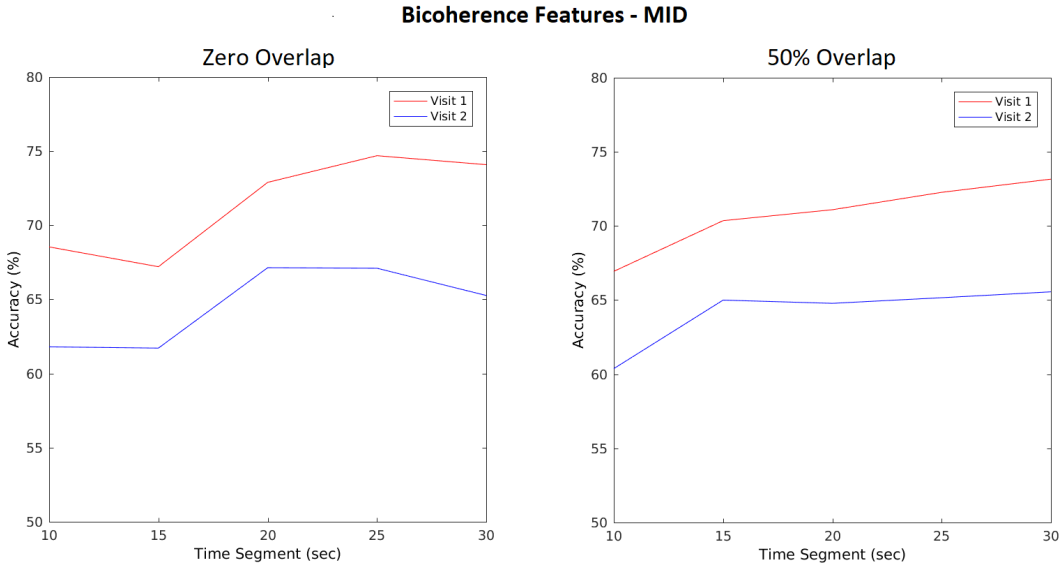

Figure 8.8: The test accuracy for zero overlap and 50% overlap time segments using the bicoherence feature set and MID feature selection algorithm. The test accuracy was calculated using the LOSO cross-validation method.

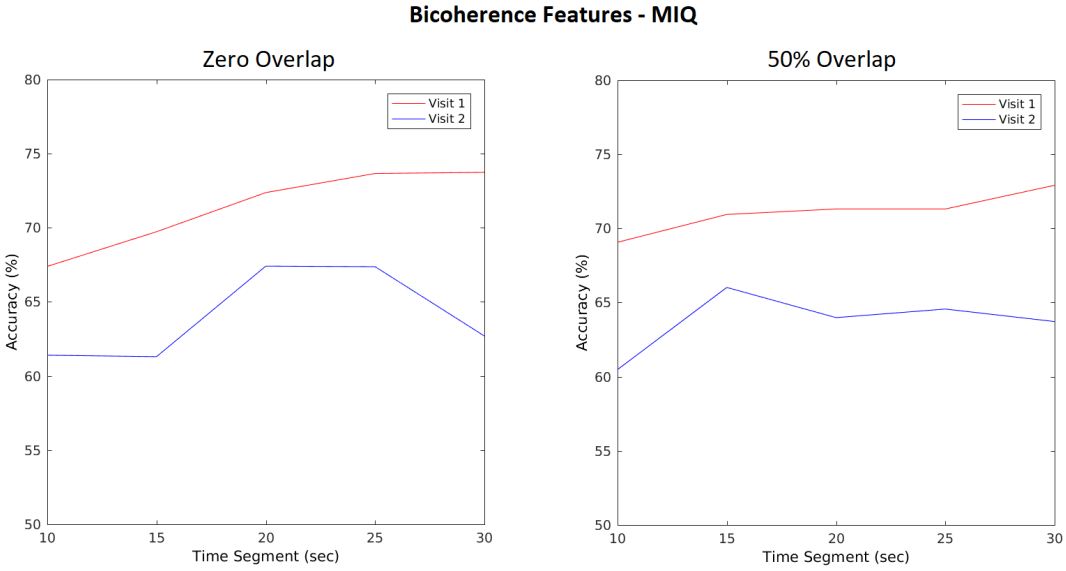

Figure 8.9: The test accuracy for zero overlap and 50% overlap time segments using the bicoherence feature set and MIQ feature selection algorithm. The test accuracy was calculated using the LOSO cross-validation method.

Table 8.6: Training and Test accuracy with respect to time segment length and amount of overlap trained on the 1-layer CNN model

	Zero Overlap		50% Overlap	
Time Segment Length	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
1	100.00	61.14	99.97	**65.18**
2	100.00	**64.99**	100.00	64.00
3	100.00	60.54	99.97	62.49
5	99.97	58.11	100.00	62.61
10	100.00	57.56	100.00	58.93
15	99.87	54.32	99.95	56.26
30	99.16	53.29	99.73	53.54

the best CNN parameters can be found in the following section. The CNNs tested here were a 1-layer and a 2-layer convolutional network with 32 filters (and 64 filters for the second layer of the 2-layer CNN) of length 2 and 10 neurons in the fully-connected dense layer. The pooling layer consisted of max pooling filter of size 2 and the activation functions used were the ReLU for the CNN and sigmoid at the output of the dense layer. Only the first visit was tested using CNNs.

Table 8.6 and Table 8.7 show the results of the time segment analysis for the 1-layer and 2-layer CNNs, respectively. The model with the best performing time segment was the 1-second segment and 50% overlap which gave a test accuracy of 65.18%. The best accuracy for the non-overlapping windows came in at a close second at 64.99% with the 2-second segment. The 2-layer CNN seemed to perform best with the 2-second overlapping and non-overlapping time segments. These results suggest that shorter time segments are ideal for use in a CNN input feature space. They also show that the choice of time segment can affect performance by about 10% as well.

The results of the time segment analyses show that the choice of EEG window

Table 8.7: Training and Test accuracy with respect to time segment length and amount of overlap trained on the 2-layer CNN model

	Zero Overlap		50% Overlap	
Time Segment Length	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
1	98.97	58.87	99.06	63.14
2	99.97	**60.53**	99.84	**64.31**
3	99.27	60.47	99.95	61.66
5	100.00	54.69	100.00	61.66
10	100.00	59.05	100.00	60.39
15	99.68	55.27	100.00	54.30
30	99.78	54.76	99.60	54.02

length can significantly affect the accuracy of the model. The best time segment will depend on the type of classification algorithm used since the CNN seemed to require a time segment that was 10 times shorter than that of SVM due to the nature of the classifier. The best performing time segments of 20 seconds for SVM and 2 seconds for CNN are then used to determine the best classification algorithm parameters in the following section . The results of the time segment analysis show that the SVM outperformed the CNN; however, further analysis of these classification algorithms is conducted in the following section.

8.3 Classification Algorithm Analysis

The machine learning algorithm used to classify between emotional states has a variety of parameters that can have an affect on the outcome of emotion recognition. Investigation into these parameters can improve the accuracy of classification and provide insights on addressing some of the challenges of EEG-based classification.

8.3.1 SVM Parameter Search

The SVM is commonly used in a variety of machine learning classification problems including EEG emotion recognition as it provides an efficient algorithm for creating separating hyperplanes and does not require as much fine tuning of parameters as their neural network counterparts. However, one issue often encountered in this study is the overfitting of the training data leading to high training accuracy but lower test accuracy. Further analysis of the number of support vectors shows that a large number of support vectors may be chosen relative to the total available trials. Figure 8.10 shows the percentage of the training trials that are chosen as support vectors with respect to the number of features selected for the best performing bicoherence feature set. It also shows the training and test accuracy for the data set. The training accuracy using far less than 10 features is high and reaches near perfect training accuracy using only one feature. The number of support vectors is also proportional to the training accuracy, making it clear that the model is overfitting to the training set.

Attempting to reduce the number of support vectors, the gamma and C parameters of the SVM were varied on the best performing bicoherence feature set. Table 8.8 and Table 8.9 show the performance of the models as a function of different values for the gamma and C parameter, respectively. In the previous test cases shown above, the default values for the parameters were chosen. For the gamma parameter, the default setting is an algorithm that automatically chooses the value for the radius of the kernel. The C parameter default setting is set at C = 1. Table 8.8 shows that while gamma may not be directly correlated to the percentage of training points that are support vectors, it does largely affect training accuracy and thus test accuracy.

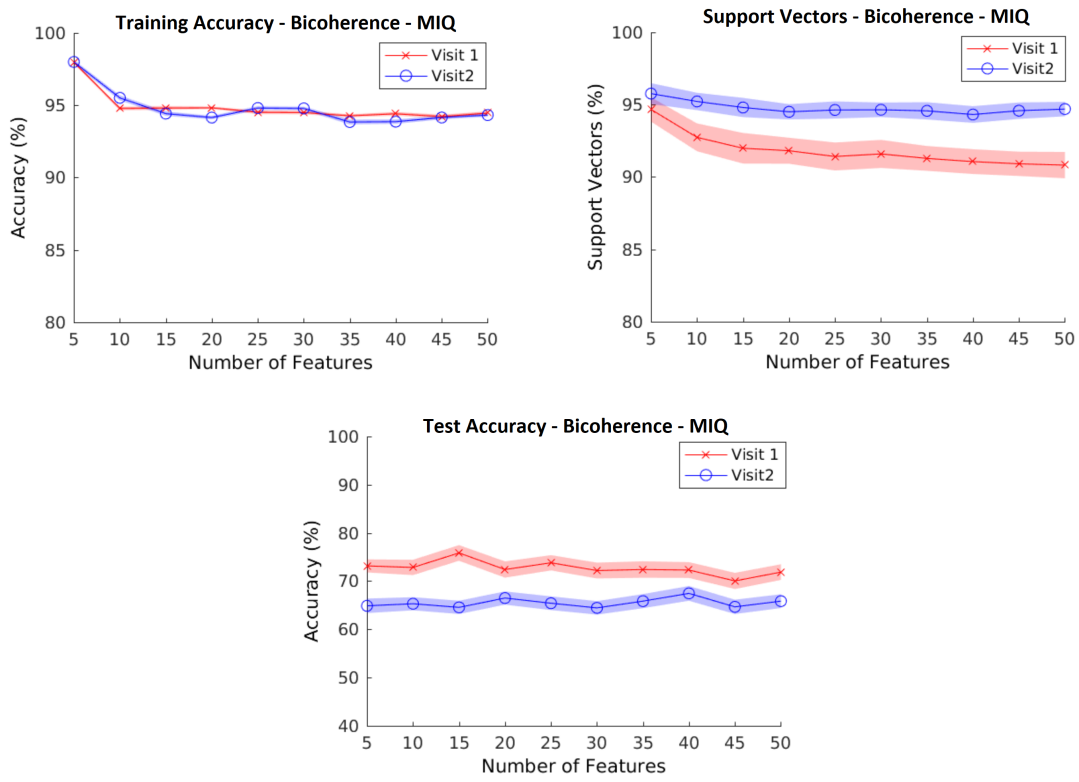

Figure 8.10: The training accuracy, test accuracy, and percentage of training samples used as support vectors for the best performing bicoherence feature set.

Table 8.8: Training accuracy, test accuracy, and percentage of training examples used as support vectors with respect to the KernelScale or gamma parameter

	Visit 1			Visit 2		
KernelScale	Mean Training Accuracy	Mean Testing Accuracy	Support Vectors	Mean Training Accuracy	Mean Testing Accuracy	Support Vectors
0.01	100.00	51.35	99.97	100.00	50.00	99.90
0.1	87.43	**73.14**	87.93	83.67	**66.75**	92.67
1	61.53	59.29	90.00	56.54	54.10	95.26
10	53.29	52.27	95.75	51.69	51.10	97.56
100	50.32	50.33	97.82	50.32	50.31	98.31

Table 8.9: Training accuracy, test accuracy, and percentage of training examples used as support vectors with respect to the BoxConstraint or C parameter

	Visit 1			Visit 2		
BoxConstraint	Mean Training Accuracy	Mean Testing Accuracy	Support Vectors	Mean Training Accuracy	Mean Testing Accuracy	Support Vectors
0.01	50.00	50.00	98.59	50.00	50.00	98.69
0.1	69.16	67.69	93.93	72.91	62.54	97.92
1	95.11	**72.37**	91.24	94.31	64.72	94.83
10	98.96	71.01	89.93	99.09	**65.60**	92.41
100	99.99	68.10	87.01	99.87	64.16	89.07

The largest test accuracy, 73.14%, occurs at gamma = 0.1 with a training accuracy of 87.43%. This is close to the original test accuracy from table 8.1 (i.e. 73.01%) with a relatively lower training accuracy in comparison (i.e. 94.6%). Table 8.9 shows that although C seems to be negatively correlated to the percentage of support vectors, a lower percentage of support vectors does not necessarily equate to higher test accuracy. These results show that a parameter search for SVM may provide a slight advantage in performance.

8.3.2 CNN Parameter Search

In terms of the CNN architecture, there are many more parameters involved in the search for the ideal parameters. Figure 8.11 shows the results of the test accuracy with respect to these parameters for a 1-layer CNN and Figure 8.12 shows the results of a 2-layer CNN architecture. All of these models were tested on the best performing CNN time segment (i.e. 2-seconds with zero overlap) using the best performing bicoherence feature set. Since this implementation of the CNN did not allow variable input lengths for the different length of the video clips, only the last n samples were taken, where n is the length of the shortest video. These figures show the results from visit 1 only and a 5-fold cross validation method was used to calculate accuracy.

For the 1-layer CNN, the best performing test accuracy was 65.4% which occurred at 32 filters of size 2 and 10 neurons in the dense layer. Using less than 10 neurons was found to decrease the test accuracy. For the 2-layer CNN, the best performing test accuracy was 64.46% which occurred at 64 filters of size 2 and 500 neurons in the dense layer. All the training accuracies were perfect or near perfect for the case of the 10-neuron dense layer. The addition of batch normalization did not improve the results of the best performing test case. Using the first n samples of the feature space as input also did not improve accuracy, and neither did using 2-second time segments with 50% overlap.

Comparing the SVM and CNN approach for the classification algorithm showed a clear advantage in performance for the SVM algorithm. Although the CNN was expected to improve accuracy due to its additional advantage of recognizing patterns between time segments of the input, it was not able to outperform the SVM.

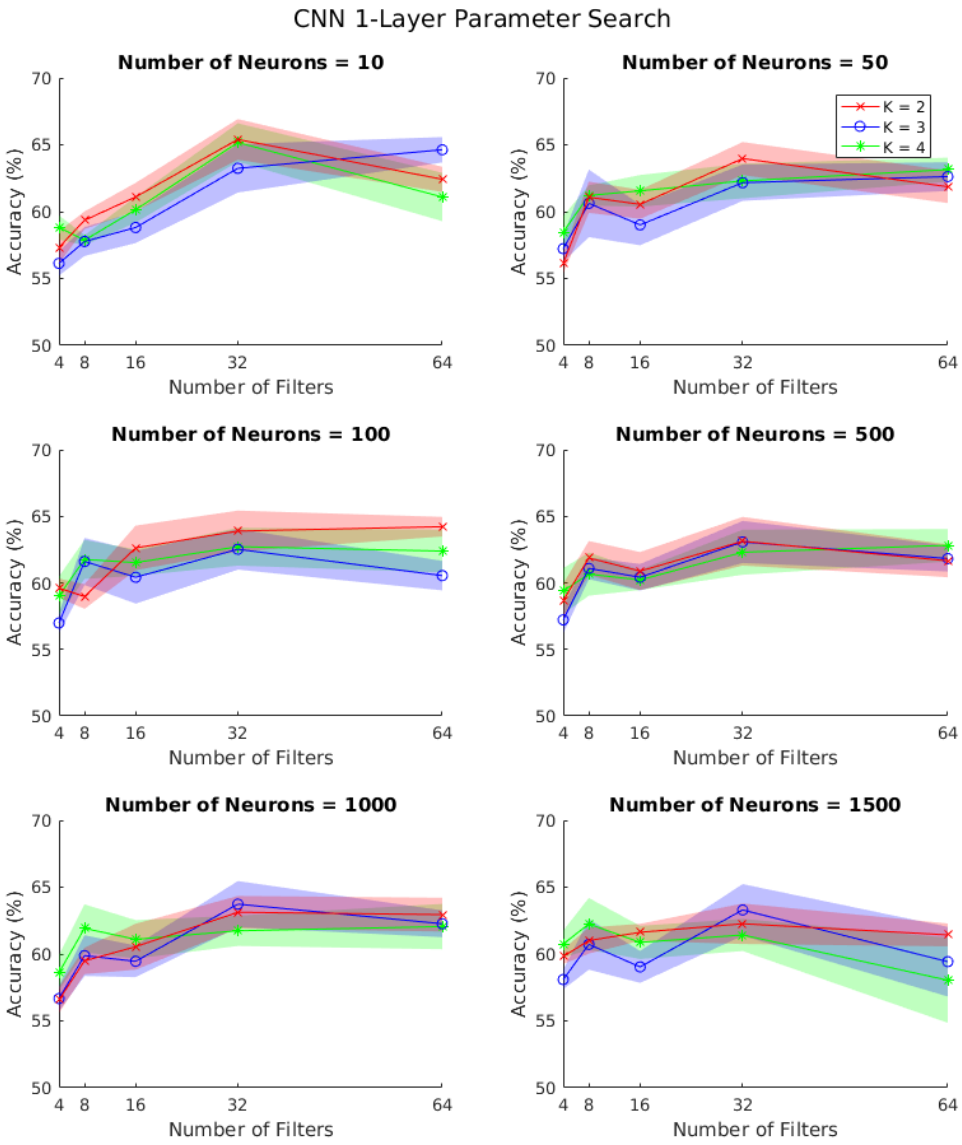

Figure 8.11: The test accuracy for the 1-Layer CNN model parameter search.

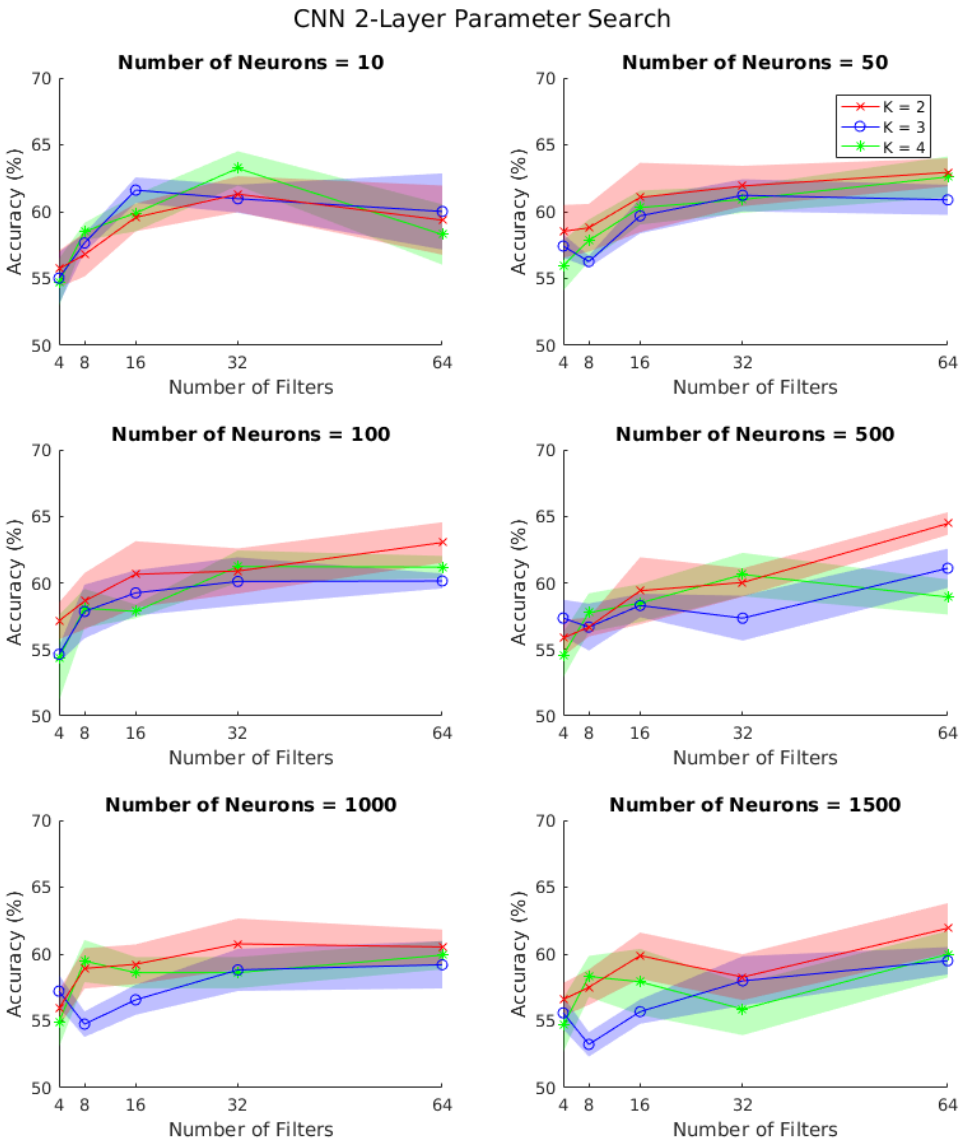

Figure 8.12: The test accuracy for the 2-Layer CNN model parameter search.

Table 8.10: Training and test accuracy for 2-way valence and arousal classification models

	Visit 1		Visit 2	
Emotion Dimension	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
Valence	89.44	**70.21**	90.06	58.47
Arousal	94.32	60.04	83.35	60.17

8.4 Valence and Arousal Model Analysis

Valence and arousal models are commonly used in emotion recognition models. Although this study focuses mainly on the classification of discrete emotions (i.e. fear and happiness), valence and arousal classification models are also tested to compare between the use of discrete and dimensional emotion models in emotion recognition. All subsequent tests are performed using the best parameters found above for feature set, time segment, and classification algorithm.

A separate model was generated for classification of valence and arousal using the all the bicoherence features and 10-fold cross validation. Table 8.10 and Table 8.11 show the results of 2-way and 3-way valence and arousal classification, respectively. The participants were asked how pleasant and arousing they found each video by rating them on a scale of 1 to 7 for valence and arousal. For the case of 2-way classification, ratings greater than 4 were taken as the high- valence or arousal class and ratings less than or equal to 4 were taken as the low- valence or arousal class. For the case of 3-way classification, ratings greater than 4 were taken as the high-valence or arousal class, ratings less than 4 were taken as low- valence or arousal, and ratings equal to 4 were rated as neutral. This grouping of the classes was found to provide the best classification accuracy for the valence and arousal models.

Table 8.11: Training and test accuracy for 3-way valence and arousal classification models

	Visit 1		Visit 2	
Emotion Dimension	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
Valence	39.11	**41.55**	33.50	35.22
Arousal	32.97	33.92	36.67	39.97

These results show worse performance for valence and arousal models as compared to the discrete emotion models for 2-way classification; however, the valence model was found to significantly outperform the arousal model for visit 1 but slightly underperform the arousal model on visit 2. The 3-way classification models were found to have worse performance overall, with some classification accuracies closer to chance accuracy (i.e. 33.33%). The 3-way valence model was also found to outperform the arousal model on visit 1 and slightly underperform the arousal model on visit 2. Training accuracy was also found to be reduced, especially for the 3-way classification of valence and arousal.

In addition, the models' performance was not improved when tested with the best 15 bicoherence features previously found for the discrete emotion models. Feature selection was also performed on the 2-way classification of valence and arousal independent of the discrete emotion model but was only found to increase the performance by about 1% for both models.

8.5 Classifiable Participant Analysis

In the creation of participant-independent models using LOSO cross validation, even the best performing model was found to perform near or below chance accuracy

for some participants for the classification between fearful and happy emotional responses. The number of participants that were classified above 60% accuracy were 98 and the number of participants that were classified at or below 60% accuracy were 18 for visit 1. In order to examine whether these results can be attributed to a difference in their emotional responses to the videos, the valence and arousal ratings were plotted for the above- and below- chance accuracy participants.

Figures 8.13 and 8.14 show the difference in ratings for the fearful and happy videos, respectively. These figures show that there are no significant differences between the average valence and arousal ratings for participants that were classified above and below 60%. The smaller range of the valence and arousal ratings is due the fact that only 18 of the 116 participants were poorly classified; however, there are no major differences in the valence and arousal ratings between the two subsets of participants. Although it is not clear why the models performed at or below chance accuracy for some participants, we can assume that this poor performance can be attributed to differences in brain activity rather than the emotional responses.

To examine this further, a separate model was created using only the data from participants that resulted in an accuracy of above 60% from the best performing bicoherence model with feature selection. This classifiable participant model was performed to examine whether training on the best performing subset of the participants can improve the test accuracy for those participants. Table 8.12 shows the results of the classifiable participant model and of the non-classifiable participant model for visit 1 and visit 2 as compared to the results of the model trained on all participants. There was an almost 3% increase in the test accuracy of the classifiable participant model in comparison to the best performing model. Although the overall

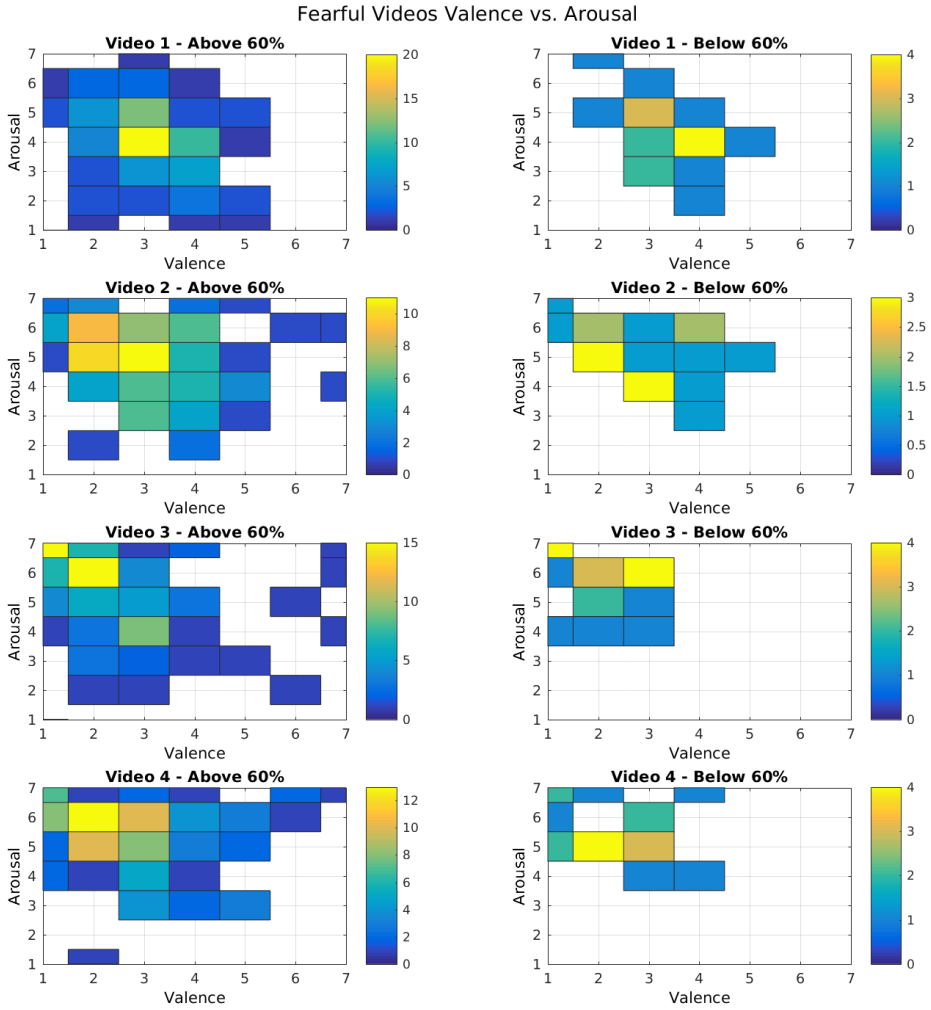

Figure 8.13: The valence and arousal ratings of fearful videos for classifiable and non-classifiable participants.

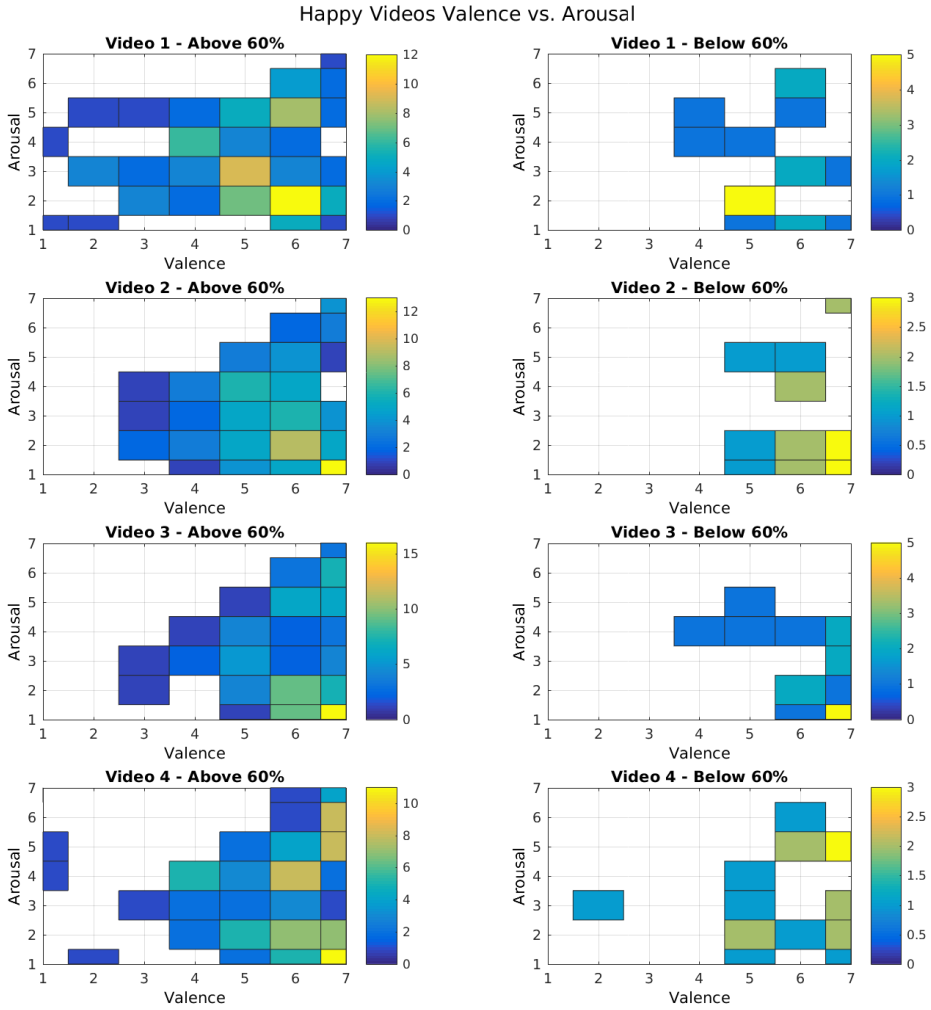

Figure 8.14: The valence and arousal ratings of happy videos for classifiable and non-classifiable participants.

Table 8.12: Training and test accuracy for models trained on classifiable participants, non-classifiable participants, and all participants

	Visit 1		Visit 2	
Model	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
Classifiable	95.87	**78.57**	97.17	72.12
Non-Classifiable	94.69	59.03	96.43	56.25
All Participants	94.82	75.86	94.43	67.46

accuracy of the model trained on the non-classifiable participants was improved when training only on the non-classifiable subset, this model continued to perform below 60% on the mean test accuracy.

8.6 Participant-Dependent Analysis

Although this study focuses mainly on participant-independent (i.e. between-participant) modelling, a short analysis was also conducted on participant-dependent (i.e. within-participant) modelling using the bicoherence features set and the best parameters identified from the participant-independent modelling analysis. For each participant, the cross-validation method used was to leave 2 videos out of the training set, one fearful video and one happy video, creating 16 possible models. The test accuracy was obtained by taking the average performance of all 16 models. Table 8.13 shows the results of the participant-dependent model in comparison to the participant-independent model.

These results show that participant-dependent modelling achieves almost the same accuracy as participant-independent modelling. Feature selection from the previous participant-independent analysis worsens the performance of participant-dependent

Table 8.13: Training and test accuracy for participant-dependent and participants-independent models

	Visit 1		Visit 2	
Model	Mean Training Accuracy	Mean Testing Accuracy	Mean Training Accuracy	Mean Testing Accuracy
Participant-Dependent	99.83	72.71	99.62	61.96
Participant-Independent	94.60	**73.01**	94.65	65.13

modelling since features are selected based off of other participants' features . Performing feature selection for each participant separately also did not improve test accuracy for participant-dependent modelling.

8.7 Between-Session Analysis

The final test conducted for emotion recognition was between-session modelling. This can be an especially difficult task due to the addition of possible variants between sessions and because the patterns identified in one visit may not generalize well to the other. The between-session models are trained on the data from the first visit and tested on the second visit. Both a participant-dependent and a participant-independent model was tested using the bicoherence feature set and the results are shown in Table 8.14.

These results show that the participant-independent model performs best for between session modelling. Further tests were conducted attempting to improve the participant-independent model accuracy by heavily weighting the test participants' own training data from the first visit; however, this did not improve test accuracy for

Table 8.14: Between-session training and test accuracy for participant-dependent and participant-independent models

Model	Mean Training Accuracy	Mean Testing Accuracy
Participant-Dependent	99.78	63.79
Participant-Independent	93.43	**67.13**

the between-session model.

Chapter 9

Discussion

This section provides a brief discussion of the results and compares our findings to previous studies. Our best performing participant-independent model was able to achieve about 76% classification accuracy between fearful and happy emotional responses. This general model classification accuracy was achieved after several investigations of feature extraction, feature selection, time segment analysis, and classification algorithm parameter search. This performance is very similar to the performance of the most similar previous study (Dhindsa and Becker, 2017), which obtained an average of 75% classification accuracy.

Overall, the use of feature selection did not seem to drastically improve test set accuracy. Feature selection methods are commonly used in emotion recognition studies as an automatic method for selecting the best features out of a large dimensional feature set. However, these results show that this may not be the best method for selecting features. It may be worthwhile to compare the accuracies of separate feature extraction methods or combinations of these methods prior to the use of feature selection. Using feature selection on the entire feature set did not choose the best

performing bicoherence features as the most relevant and least redundant features. Since this feature selection method is blind to the actual output accuracy and is only based on a priori computations of mutual information, it does not guarantee improved accuracy over manual feature selection.

One interesting finding through the exploration of feature extraction methods is the superior performance of the models trained on the bicoherence features. This suggests that higher order spectral features may be worth exploring in future studies and could become one of the more commonly used methods in emotion recognition. Further exploration of input parameters to these HOSA features may be explored in future studies to determine the optimal bicoherence parameters in EEG-based emotion recognition.

Although there are some exceptions, it was found that performance tends to be worse on the models trained on the data from the second visit when compared to the first visit throughout the entire study. This suggests that previous exposure to a stimulus may contribute to difficulties with emotional classification. Emotion cues may be less salient in the second visit as previous exposure to a stimulus can result in some adaptation, making it more difficult for the model to classify the elicited emotion.

The results of the time segment analysis show that the choice of EEG window length can significantly affect the accuracy of the model. The best time segment will depend on the type of classification algorithm used since the CNN seemed to require a time segment that was 10 times shorter than that of SVM. Longer time segments may be required for SVM in order to capture the emotional response within each segment, since they are all regarded as independent training samples. The CNN may benefit

from shorter time segments as more samples means that a higher resolution pattern can be found across the video in time, since the time segments are not independent.

Comparing the SVM and CNN approach for the classification algorithm showed a clear advantage in performance for the SVM algorithm. Although the CNN was expected to improve accuracy due to its additional advantage of recognizing patterns between time segments of the input, it was not able to outperform the SVM. Furthermore, while the SVM parameter search was not able to improve results, changing the values of the parameters of the CNN had a large impact on performance.

Using the valence and arousal emotion models for 2-way classification was found to produce an accuracy of about 70% for valence and 60% for arousal. These results suggest that a participants' rating for pleasantness (i.e. valence) may be easier to classify than their perceived intensity of emotion (i.e. arousal). Although a participant may rate a fearful video as pleasant (i.e. high-valence), it may be easier to classify their EEG response as a fearful response rather than a pleasant response, thus explaining why there may be a disparity between discrete and dimensional emotion classification accuracy. Although this model underperformed the discrete emotion model of classification between fearful and happy emotional responses, it was able to outperform models reported in similar studies for valence accuracy (i.e. 61.17% for 2-way valence classification) (Kumar *et al.*, 2016). Unlike this previous study, our valence model outperformed the arousal model. However, one study using the DEAP dataset was able to achieve about 74% and 73% for 2-way valence and arousal classification, respectively(Atkinson and Campos, 2016), while another study was able to successfully perform 3-way classification models for valence and arousal (Soleymani *et al.*, 2012). Both these studies suggest that our results may be improved by using

more EEG channels or other feature extraction methods such as fractal dimension.

Creating a model using the classifiable participants was found to increase the overall performance to about 79%. This shows that choosing a subset of the participants can increase the overall accuracy of the model and agrees with the findings of previous work that studied classifiable participant models (Dhindsa and Becker, 2017). Although the non-classifiable participant model was able to improve the overall accuracy for those participants, it still performed poorly overall and did not result in the majority of participants becoming classifiable when trained on this subset only. This suggests that there is further variability between the non-classifiable participants that is causing poor performance beyond the variability between the non-classifiable and classifiable participants. Further work may be done to determine why certain participants were non-classifiable and how to improve the model's performance in this situation.

The participant-dependent models did not perform better than the participant-independent models as expected but had a similar accuracy of about 73%. However, the parameters used (i.e. feature set, time segment) were not optimized for the participant-dependent case, and therefore performance may be increased with further analysis. Feature selection did not help improve the accuracy even though feature space dimensionality reduction is especially important with a decreased number of training examples. Similar to the participant-independent model, some participants were found to be non-classifiable and performed below chance accuracy when trained on their own data.

Finally, the between-session model was able to obtain an accuracy of about 67% using participant-independent modelling, which outperformed the participant-dependent

between-session model. This model was expected to underperform previous models due to the effects of familiarity to the stimuli (Thammasan *et al.*, 2017) but showed that it was still able to predict at above chance accuracy. However, this model underperformed in comparison to other studies that examined between session modelling (Zheng *et al.*, 2016). This disparity may be due to the use of many more EEG channels (i.e. 62 channels) or the use of fewer participants (i.e. 15 participants), but it does suggest that improved accuracy may be achieved using other feature extraction methods (e.g. differential entropy), classification algorithms (e.g. graph-regularized extreme learning machine), or feature smoothing (e.g. linear dynamical smoothing).

Chapter 10

Conclusion and Future Outlook

In this study, we have systematically evaluated the performance of various commonly used feature extraction methods, feature selection methods, and pattern classification methods for participant-independent emotion recognition on a large dataset of 116 participants. From our experimental results, we have found that a model trained using bicoherence features with feature selection and an SVM classifier outperforms other methods and the ideal time segment length of input to the classifier is 20 seconds. The best average classification accuracy of 75.86% was obtained with LOSO cross-validation for determining between fearful and happy emotional responses which outperformed the valence and arousal emotion models. In addition, we found that it is possible to obtain an increased classification accuracy of 78.57% when training on the subset classifiable participants. Further analysis is required to determine why certain participants were non-classifiable and how to increase performance on this subset of individuals. Moreover, we were able to achieve an accuracy of 72.71% and 67.13% for participant-dependent and between-session models, respectively.

We found that overall performance was similar to comparable previous work.

Studies that outperformed our model were found to either have included 32 or more EEG channels or consisted of about one forth the number of participants or fewer. Also, future work may focus on finding the minimum number of participants required for adequate generalization using a large dataset (Kambeitz *et al.*, 2018). Some studies were able to achieve improved accuracy with the addition of multimodal physiological signals such as ECG, EOG, EMG, GSR, etc. (Soleymani *et al.*, 2015). Although there exist some EEG emotion recognition datasets for comparison of classification methods (e.g. DEAP dataset), there does not exist a database of videos for emotion elicitation as there does for images and sounds for the purpose of more direct comparison between studies. This problem may be addressed in future studies with the use of the novel One-Minute Gradual Emotional Behaviour dataset (OMG-Emotion dataset) which consists of one-minute videos designed to elicit several emotions and were rated on valence and arousal by several participants across various points in the videos, allowing for possible applications of real-time tracking of emotional response (Barros *et al.*, 2018). Finally, other potential improvement of classification accuracy on this dataset may be achieved with the addition of better artifact rejection methods (e.g. filter-bank artifact rejection), various feature extraction methods (e.g. fractal dimension or differential entropy), feature smoothing (e.g. linear dynamical systems), or other classification methods (e.g. graph-regularized extreme learning machine).

Bibliography

Aftanas, L. I., Varlamov, A. A., Pavlov, S. V., Makhnev, V. P., and Reva, N. V. (2002). Time-dependent cortical asymmetries induced by emotional arousal: EEG analysis of event-related synchronization and desynchronization in individually defined frequency bands. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, **44**(1), 67–82.

Atkinson, J. and Campos, D. (2016). Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Systems with Applications*, **47**, 35–41. Publisher: Elsevier Ltd ISBN: 09574174.

Başar, E., Schürmann, M., and Sakowitz, O. (2001). The selectively distributed theta system: functions.

Balconi, M. and Lucchiari, C. (2006). EEG correlates (event-related desynchronization) of emotional face elaboration: A temporal analysis. *Neuroscience Letters*, **392**(1-2), 118–123. ISBN: 0304-3940 (Print)\r0304-3940 (Linking).

Balconi, M. and Pozzoli, U. (2009). Arousal effect on emotional face comprehension: Frequency band changes in different time intervals. *Physiology & Behavior*, **97**(3), 455–462.

Barrett, L. F., Mesquita, B., Ochsner, K. N., and Gross, J. J. (2007). The Experience of Emotion. *Annual Review of Psychology*, **58**(1), 373–403. ISBN: 0066-4308.

Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., and Wermter, S. (2018). The OMG-Emotion Behavior Dataset. arXiv: 1803.05434.

Bos, D. P.-O., Reuderink, B., Laar, B. v. d., Grkk, H., Mhl, C., Poel, M., Nijholt, A., and Heylen, D. (2010). Brain-Computer Interfacing and Games. In *Brain-Computer Interfaces*, Human-Computer Interaction Series, pages 149–178. Springer, London.

Bradley, M. and Lang, P. (1999). International affective digitized sounds (IADS): stimuli, instruction manual and affective ratings.

Cecotti, H. and Gräser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(3), 433–445. ISBN: 1939-3539 (Electronic)\r0098-5589 (Linking).

Chanel, G., Kronegg, J., Grandjean, D., and Pun, T. (2006). Emotion Assessment: Arousal Evaluation Using EEGs and Peripheral Physiological Signals. pages 530–537. ISBN: 3540393927.

Chanel, G., Kierkels, J. J., Soleymani, M., and Pun, T. (2009). Short-term emotion assessment in a recall paradigm. *International Journal of Human Computer Studies*, **67**(8), 607–627. ISBN: 1071-5819.

Choppin, A. (2000). EEG-Based Human Interface for Disabled Individuals : Emotion Expression with Neural Networks Submitted for the Master Degree. *Emotion*.

Cohen, M. X. (2014). *Analyzing Neural Time Series Data*. The MIT Press.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, **18**(1), 32–80. ISBN: 1053-5888.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA.

Delorme, A., Mullen, T., Kothe, C., Akalin Acar, Z., Bigdely-Shamlo, N., Vankov, A., and Makeig, S. (2011). EEGLAB, SIFT, NFT, BCILAB, and ERICA: New Tools for Advanced EEG Processing. *Computational Intelligence and Neuroscience*.

Dhindsa, K. and Becker, S. (2017). Emotional reaction recognition from EEG. *2017 International Workshop on Pattern Recognition in Neuroimaging, PRNI 2017*. Citation Key: Dhindsa2017 ISBN: 9781538631591.

Ding, C. and Peng, H. (2005). Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, **3**(2), 523–528. Citation Key: Ding2005 ISBN: 0-7695-2000-6.

Duan, R.-n., Wang, X.-W., and Lu, B.-L. (2012). EEG-Based Emotion Recognition in Listening Music by Using Support Vector Machine and Linear Dynamic System. *Neural Information Processing*, pages 468–475. Citation Key: Duan2012.

Ekman, P. (1992). Are There Basic Emotions? *Psychological Review*, **99**(3), 550–553. arXiv: 1011.1669v3 ISBN: 1939-1471 (Electronic); 0033-295X (Print).

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, **18**(12), 1050–1057.

Frantzidis, C. A., Bratsas, C., Papadelis, C. L., Konstantinidis, E., Pappas, C., and Bamidis, P. D. (2010). Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE Transactions on Information Technology in Biomedicine*, **14**(3), 589–597. ISBN: 1089-7771.

Glassman, E. L. and Member, S. (2005). A Wavelet-Like Filter Based on Neuron-Action Potentials for Analysis of HumanScalp Electroencephalographs. **52**(11), 1–12.

Glauser, E. S. D. and Scherer, K. R. (2008). Neuronal Processes Involved in Subjective Feeling Emergence: Oscillatory Activity During an Emotional Monitoring Task. *Brain Topography*, **20**(4), 224–231.

Goncharova, I. I., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2003). EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, **114**(9), 1580–1593.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Grandjean, D. and Scherer, K. R. (2008). Unpacking the cognitive architecture of emotion processes. *Emotion (Washington, D.C.)*, **8**(3), 341–351.

Halterman, M. W. (2005). Neuroscience, 3rd Edition. *Neurology*, **64**(4), 769.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: Controversies and consensus. *Trends in Cognitive Sciences*, **16**(9), 458. ISBN: 1879-307X (Electronic)\n1364-6613 (Linking).

Hammond, D. C. (2005). Neurofeedback with anxiety and affective disorders. *Child and Adolescent Psychiatric Clinics of North America*, **14**(1 SPEC.ISS.), 105–123. ISBN: 1056-4993 (Print)\n1056-4993 (Linking).

Harmon-Jones, E. (2003). Clarifying the emotive functions of asymmetrical frontal cortical activity. *Psychophysiology*, **40**(6), 838–848.

Harmon-Jones, E., Gable, P. A., and Peterson, C. K. (2010). The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update. *Biological Psychology*, **84**(3), 451–462.

Hidalgo-Muñoz, A. R., Lopez, M. M., Pereira, A. T., Santos, I. M., and Tom, A. M. (2013). Spectral turbulence measuring as feature extraction method from EEG on affective computing. *Biomedical Signal Processing and Control*, **8**(6), 945–950. ISBN: 17468094 (ISSN).

Hosseini, S. A. and Naghibi-Sistani, M. B. (2011). Classification of emotional stress using brain activity. *Applied Biomedical Engineering*, (2004), 313–336. ISBN: 9789537619824.

Hua, J., Xiong, Z., Lowey, J., Suh, E., and Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics (Oxford, England)*, **21**(8), 1509–1515.

J Lang, P., M Bradley, M., and N & Cuthbert, B. (2008). International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual (Rep. No. A-8). *Technical Report A-8*.

Jirayucharoensak, S., Pan-Ngum, S., Israsena, P., Jirayucharoensak, S., Pan-Ngum, S., and Israsena, P. (2014). EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation, EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation. **2014, 2014**, e627892.

Kambeitz, J., Cabral, C., Sacchet, M. D., Gotlib, I. H., Zahn, R., Serpa, M. H., Walter, M., Falkai, P., and Koutsouleris, N. (2018). Reply to: Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biological Psychiatry*, pages 2017–2018. Publisher: Society of Biological Psychiatry.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and Emotion*, **5**(3), 263–291. ISBN: ISSN˜˜0146-7239.

Knyazev, G. G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience and Biobehavioral Reviews*, **36**(1), 677–695. ISBN: 1873-7528 (Electronic)\n0149-7634 (Linking).

Koelstra, S., Mhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T.,

Nijholt, A., and Patras, I. (2012). DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing*, **3**(1), 18–31. ISBN: 1949-3045 VO - 3.

Kumar, N., Khaund, K., and Hazarika, S. M. (2016). Bispectral Analysis of EEG for Emotion Recognition. *Procedia Computer Science*, **84**, 31–35. Publisher: Elsevier Masson SAS Citation Key: Kumar2016.

Lan, Z., Sourina, O., Wang, L., and Liu, Y. (2016). Real-time EEG-based emotion monitoring using stable features. *Visual Computer*, **32**(3), 347–358. Publisher: Springer Berlin Heidelberg.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**(7553), 436–444. arXiv: 1312.6184v5 ISBN: 9780521835688.

LeDoux, J. E. (1996). *The emotional brain: the mysterious underpinnings of emotional life*. Simon & Schuster, New York.

LeDoux, J. E. (2000). Emotion Circuits in the Brain. pages 155–184.

Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., and Chen, J. H. (2010). EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, **57**(7), 1798–1806. ISBN: 0018-9294 VO - 57.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique, Second Edition*. The MIT Press.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors (1983). *Machine*

Learning: An Artificial Intelligence Approach. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, **1**(2), 66–84. Publisher: Taylor & Francis Citation Key: Muhl2014 ISBN: 2326-263X.

Murugappan, M., Juhari, M. R. B. M., Nagarajan, R., and Yaacob, S. (2009). An investigation on visual and audiovisual stimulus based emotion recognition using EEG. *International Journal of Medical Engineering and Informatics*, **1**(3), 342. Citation Key: Murugappan2009.

Murugappan, M., Ramachandran, N., and Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, **03**(04), 390–396.

Neuhaus, A. H. and Popescu, F. C. (2018). Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biological Psychiatry*, **0**(0).

Nie, D., Wang, X.-W., Shi, L.-C., and Lu, B.-L. (2011). EEG-based Emotion Recognition during Watching Movies. pages 667–670. ISBN: 9781424441419.

Olofsson, J. K., Nordin, S., Sequeira, H., and Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, **77**(3), 247–265. arXiv: NIHMS150003 ISBN: 0301-0511 (Print)\r0301-0511 (Linking).

Onton, J. (2009). High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in Human Neuroscience*, **3**(December), 1–18. ISBN: 1662-5161 (Electronic)\r1662-5161 (Linking).

Panksepp, J. (1998). *Affective Neuroscience: The foundations of human and animal emotions*. Oxford University Press, New York.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **27**(8), 1226–1238. arXiv: f Citation Key: Peng2005 ISBN: 0162-8828.

Petrantonakis, P. C. (2010). Emotion recognition from EEG using higher order crossings. *Technology*, **14**(2), 186–197. ISBN: 1089-7771 VO - 14.

Petrantonakis, P. C. and Hadjileontiadis, L. J. (2011). 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. ISBN: 1089-7771 VO - PP.

Plutchik, R. (1980). *Emotion : a psychoevolutionary synthesis*. New York : Harper & Row.

Reuderink, B., Mhl, C., and Poel, M. (2013). Valence, arousal and dominance in the EEG during game play. *International Journal of Autonomous and Adaptive Communications Systems*, **6**(1), 45.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, **39**(6), 1161–1178.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, **76**(5), 805–819.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, **38**(11), 5391–5420. arXiv: 1703.05051 ISBN: 978-1-5386-4873-5.

Sigl, J. C. and Chamoun, N. G. (1994). An introduction to bispectral analysis for the electroencephalogram. *Journal of Clinical Monitoring*, **10**(6), 392–404. Citation Key: Sigl1994 ISBN: 0748-1977 (Print)$\$r0748-1977 (Linking).

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, **3**(1), 42–55. ISBN: 1949-3045.

Soleymani, M., Pantic, M., and Pun, T. (2015). Multimodal emotion recognition in response to videos (Extended abstract). *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, **3**(2), 491–497. Citation Key: Soleymani2015 ISBN: 9781479999538.

Swami, A., Mendel, J. M., and Nikias, C. L. (1993). *Higher-Order Spectral Analysis Toolbox*.

Takahashi, K. (2004). Remarks on SVM-based emotion recognition from multi-modal bio-potential signals. *RO-MAN 2004. 13th IEEE International Workshop on Robot*

and Human Interactive Communication (IEEE Catalog No.04TH8759), pages 95–100. ISBN: 0-7803-8570-5.

Thammasan, N., Moriyama, K., Fukui, K. i., and Numao, M. (2017). Familiarity effects in EEG-based emotion recognition. *Brain Informatics*, **4**(1), 39–50. arXiv: 1611.10120v1 Publisher: Springer Berlin Heidelberg ISBN: 2198-4018 (Print)\r2198-4026.

Van Den Broek, E. L. (2012). Affective computing: A reverence for a century of research. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **7403 LNCS**, 434–448. ISBN: 9783642345838.

Wang, X., Nie, D., and Lu, B. (2011). EEG-based emotion recognition using frequency domain features and support vector machines. *International Conference on Neural Information Processing*, pages 734–743. ISBN: 9783642249549.

Westermann, R., Spies, K., Stahl, G., and Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: a meta-analysis. *European Journal of Social Psychology*, **26**(4), 557–580.

Wolpaw, J. R. (2013). *Brain-computer interfaces*, volume 110. arXiv: 1011.1669v3 Publication Title: Handbook of clinical neurology Issue: Journal Article PG - 67-74 ISSN: 00189162.

Xu, H. and Plataniotis, K. N. (2016). Affective states classification using EEG and semi-supervised deep learning approaches. *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. ISBN: 978-1-5090-3724-7.

Zheng, W.-L., Zhu, J.-Y., Peng, Y., and Lu, B.-L. (2014). EEG-Based Emotion Classification Using Deep Belief Networks. *Multimedia and Expo (ICME)*, pages 1–6. ISBN: 978-1-4799-4761-4.

Zheng, W.-L., Zhu, J.-Y., and Lu, B.-L. (2016). Identifying Stable Patterns over Time for Emotion Recognition from EEG. pages 1–15. arXiv: 1601.02197 Citation Key: Zheng2016.

Zhou, F., Qu, X., Jiao, J., and Helander, M. G. (2014). Emotion prediction from physiological signals: A comparison study between visual and auditory elicitors. *Interacting with Computers*, **26**(3), 285–302. ISBN: 1111111111.