AFFECT RECOGNITION USING ELECTROENCEPHALOGRAPHY FEATURES

A Thesis

by

SUSHIRDEEP NARAYANA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Peng Li
Committee Members,	Gwan Choi
	Yoonsuck Choe
	Xiaoning Qian
Head of Department,	Miroslav M. Begovic

May 2017

Major Subject: Electrical Engineering

ABSTRACT

Affect is the psychological display of emotion often described with three principal dimensions: 1) valence 2) arousal and 3) dominance. This thesis work explores the ability of computers to recognize human emotions using Electroencephalography (EEG) features. The development of computer systems to classify human emotions using physiological signals has recently gained pace in the research and technological community. This is because by using EEG to analyze the cognitive state one will be able to establish a direct communication channel between a computer and the human brain. Other applications of recognizing the affective states from EEG include identifying stress and cognitive workload on individuals and assist them in relaxation.

This thesis is an extensive study on the design of paradigms that help computer systems recognize emotional states given a multichannel Electroencephalogram (EEG) segment. The process of first extracting features from the EEG signals using signal processing and then constructing a predictive model via machine learning is often referred to as paradigms. In this work, we will first present a brief review of the state-of-the-art paradigms that have contributed to the topic of emotional affect recognition. Then the proposed paradigms to recognize the principal dimensions of affect are detailed. Feature selection is also performed in order to select the relevant features. The evaluation of the models created to predict the affective states will be performed quantitatively by calculating the generalization accuracy and qualitatively by interpreting them.

DEDICATION

To my parents – for their unwavering support, affection and assistance.

ACKNOWLEDGEMENTS

I am grateful to my research supervisor, Dr. Peng Li, for his constant and thoughtful guidance, encouragement and mentorship throughout my research. I am also most thankful to Dr. Peng Li for providing all the facilities and supporting me to complete this project. I sincerely thank all the members of my Thesis Committee, Dr. Gwan Choi, Dr. Yoonsuck Choe, Dr. Xiaoning Qian and Dr. Sung Il Park for their help, contribution and encouragement. I am also thankful to Dr. Xiaoyu Xu and Dr. Song Qiu for their support to complete this project.

I would also like to thank all my colleagues in the research group of Dr. Peng Li for their support and friendship. Finally, I would like to thank Dr. J. Silva-Martinez for his encouragement and assistance in providing the administrative support during the M.S degree course.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. Peng Li, Dr. Gwan Choi and Dr. Xiaoning Qian of the Department of Electrical and Computer Engineering and Dr. Yoonsuck Choe of the Department of Computer Science and Engineering of Texas A&M University, College Station.

All the work conducted for this thesis was completed by the student independently.

Funding Sources

This work was made possible in part by the Graduate Merit Scholarship given by the Department of Electrical and Computer Engineering of Texas A&M University, College Station.

NOMENCLATURE

BCI	Brain- Computer Interface
EEG	Electroencephalography
AC	Alternating Current
PSD	Power Spectral Density
DWT	Discrete Wavelet Transform
CSP	Common Spatial Patterns
SVM	Support Vector Machines
DEAP	Database for Emotional Analysis using Physiological signals
mRMR	Maximum Relevancy and Minimum Redundancy

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION AND BACKGROUND

A system which takes neural signals measured from a person to predict some abstract aspect of the person's cognitive state is a Brain-Computer Interface (BCI). BCIs enable us to investigate brain activity as an independent variable. The aspects of the human brain state that can be predicted include motor functions, alertness, emotions, task involvement, etc. BCIs provide an additional channel that allow neuronal activity of the brain to interact with computer systems. Brain activity can be recorded intra-cortically with multielectrode arrays or single electrode or subdurally from the cortex or from the scalp. The type of recording employed classifies the BCI [30] into two types as described below: -

(1) Invasive BCI: In this approach, the brain activity within the cortex and activity from the surface of the cortex is measured using surgical implants. Examples include Electrocardiogram (ECoG) and microarrays and neural chips placed on the spinal cord.

(2) Non-Invasive BCI: In this method, the activity of the brain is recorded using noninvasive techniques. Examples based on this approach include Electroencephalography (EEG), Magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI).

Since the BCIs provide an additional communication channel between the brain and a computer system they have the potential to bridge the gap of Human-Computer Interactions. Figure 1 shows a BCI system setup. Many of the applications of BCIs can be divided into two varieties which are clinical applications and non-clinical applications. The clinical applications of BCI [30] include constructing systems with the help of BCI for motor control and other functions that are required for patients. Patients who have lost most of their voluntary muscular movement, suffer from motor neuron disease, spinal cord injury or traumatic brain injury which may lead to severe motor paralysis. Depending on the state of the patient one can construct a BCI that restores some of their motor functions. An example of a popular clinical BCI is the P300 Speller BCI based on the odd-ball paradigm. The P300 [30] wave of Electroencephalography signal is central to the idea of a Memory and Encoding Related Multifaceted Electroencephalographic Response (MERMER) developed by Dr. Lawrence Farwell. In the P300 [30] Speller BCI, a 6x6 matrix of characters and commands are displayed to the user and the subject must focus his/her attention on a specific character as the rows and columns are flashed in a random order. The character is identified by the computer by recognizing the row and the column of the matrix the user is focusing on by analyzing the subject's P300 [30] signal response.

There are several non-clinical applications of BCI [30] as well. The video gaming industry involving biofeedback games is one such application. Neurorehabilitation which deals with cognitive workload monitoring is another example where BCIs are used to analyze the amount of stress an individual is experiencing. BCIs can also be used in the Forensics

department as lie detection monitors to evaluate trust. Many researchers use BCIs to gather information content that can help them answer some of the open problems about neuroscience that have not yet been solved.

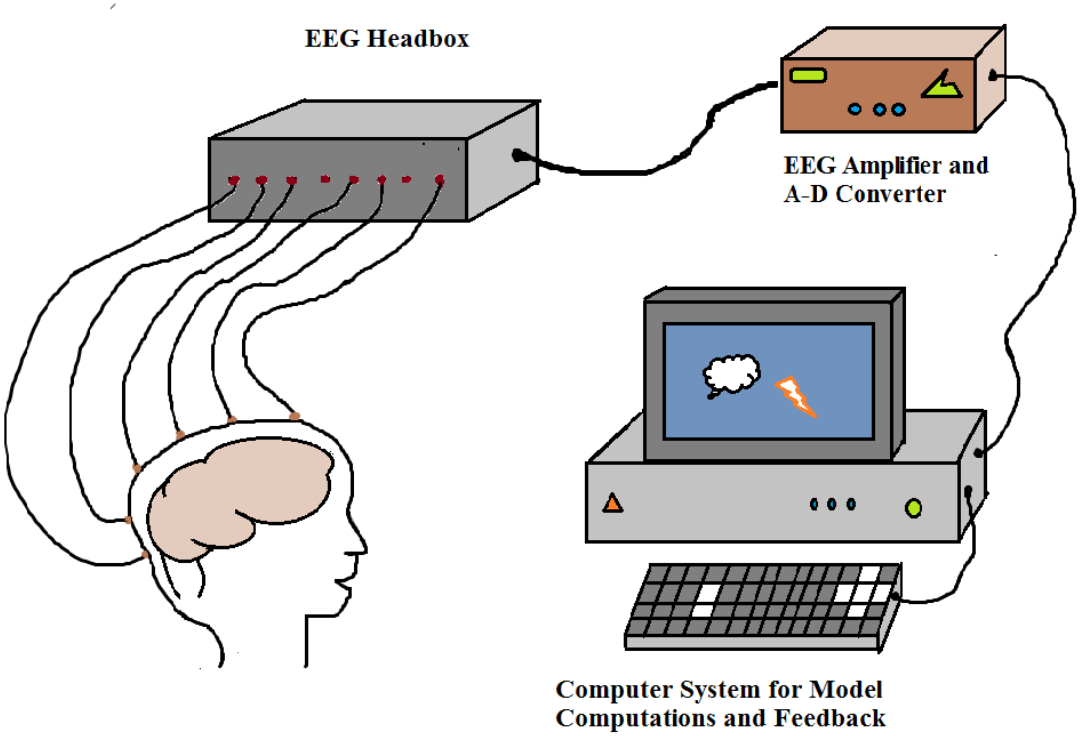

Figure 1: Brain-Computer Interface System Setup

There are several challenges involved in data processing and recognition from BCI applications. There are lot of variabilities encountered when processing EEG signals acquired from BCIs. A portion of this variability is caused by the difference in sensor locations when collecting the signals across different sessions. The low Signal-to-Noise ratio of EEG signals recorded makes the relevant neural activity of the brain very small

compared to the interfering artifacts and noise. Also, building a robust and useful BCI model via machine learning when the brain dynamics are non-stationary at all time scales is a major challenge. This thesis work will help outline some promising directions to address these hardships.

In this thesis, sophisticated signal processing and machine learning techniques are leveraged to address the challenges encountered on emotional affect recognition with EEG based BCIs. The process of first extracting features from the EEG signals using signal processing transforms and filters and then constructing a predictive model via machine learning is known as paradigms. The current chapter gives an introduction and a brief background on the topic. Chapter II discusses the signal processing techniques applied to extract features from EEG signals acquired from BCIs. Details about the calculations of Power Spectral Density, Discrete Wavelet Transforms and computations of spatial filters will be provided. Later in Chapter II, the Support Vector Machines (SVM) algorithm will be discussed. The SVMs are considered to be one of the state-of-the art Supervised Machine Learning algorithms. They are formed by combining the kernel trick with a modified loss function. Chapter III discusses generation of semi-simulated EEG signal data using modeled artifact signals and noise. In Chapter IV we will examine the processes of Feature Selection method employed to eliminate the redundant features and clearly identify the relevant features present in the SVM model. Chapter V gives a description of the experiments performed with the open source datasets and the results obtained from the designed paradigms. Inferences from these experiments and results will be discussed to

get a qualitative interpretation of the SVM model computed. Chapter VI gives a brief conclusion to the thesis work.

The organization of the remaining of this chapter is described below. Section 1 discusses about the different techniques used for recording brain activity. Section 2 describes about the basics of Electroencephalography (EEG) recordings. In Section 3 an introduction to Affective Computing is provided.

1.1 Recording Brain Activity

The primary motivation for recording brain activity [30] is to study neural processes in the normal working brain. Over the past few years, functional brain imaging has developed into a multidisciplinary research field encompassing techniques that help better understand the processes that underlie normal and pathological brain function. In this section, we will be covering mostly on non-invasive techniques of recording brain activity.

In Positron Emission Tomography (PET) [30] radioactive labeled organic molecules are used to capture the dynamic changes of the spatial distribution. The PET images have spatial resolution as high as 2mm, but the temporal resolution is limited by the dynamics of the process being studied which can last for several minutes. In the past few years, it has been possible to use blood oxygen level dependent (BOLD) response as an input signal to a BCI using functional Magnetic Resonance Imaging (fMRI). fMRI [30] can be

performed using a standard 1.5 T clinical MRI magnet. However, many studies nowadays, use higher field machines around 3 – 4.5 T machines for improved Signal-to-Noise Ratio (SNR) and spatial resolution. fMRI studies can obtain spatial resolutions as high as 1 mm - 3 mm, but their temporal resolution is relatively low, in the order of 2 s- 4 s. It should be noted that the regions representing BOLD changes in fMRI images do not hold a one-to-one relationship with the electrical neural activity regions of the brain. The BOLD signal of fMRI is essentially a qualitative signal because its dependence with brain activity is a very complex dependence with changes of blood flow and oxygen mechanism. The EEG and Magnetoencephalography (MEG) [30] recoding techniques possess a superior temporal resolution of brain activity when compared to other non-invasive techniques of recording like PET and fMRI. The characteristic magnetic induction produced by the neural currents of the brain is weak, on the order of femtoTeslas, which necessitates sophisticated sensing technology and a very expensive setup to measure MEG signals. Acquiring signals through Electroencephalography (EEG) is comparatively easier and cheaper. The next section gives a detailed description of EEG.

1.2 Basics of Electroencephalography

Electroencephalography (EEG) is the measurement of a set of electric potential differences between pairs of electrodes attached to the scalp. The sensors may be either glued to the skin at selected locations using a water based gel or fitted in an elastic cap for a quick attachment and a uniform coverage of the entire scalp. Researchers use 32 to 256

electrodes. The first recording of EEG signals of the brain was performed by a German physician Hans Berger in 1924. Today's technological developments have made the EEG the most widely known non-invasive brain imaging modality. The benefits of using EEG as a form of recording brain activity include its portability, affordability and a straightforward setup. EEG has been very successful in detecting epileptic seizures as the seizures are characterized by unusual electric activity in epileptogenic regions of the brain.

EEG signals recorded can only detect large scale neural dynamics. In other words, EEG signals are actively recorded when around 50, 000 to millions of neurons are firing in near synchrony. When the large number of neurons mentioned above fire in near synchrony the electromagnetic fields of the co-aligned neurons add up resulting in a detectable EEG signal. The events where the synchronized firing can occur are given below:

a) An external event from the surroundings and the environment of the subject generates a cascade of neural processes. Examples of these types of event triggers include watching a video, listening to music, etc.
b) An internal event due to the subjects thinking triggers a series of neural processes. Examples of these types of internal event triggers include recollection of previous incidents and memories.
c) Several neural populations enter a synchronized steady state firing pattern. This phenomenon is also called idle oscillations.

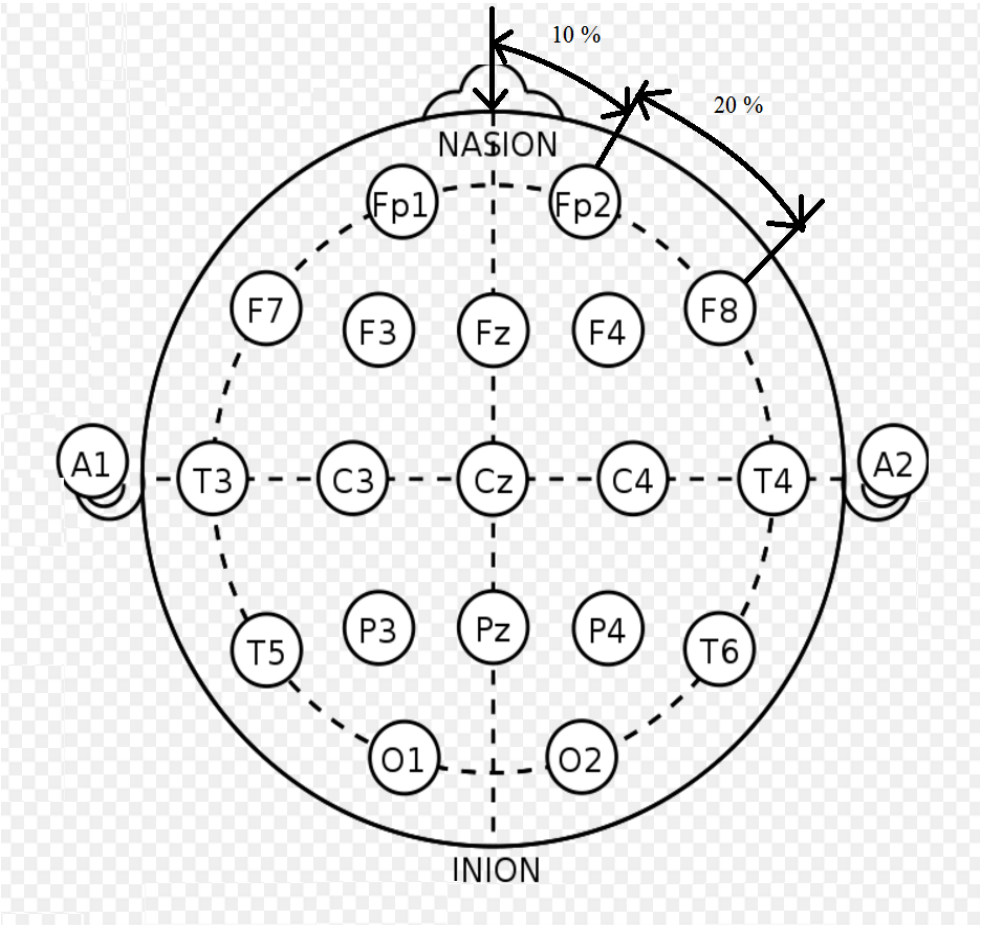

Figure 2: 10-20 Electrode Placement System. Adapted from [29]

Analyzing the collected EEG signals is very challenging because the root cause is not directly identifiable. The electrical activity recorded by the EEG instrument is the neural current through the brain volume to the scalp and the sensor electrodes. Each sensor measures a weighted sum of several neurons activity.

Multichannel EEG Recordings are acquired by using the 10-20 electrode placement system. This approach was developed to enable standardized reproducibility and so that an individual's EEG waves could be compared over time. Also, the EEG waves of different subjects can be compared to each other. The 10-20 international [24] placement of electrodes relies on percentages and not on absolute measurements which give it a distinct advantage over other positioning techniques because absolute measurements fail when it comes to comparing the heads of different sizes and reproducing results over time as the individual's age (over the years).

The name 10-20 is derived from the concept that the adjacent electrodes are placed apart by a distance that equals 10% or 20% of the distance from two anatomical points. The two anatomical points are the nasion (the point where the bridge of the nose meets the forehead) and the inion (the lowest point on the occiput). The Nomenclature of the 10-20 System [24] is composed such that the letter represents the lobe of the brain the electrode lies on and the number denotes the hemisphere the electrode is placed. Figure 2 [31] shows 21 primary electrodes positioned according to the 10-20 international system. The common letters used in the 10-20 system are F for Frontal, P for Parietal, C for Central, Fp for Frontopolar and O for Occipital. The letter A is used to denote electrodes placed on the Earlobe. It should be noted that there is no central lobe and the letter C is used only to help in identifying the location of the electrodes. The letters of the electrodes positioned in the 10-20 system can be recognized by looking into the five transverse planes. The numbers accompany the letter based on the hemisphere the electrode is located. The odd

numbers (1,3,5,7) represent the left hemisphere, and the even numbers (2,4,8) represent the right hemisphere. The letter z (zero) denotes the midline longitude of the brain. The numberings of the electrodes can be identified by examining the electrode placement through the sagittal planes.

When needed for a high-resolution recording of brain activity with EEG, a more detailed recording with a larger number of electrodes is carried out. The number of electrodes utilized is increased to 81 and 345 by following the 10-10 system and the 10-5 systems respectively. The electrode locations in the new modified systems are obtained by placing them midway intermediate to the adjacent electrodes of the 10-20 international system. The new letter codes for the electrode names are obtained by combining the letters of the adjacent electrodes. For example, the FC4 electrode denotes the electrode placed between the F4 and the C4 electrodes.

When dealing with multichannel EEG signals one comes across non-brain artifacts [24] a large number of times. An artifact is a waveform of the EEG signal which is not of cerebral origin. Mistaking an artifact for an EEG wave is not an uncommon error for EEG waveform readers. In Chapter IV we will discuss in detail the various types of non-cerebral artifacts and noise signals present in a multichannel EEG segment.

1.3 Affective Computing

Affective Computing [10] is the study of the systems that acquire, process, interpret and recognize aspects related to human emotion. The origins of this branch of research can be traced back to Rosalind Picard's 1995 paper [28] on affective computing. The primary motivation for studying affective computing is to provide the computers with the ability to simulate empathy. Human emotions can be measured either as discrete states where the emotional states are distinct entities such as happy, sad, anger, etc. or as continuous scales in three dimensions [4] which are valence, arousal, and dominance. Valence is the dimension of affect that captures the amount of pleasantness expressed in an emotion. Positive valence represents happiness while negative emotion denotes sadness. Arousal is associated with the amount of activeness one expresses with his/her emotion. Positive arousal corresponds to excitement whereas negative arousal corresponds to boredom. Dominance is the dimension of affect that represents the quantity of control the person has over the emotions being expressed. Positive dominance denotes confidence while negative dominance shows fear in the emotion displayed. In this thesis, we will be viewing affect in terms of valence, arousal and dominance levels. Figure 3 shows a mapping between the discrete emotional states vs. valence and arousal across a 2D plane.

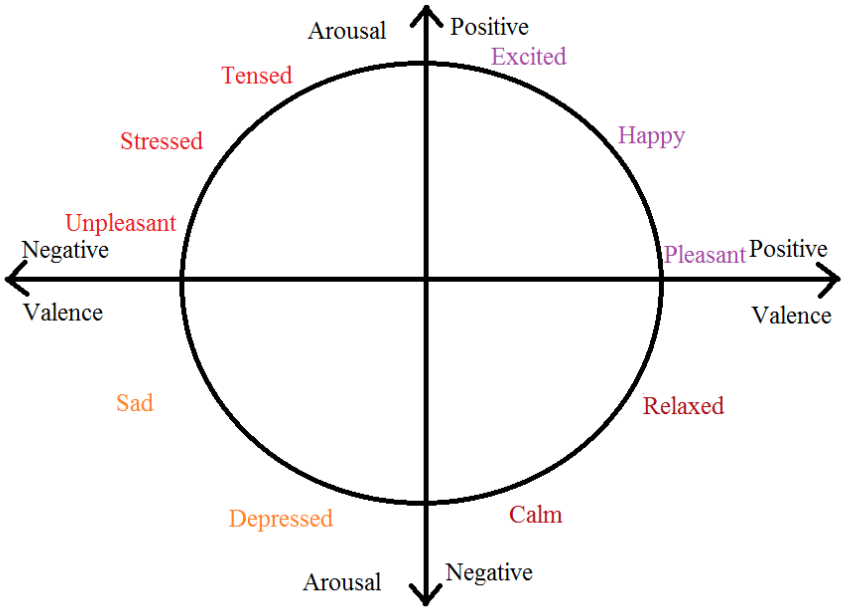

Figure 3: Mapping between Valence and Arousal vs. Discrete Emotional States

CHAPTER II

SIGNAL PROCESSING AND MACHINE LEARNING IN BRAIN-COMPUTER INTERFACES

The collective process of first pre-processing the acquired multichannel EEG signals and extracting the relevant features representing the signals belonging to the particular classes using filters and signal processing techniques and learning a classifier to predict new test multichannel EEG signal via machine learning is called Brain-Computer Interface Paradigms. The workflow involved in BCI paradigms is shown in Figure 4. This chapter primarily emphasis on the part of feature extraction from EEG signals using signal processing transformations and constructing a machine learning model from the extracted features. We will discuss briefly some of the common approaches to feature representations of multichannel EEG signals that have been used and later describe in detail the following approaches to feature representations of EEG signals

1) Power Spectral Density Features
2) Spatial Projection Features
3) Discrete Wavelet Transforms

Before we dive into the feature extraction using the above three techniques we will first briefly discuss prior work on EEG BCI paradigms for emotion recognition.

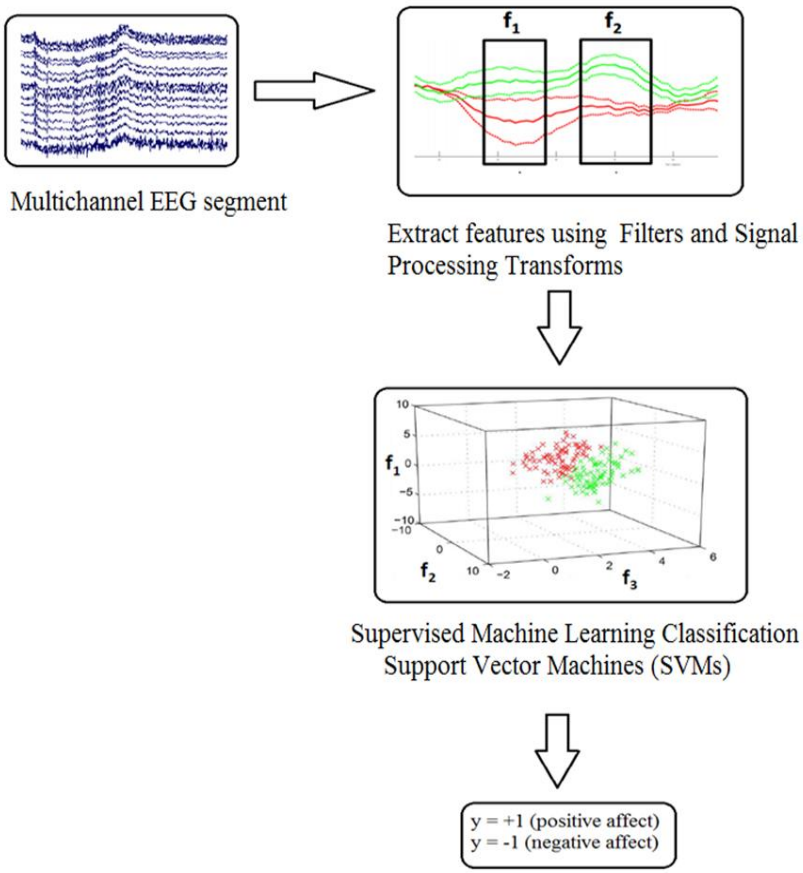

Figure 4: Workflow in Brain-Computer Interface Paradigms

2.1 Prior Work on Affective Electroencephalography based Brain-Computer Interface Paradigms

The most traditional method is to extract the power spectrum features [1], [2] from the channels and average them over the commonly studied oscillatory frequency bands. In [1], the difference between the spectral power of all the symmetrical pairs of electrodes on the

right and left hemispheres are also computed and added to the existing features. The commonly studied oscillatory bands are given below

a) theta (4 - 8 Hz)

b) slow- alpha (8 - 10 Hz)

c) alpha (8 - 13 Hz)

d) beta (13 - 30 Hz)

e) gamma (30 - 44 Hz)

The authors of [1] have used Naïve Bayes classifier with the above features to perform single trial classification.

In [5], the features corresponding to the fractal dimensions were extracted by applying the Higuchi algorithm to the dataset they collected. The authors in [5] assume thee EEG signals to be nonlinear and chaotic and hence employ the Higuchi algorithm from the time series data to collect the features representing the EEG data. Let X_m^k denote the time series of a single channel EEG data as shown below

$$X_m^k : X(m), X(m+k), \ldots, X\left(m + \left[\frac{N-m}{k}\right].k\right) \tag{2.1}$$

where *m* is the initial time and *k* is the interval time.

k sets of $L_m(k)$ are calculated as follows

$$L_m(k) = \frac{\left\{\left(\sum_{i=1}^{\left[\frac{N-m}{k}\right]} |X(m+ik) - X(m+(i-1)k)|\right)\frac{N-1}{\left[\frac{N-m}{k}\right].k}\right\}}{k} \tag{2.2}$$

and let $\langle L(k) \rangle$ denote the average of $L_m(k)$. The fractal dimension *D* is related to the average of $L_m(k)$ as depicted in equation (2.3).

$$< L(k) > \propto k^{-D} \quad (2.3)$$

The fractal dimensions were given as inputs to a Support Vector Machine classifier. The results reported after test classification showed significantly high recognition scores.

Murugappan et. al [7] used the Discrete Wavelet Transform (DWT) to extract information from the EEG signal data. The emotions were classified into discrete emotional states, namely disgust, happy, surprise, fear and neutral. The wavelet features were represented in the form of conventional and modified energy based features. Linear Discriminant Analysis (LDA) and k-Nearest Neighbors (kNN) pattern classification was employed to perform pattern classification on the extracted feature representations. Signal statistics features were explored by the authors of [8]. These features primarily included the mean of the channel signal, standard deviation, signal difference, Hjorth Features, and Histogram of Crossings (HOC). Feature selection was also applied. The authors in [8] selected a Quadratic Discriminant Analysis (QDA) with diagonal covariance estimate as their classifier.

In [14], the authors employed an unconventional approach by using subject information as privileged information along with the power spectrum features from the five frequency bands in [1]. In [14] more emphasis was put into constructing the classifier using Bayesian Networks to recognize emotions. One three node Bayesian Network applied was triangle structured, while the other was V structured. The models for each of the techniques were computed using Maximum Likelihood Estimate from the training set. During testing,

given a new data, the EEG features were calculated and the class label y was assigned based on the values of posterior probability. In [17], the authors apply a similar strategy for implicit hybrid video tagging using Brain Computer Interfaces.

In [9], the authors implemented Filter Bank Common Spatial Patterns algorithm which is very computationally extensive algorithm. The multichannel EEG segment was first spectrally filtered using five filter banks and later spatially filtered with eight spatial filters. The variance of the filtered signals was extracted as features and a Generalized Linear model with a Logistic link function was used as a classifier. The performance using this paradigm on their collected dataset was remarkable.

The preprocessing of the multichannel EEG segment was performed by applying a High Pass Infinite Impulse Response (IIR) Filter with 0.1 – 1Hz transition band. This was employed to insure the DC noise was removed from the acquired signals. We will now discuss the features experimented with EEG signals in this thesis work using Power Spectral Density, Discrete Wavelet Transforms and Common Spatial Patterns in Sections 2, 3, and 4 respectively. The details of computing a Support Vector Machines model is detailed in Section 5.

2.2 Power Spectral Density Features

In this thesis work two types of features involving Power Spectral Density (PSD) was explored. The Welch's algorithm [11] was used to compute the Power Spectral Density from multichannel EEG signals. The Welch's algorithm is based on the Parseval's theorem depicted by equations (2.4) and (2.5).

$$F\{x(t) * x(t)\} = X(f).X^*(f) \tag{2.4}$$

$$F\{x(t) * x(t)\} = |X(f)|^2 \tag{2.5}$$

where, x(t) is the discrete time signal and X(f) is the DFT of x(t).

The steps involved in the Welch's algorithm [11] are given as follows:

a) The discrete time signal is split into several segments with 50% overlap between each of the segment.
b) The segments are then windowed using a window function.
c) The spectrogram is calculated by computing the Discrete Fourier Transform (DFT) and then the squared magnitude of the DFT signal is taken.

After employing the Welch's algorithm an array of power measurements vs. frequency bin is obtained. There were two types of feature vectors composed using the PSD computations. The logarithmic PSD for all the channels were summed up with respect to the same frequency bins to obtain the feature vector for a trial. It should be noted that the feature vector formed in this fashion contains information of PSD only as a function of frequency [6]. In order to see the impact of any spatial information on the recognition of

affect another type of feature vector was constructed using procedure described as follows. After computing the PSD for each channel as a function of frequency. The logarithmic average of the PSD for each channel is calculated for each of the frequency bands given below:

1) theta (4 - 8 Hz)
2) slow- alpha (8 - 10 Hz)
3) alpha (8 - 13 Hz)
4) beta (13 - 30 Hz)
5) gamma (30 - 44 Hz)
6) (44 – 54 Hz)
7) (54 – 64 Hz)

The additional high frequency bands of (44 – 54 Hz) and (54 – 64 Hz) were chosen so that the role of high frequency activities can be examined in emotion recognition from EEG activity. The feature vectors from this kind of formation have the form [# channels x # frequency bands] compared to the previous approach which take the form [# frequency range] for each sample trial.

2.3 Discrete Wavelet Transforms

The Wavelet transform is a technique to decompose an input signal into a set of elementary waveforms called “wavelets” and provides a way to analyze the signals by examining the coefficients\ weights of these wavelets. The Fourier transforms have a major disadvantage

of containing only the frequency information of the signal with no information on the time resolution. In other words, by applying the Fourier transforms we will be able to determine all the frequencies present in the signal but not know when they are present. This difficulty is resolved by employing the Discrete Wavelet Transforms (DWT) as they represent joint time - frequency representations. The key advantage of Discrete Wavelet Transform [21] is the inherent multi-resolution nature. The DWT [21] partitions the frequency axis smoothly and recursively so that the transformation can analyze each segment (i.e. frequency band) of the signal with a resolution matched to its scale. The process of obtaining the coefficients of the DWT is by performing convolution through a cascade of filters.

The input signal is first split into low and high frequency bands by convolution and subsampling operations with a pair of filters consisting of a "lowpass" filter $\{h_k\}$ and a "highpass" filter $\{g_k\}$ on the discrete time domain. While obtaining the coefficients of DWT, this decomposition process is iterated only on the low frequency bands and each time the high frequency coefficients are retained. Each high frequency sub-band is spanned by a set of translated versions of a single elementary waveform commonly referred to as the "mother wavelet" with a specific scaling parameter. Each low frequency sub-band is spanned by a set of translated versions of another single elementary waveform known by the term "scaling function". Figure 5 shows the shapes of the mother wavelets for Daubechies, Haar, Symlet and Mexican hat wavelet families.

Time and frequency usually are viewed as two different domains when representing signals as functions, but they are inextricably linked. If one attempts to gather precise information of the signal with respect to time, they must accept some uncertainty in frequency, and vice versa. This is the Heisenberg's Uncertainty principle in signal processing. The multichannel EEG signal is decomposed using two-dimensional wavelet transform. A two-dimensional scaling function φ(x,y), and three two-dimensional wavelets, $\psi^H(x,y)$, $\psi^V(x,y)$, $\psi^D(x,y)$ are required to perform a two-dimensional wavelet decomposition.

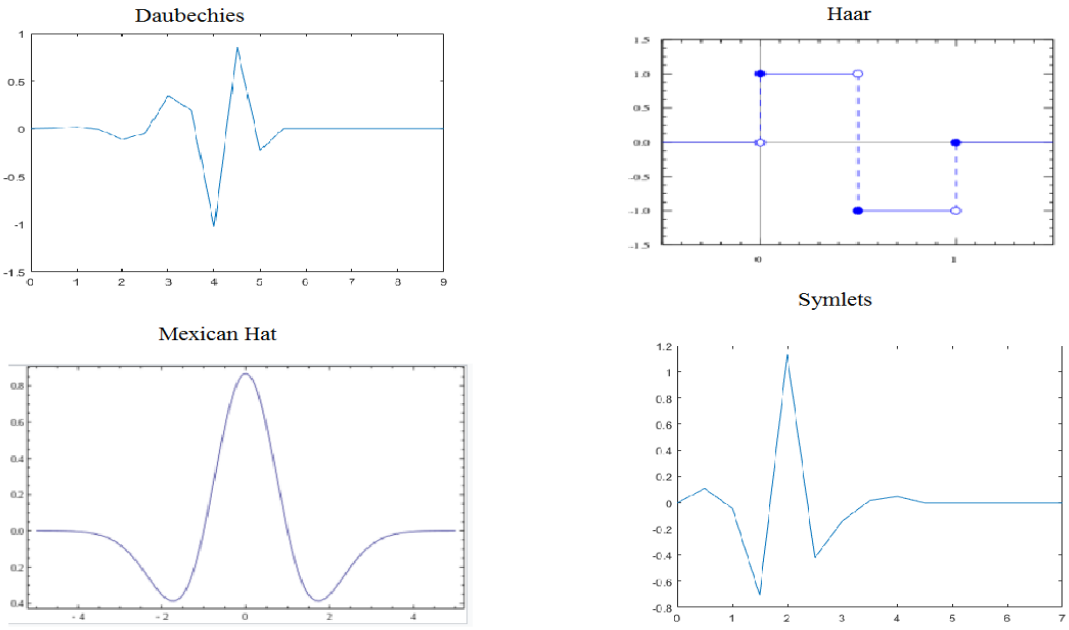

Figure 5: Mother Wavelets of Daubechies, Haar, Mexican Hat and Symlets Wavelet Families

The scaling functions and the "directionally sensitive" wavelets chosen for obtaining the coefficients of DWT [21] are separable as indicated by the equations (2.6) – (2.9). These wavelets measure functional variance along different directions: ψ^H measures variations along horizontal edges, ψ^V responds to variations along the vertical edges and ψ^D indicates variations along the diagonals. The block diagram representing the two-dimensional Wavelet Transform is shown in Figure 2.2.

$$\varphi(x,y) = \varphi(x)\varphi(y) \tag{2.6}$$

$$\psi^H(x,y) = \psi(x)\varphi(y) \tag{2.7}$$

$$\psi^V(x,y) = \varphi(x)\psi(y) \tag{2.8}$$

$$\psi^D(x,y) = \psi(x)\psi(y) \tag{2.9}$$

In order to define the equations used to obtain the two-dimensional wavelets, we must first define the scaled and translated basis functions displayed in equations (2.10) and (2.11). Figure 6 shows the flow diagram of computing the 2-dimensional DWT coefficients for a particular level of decomposition.

$$\varphi_{j,m,n}(x,y) = 2^{\frac{j}{2}}\varphi\left(2^j x - m, 2^j y - n\right) \tag{2.10}$$

$$\psi^i_{j,m,n}(x,y) = 2^{\frac{j}{2}}\psi^i\left(2^j x - m, 2^j y - n\right), \quad i = \{H,V,D\} \tag{2.11}$$

Please note that the index i refer to the directions of the wavelet coefficients.

The Discrete Wavelet Transform of a multichannel EEG segment represented by f(x,y) of size MxN where, x denotes the channel , y denotes the time point is given by equations (2.12) and (2.13).

$$W_{\varphi}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \varphi_{j0,m,n}(x, y) \tag{2.12}$$

$$W_{\psi}^{i}(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\, \psi_{j,m,n}^{i}(x, y), \quad i = \{H, V, D\} \tag{2.13}$$

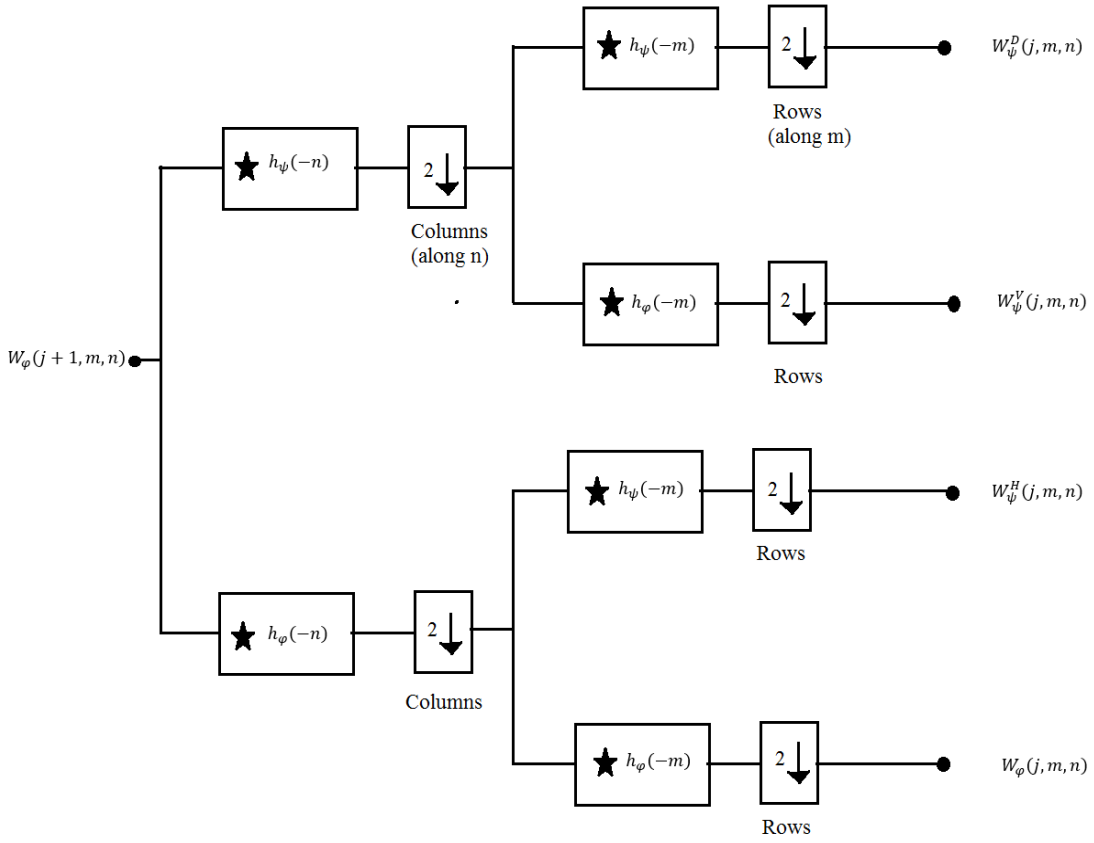

Figure 6: The Two-Dimensional Discrete Wavelet Transform

In this work, 5 levels of decomposition using Daubechies wavelet family of order 5 was employed. The DWT transform first decomposes the multichannel EEG signal into approximation coefficients and detailed coefficients. The approximation coefficients are

subsequently split into new approximation and detailed coefficients. This process is repeated five times to obtain five levels of decomposition. As the coefficients of DWT represent frequency characteristics of the multichannel EEG segment. The variance of these coefficients along the time axis was taken from each level of decomposition to extract multi-resolution time-frequency representation. The features were converted into logarithmic scale.

2.4 Common Spatial Patterns

Due to volume conduction, inferring the brain activity from multichannel EEG recordings is similar to identifying components from a highly-blurred image. Common Spatial Patterns (CSP) [12] refers to the technique of learning spatial filters from various acquired EEG signal recordings. This technique was introduced by the Berlin BCI project [19] to perform the task of Motor Imagery using EEG based BCIs. It is considered as one of the state of the art techniques for Motor Imagery BCIs. We apply this technique for Emotion Recognition and compare the results from the other proposed methods. This method is related to the Principal Component Analysis in machine learning. A brief outline on the construction of spatial filters is provided for the remainder of this section.

The method used to design the spatial filters is based on the simultaneous diagonalization of the two covariance matrices. CSP [12] is a supervised decomposition of a signal parametrized by the matrix $W \in R^{C*m}$ (C represents the number of channels and m the

number of spatial filters). The matrix W projects $x(t) \in R^C$ in the original sensor space to a substitute sensor space $x_{CSP}(t) \in R^m$ as shown by equation (2.14).

$$x_{CSP}(t) = W\, x(t) \tag{2.14}$$

The multichannel EEG signal is first band-passed filtered in the frequency range of [0.3Hz - 54 Hz]. The signal is then normalized to zero mean and unit variance. Pooled covariance matrices $\Sigma^{(+)}$ and $\Sigma^{(-)}$ are computed using the above preprocessed EEG signal as per equation (2.15).

$$\Sigma^{(c)} = \frac{1}{|I_c|}\sum_{i \in I_c} XiXi^T \quad c \in \{+, -\} \tag{2.15}$$

I_c is the set of indices that belong to class c and $|I_c|$ denotes the number of examples in class c. The spatial filter W is then obtained by the simultaneous diagonalization of the above two covariance matrices. This process is detailed by the equations below

$$W^T\, \Sigma^{(+)}\, W = \Lambda^{(+)} \tag{2.16}$$

$$W^T\, \Sigma^{(-)}\, W = \Lambda^{(-)} \tag{2.17}$$

such that $\Lambda^{(+)} + \Lambda^{(-)} = I$. The W matrix is obtained by solving the generalized eigenvalue problem as shown below

$$\Sigma^{(+)} w = \lambda\, \Sigma^{(-)} w \tag{2.18}$$

where *w* denotes the columns of W and represent the generalized eigen vectors and λ_i^c's are the generalized eigenvalues which are the diagonal elements of Λ^c. Once the matrix W is obtained the multichannel EEG segment is spatially filtered using equation (2.14).

Let S_d and S_c represent discriminative and common activity of the EEG signals between the two classes. The mathematical formulation of S_d and S_c are shown in equations (2.19) and (2.20).

$$S_d = \Sigma^{(+)} - \Sigma^{(-)} \tag{2.19}$$

$$S_c = \Sigma^{(+)} + \Sigma^{(-)} \tag{2.20}$$

The process of obtaining the spatial filters w can be described as solving the optimization problem defined below.

$$maximize_w \quad \frac{w^T S_d\, w}{w^T S_c\, w} \tag{2.21}$$

2.5 Support Vector Machines

Now, given a set of features from multichannel EEG signals and class labels corresponding to the affect levels of the subject our primary goal is to learn a parametric model M, that encodes the mapping between the extracted features and the affect labels. This parametric model M is learnt using a Support Vector Machine classifier. The parametric model learnt is used to predict affect labels on future trials of the subject. This process is illustrated in Figure 7.

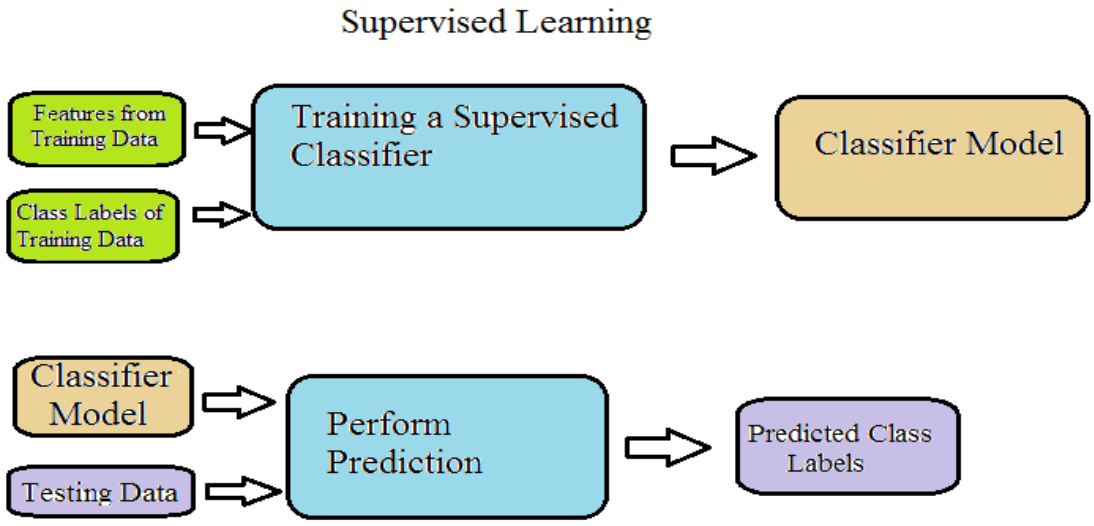

Figure 7: Illustration of Supervised Machine Learning

In this section, we will discuss the Support Vector Machine (SVM) learning algorithm [18], [22]. This section primarily deals with the training and testing involved when employing an SVM algorithm [18], [22] to the features extracted using the procedures described in the previous sections. In the later part of this section details on selecting certain parameters of the SVM is also provided. SVMs are believed to be the state of the art "off-the-shelf" supervised learning algorithms.

In an SVM the predictions on the test feature set depend only on a subset of the training data. This subset of the training data is commonly referred to as the support vectors. The SVMs also make use of kernel functions which are denoted by $\kappa(x, x')$. This combination of the kernel trick and a modified loss function due to subset of the training samples is what makes the SVMs very powerful. The SVM models encode sparsity in the loss

function by considering the support vectors. SVMs are very unnatural from a probabilistic point of view, but they become more appropriate when we talk about margins and the idea of separating the data with a large “gap”.

First, we will formulate SVM [18], [22] as an optimal margin classifier. The solution to a classification problem can be visualized as the process of obtaining a decision boundary that separates the two classes of data. While training an SVM classifier we try to find a decision boundary that maximizes the margin between the two classes of data, since this would predict a confident set of predictions on the training set and a good “fit” to the training data. This results in a classifier that separates the positive and negative training samples with a large margin. Figure 8 illustrates the large margin principle introduced above.

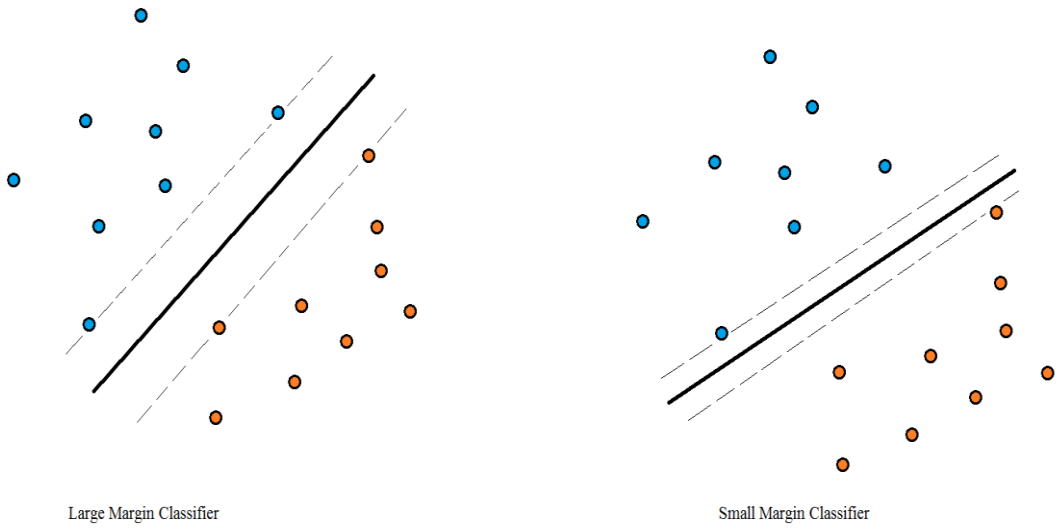

Figure 8: Illustration of a Large Margin Classifier and a Small Margin Classifier

Let the training dataset be represented as $\{(x_i, y_i)\}$ where $i = 1,2,3 \ldots, N$, where there are N examples, x_i represents the features and y_i denotes class labels. Picking the best separating hyper-plane as the one that maximizes the margin is illustrated by equation (2.22).

$$max_{w,w0}\ min_{i=1}^{N}\ \frac{y_i(w^T x_i + w0)}{||w||} \tag{2.22}$$

Note that *w, w0* represent the parameters of the hyperplane and that rescaling the parameters *w* to *kw* and *w0* to *kw0* does not change the distance of any point to the boundary since the *k* factor cancels out when we divide by $||w||$. Let $f_i = w^T x_i + w0$, and let us assume a scaling factor of $y_i f_i = 1$ for the point that is closest to the decision boundary. Then, $y_i f_i \geq 1$ for all *i*. Finally, we should note that, maximizing $(1/||w||)$ is equivalent to minimizing $||w||^2$. With the above changes in mind, the new objective of the SVM formulation is given by the equation below.

$$min_{w,w0}\ \frac{1}{2}||w||^2 \quad such\ that \quad y_i(w^T x_i + w_0)\ \geq 1, i = 1,2, \ldots, N \tag{2.23}$$

The above equation is a constrained optimization problem with a convex quadratic objective function with linear constraints. This kind of constrained optimization problem can be solved using Lagrange duality. Equation (2.23) is often referred to as the primal problem of an SVM. Using equation (2.23), the Lagrangian is formulated and the Karush-Kuhn-Tucker (KKT) conditions are applied to obtain the dual form of the SVM as described by the equations below. Let α_i's and β_i's denote the Lagrange multipliers, and

let *L (w, α, β)* represent the Lagrangian function. The KKT conditions for the primal SVM problem are summarized below.

$$\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) = 0,\ i = 1,2, \dots, d \tag{2.24}$$

$$\frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) = 0,\ i = 1,2, \dots, l \tag{2.25}$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1,2, \dots, k \tag{2.26}$$

$$g_i(w^*) \leq 0, \quad i = 1,2, \dots, k \tag{2.27}$$

$$\alpha^* \geq 0, \quad i = 1,2, \dots, k \tag{2.28}$$

w*, α*, β* which satisfy the above KKT conditions are the solution to the primal and dual problem of SVM. One should also note that equation (2.26) implies that for the points where $\alpha_i^* > 0$, the $g_i(w^*) = 0$ (i.e., equation (2.27) holds with the equality constraint) which are the key for obtaining the "support vectors". The Lagrangian for our SVM problem is shown below.

$$L(w, w0, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i [y_i(w^T x_i + w0) - 1] \tag{2.29}$$

After employing the KKT conditions to the Lagrangian function defined above, we obtain the dual formulation of the SVM described by the equation below.

$$max_\alpha\ W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j\ \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$such\ that\ \alpha_i \geq 0, \quad i = 1,2, \dots, N\ and\ \sum_{i=1}^{N} \alpha_i\ y_i = 0 \tag{2.30}$$

As mentioned before, the points where $\alpha_i > 0$ help determine the support vectors and $\langle x_i, x_j \rangle$ denotes a dot product which represents a linear kernel. If one wants to apply a

different kernel then $< x_i, x_j >$ must be replaced with the desired kernel function $\kappa (x_i, x_j)$. The different kernels commonly used with SVM are listed in Table 1.

Kernel Type	Description	Mathematical Representation
Linear kernel	Most commonly used kernel	$\kappa (u, v) = u'v$
Polynomial kernel	Polynomial kernel of order p	$\kappa (u, v) = (1 + u'v)^p$
Gaussian kernel	Also, known as Radial Basis Function (RBF) kernel	$\kappa (u, v) = exp (\gamma \| u - v \|^2)$

Table 1: Descriptions of Kernel Functions Commonly Used in SVMs

The primal and dual problems were formulated under the assumption that the feature dataset was linearly separable. If the feature dataset was not linearly separable, then there will be no feasible solution to the primal and dual problems of SVM [18]. Hence, we introduce slack variables $\xi_i > 0$ so that the SVM algorithm works for non-linear feature datasets and will not be sensitive to outliers. The points are assigned a slack value depending on the portion of the decision boundary and the side of the margin they occupy in the feature space. If the point is on the correct side of the margin boundary, then $\xi_i = 0$. If $0 < \xi_i \leq 1$, the point lies inside the margin but on the correct side of the decision boundary. The point lies on the wrong side of the separating hyperplane if $\xi_i > 1$. The assignment of the slack variables ξi is illustrated in Figure 9. This formulation of the SVM objective function with slack variables is known as SVM with soft margin constraints. The

primal and dual SVM objective functions with soft margin constraints are respectively given below.

$$min_{w,w0,\xi}\ \frac{1}{2}||w||^2 + C\ \sum_{i=1}^{N}\xi_i \quad such\ that, \xi_i\ \geq 0,\ y_i(x_i^T w + w0) \geq 1 - \xi_i \quad (2.31)$$

$$max_{\alpha}\ W(\alpha) = \sum_{i=1}^{N}\alpha_i\ - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$such\ that\ \ 0 \leq \alpha_i\ \leq C,\ \sum_{i=1}^{N}\alpha_i y_i\ = 0 \quad (2.32)$$

The parameter C acts as a regularization parameter that controls the number of errors we are willing to tolerate on the training feature set. The Sequential Minimal Optimization (SMO) algorithm developed by John Platt [27] gives an efficient technique to solve the dual formulation of SVM.

The training of an SVM model [18] by employing the features extracted from the multichannel EEG signals and the affect labels requires the specification of the kernel function and the C parameter value. The value of C is chosen by cross-validation on each subject. The C parameter interacts strongly with the kernel function of the SVM. For example, suppose we are using a Gaussian kernel with $\gamma = 5$ which corresponds to a narrow kernel, then the value of C which will give us an optimum classifier is small as we need heavy regularization. The optimum C was obtained using cross-validation over a one-dimensional grid with $C \in \{10^{-2}, 10^{-1}, 1, 10, 100\}$ similar to the procedure adopted by [18]. The optimum C and γ when using a Gaussian kernel are obtained by performing cross-validation over a two-dimensional grid with $C \in \{10^{-2}, 10^{-1}, 1, 10, 100\}$ and $\gamma \in \{10^{-2}, 10^{-1}, 1, 10, 100\}$ as in [18].

Please also note that the feature set was standardized (zero mean and unit variance) before given as inputs to the SVM classifier.

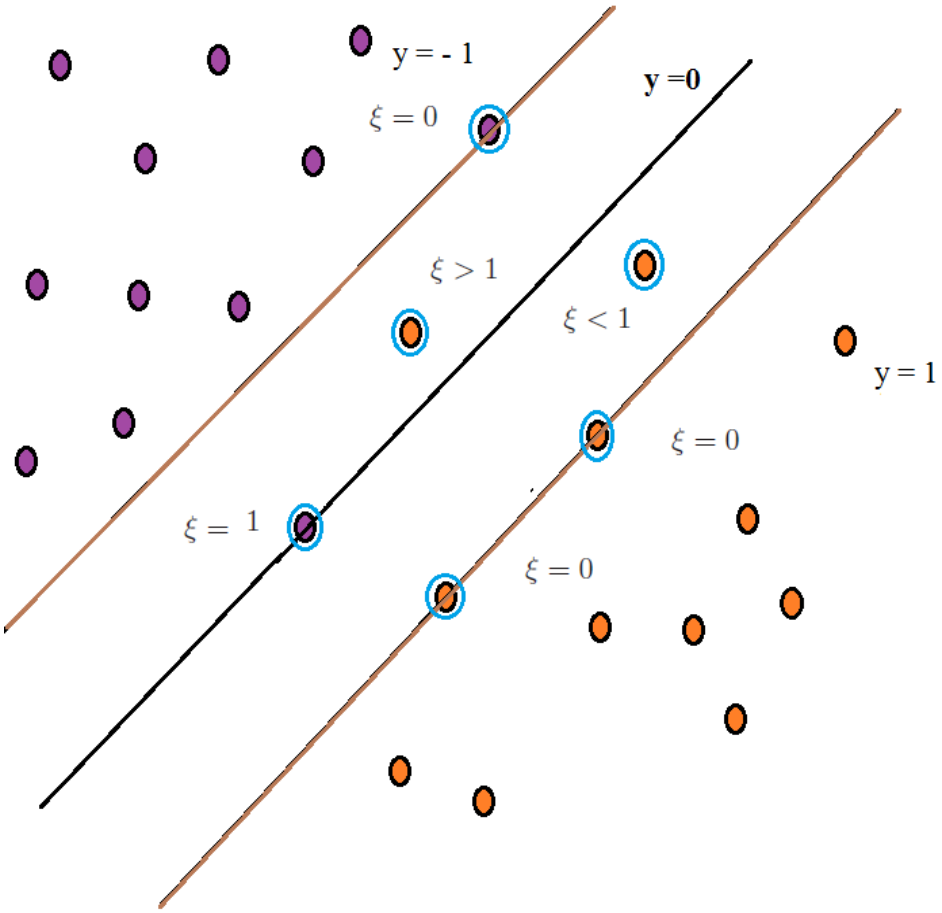

Figure: 9: Illustration of the Principle of Soft Margin SVM

The trained SVM models were evaluated using k-Fold Cross Validation. The feature dataset is split into K folds. For each fold k ∈ {1, 2, …, K} we train an SVM model on k-1 folds and test the model on the kth fold in a round robin fashion This process is

illustrated in Figure 10. Commonly used value of K is K=5, also known as 5-fold cross validation. When K = N, where N denotes the number of data samples present the technique is called as Leave-one Out Cross Validation (LOOCV) since in fold j, we train on all the samples except for the j^{th} sample and then test on the j^{th} sample. The classifier model was computed for each subject as emotions are personal and greatly vary depending on the person.

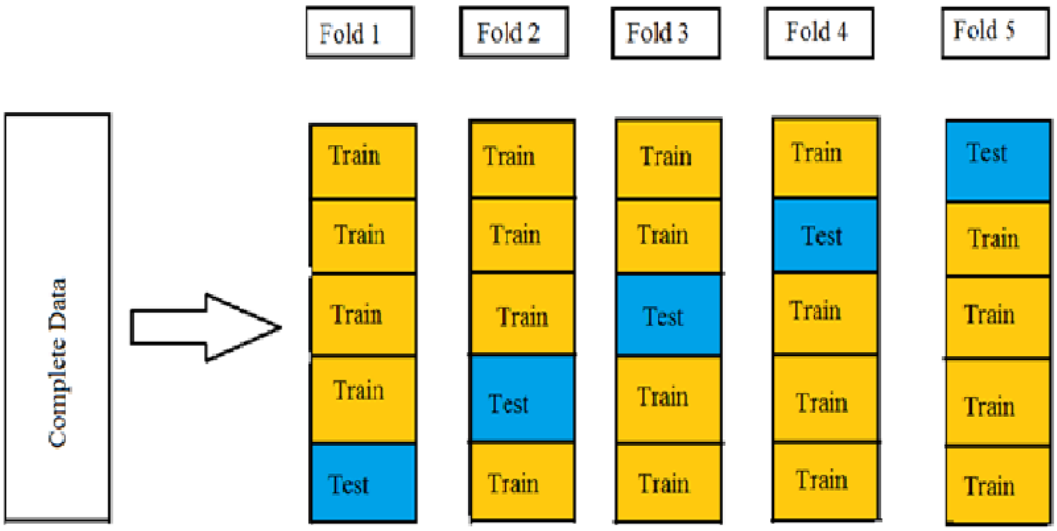

Figure 10: k-Fold Cross Validation with k =5

CHAPTER III

GENERATING SEMI-SIMULATED ELECTROENCEPHALOGRAPHY SIGNAL DATA

In many BCI experiments the amount of trials conducted for each subject is small in number. The amount of EEG signal data present for each subject is less compared to the quantity of data encountered in a common machine learning problem. As gathering more data is said to be more advantageous than designing a cleverer algorithm, in this chapter we will discuss the process of acquiring more EEG signal data by generating semi-simulated EEG data. The semi-simulated EEG data is computed by adding modeled artifact signals to the assumed artifact free multichannel EEG signals. In this chapter, we will first consider Electroencephalographic artifacts, examine how they are modeled and later on describe how the semi-simulated data was computed from these modeled artifacts.

3.1 Modeling Electroencephalographic Artifacts

Portions of waveforms in EEG that are not of cerebral origin are known as EEG artifacts. Artifact waveforms [24] are very common in EEG signals that it is, in fact, a common error for inexperienced EEG waveform readers to mistake an artifact for an EEG wave. Some EEG artifacts are easy to recognize like the muscle artifacts in the frontal and the temporal channels because of their distinctive appearance, but a few EEG artifacts are difficult to identify like the eye-blink artifacts, AC sinusoidal noise interference, etc.

Certain artifacts in multichannel EEG signals such as eye-blink artifacts, temporal muscle artifacts and electrode pops have characteristic shape and appearance which can be obtained using a combination of techniques of wave pattern generation and waveform topography. In this thesis, we will be modeling the following artifacts for computing the semi-simulated EEG signal data: -

(a) Eye-Blink Artifacts

(b) Temporal Muscle Artifacts

(c) AC Sinusoidal Noise

(d) Unfiltered White Gaussian Noise

3.1.1 Eye-Blink Artifact

When an individual close their eyes, the globes of the eyes deviate upwards. This upward movement of the globes is hidden from view by the closed eyelids. The eye-blink EEG artifact [24] caused by the upward movement of the globes during eye-blinking or eye closure cannot be understood without knowing that the globe of the eye contains a dipole distribution of charge. The globe of the eye contains charge distribution such that the cornea of the eye carries a net positive charge relative to the posterior of the eye which carries a net negative charge. When the eyes blink, the net positive charge of the cornea get shifted upward. This upward shift of the positive charge is reflected in the frontal channels of EEG [24], primarily Fp1, AF3, F3, Fp2, AF4, F4. It should be noted that as F3 and F4 are more distant from the cornea, they pick up less amount of positivity shift.

An extension of the same concept involving the motion of positive charge on the anterior aspect of the globe can be used to comprehend the lateral eye movement artifact. The anterior temporal electrodes F7 and F8 capture the lateral movement of charge. The above-mentioned characteristics of eye-artifacts are captured by passing random noise through a band pass Finite Impulse Response (FIR) Filter with a pass band of 1 Hz to 3 Hz [23]. The FIR filter was designed using the Parks McClellan Algorithm [26]. The modeled eye artifacts were added only to the frontal channels of EEG segments, primarily Fp1, AF3, F3, Fp2, AF4, F4, F7 and F8. The amplitude of the artifacts was considered to be in the range between -5 dB to -20 dB. The addition of the artifact to the artifact free EEG dataset was performed using Signal-to-Noise Ratio (SNR) statistics. Similar procedure for generating semi-simulated data was adopted by [23]. Figure 11 shows a sample modeled eye-blink artifacts and the semi-simulated data using the eye-blink artifact on the Fp1 channel of the EEG segment.

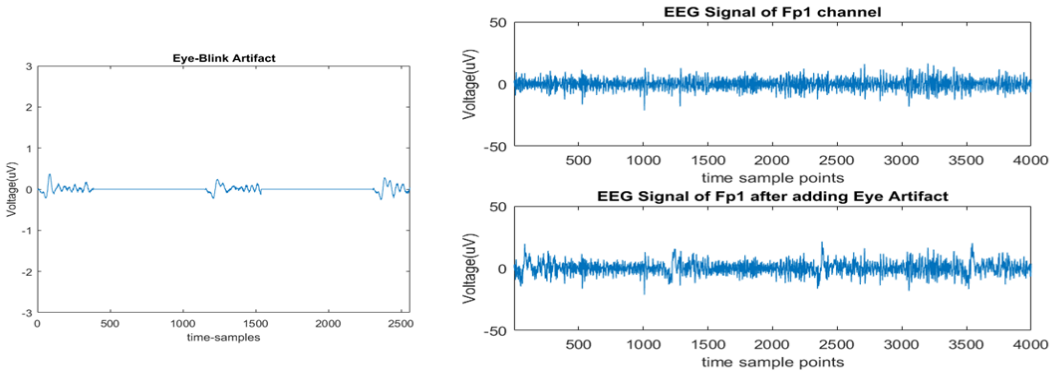

Figure 11: Modeled Eye-Blink Artifact and Semi-Simulated Data Illustrating the Eye-Blink Artifact

3.1.2 Muscle Artifact

Muscle artifact is the most commonly encountered artifact in multichannel EEG segments. Muscle artifacts [24] are captured by the electrodes that overlie the muscles of the scalp. These muscles include frontalis, temporalis, and occipitalis muscles. The appearance of muscle artifacts resembles a mixture of fast waves with spike-like potentials having variable heights and shape. These artifacts arise due to contamination from the EMG (Electromyogram) signals which occur from muscle movement from the jaws and the neck.

The temporal muscle artifacts were modeled by first generating random noise and filtering it via a Band-pass FIR filter with a pass band in the range of 20 Hz to 40 Hz [23]. The FIR Band-pass Filter was designed using the Parks-McClellan Algorithm [26]. The modeled artifacts were added to the artifact free EEG signals based on the muscle scalp map. The artifact was added with a high gain at the temporal and occipital electrode positions (around -6 dB to -10 dB) and smaller gains at the other electrode positions. Figure 12 illustrates the temporal muscle artifact generation and summation into the artifact free multi-channel EEG segment.

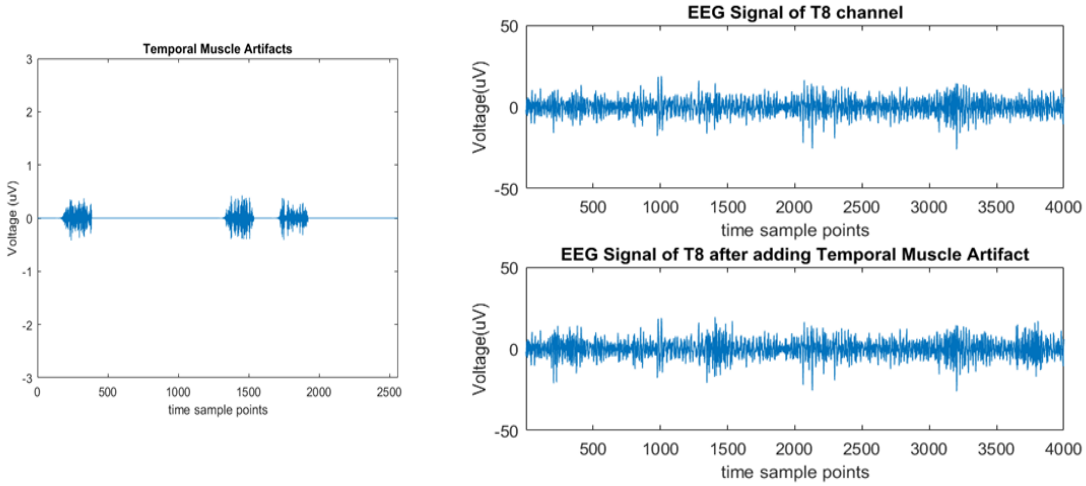

Figure 12: Illustration of Generating Modeled Temporal Muscle Artifact and Adding it to Compute the Semi-Simulated EEG Segment

3.1.3 Alternating Current Sinusoidal Noise

Small amounts of Alternating Currents (AC) from the power mains that surround the patient can flow through the patient's body accounting for this type of artifact. In a proper EEG setting, the AC current artifact is present in similar amounts on all the channels which gets subtracted out by the common mode rejection amplifiers used later. However, when one or some of the electrodes have been poorly applied the AC sinusoidal artifact will not be subtracted out in equal proportions. This causes the sinusoidal AC noise artifact to appear in the EEG segment [24]. The AC frequency current is either 50 Hz (Europe) or 60 Hz (United States) depending on the country where the EEG signals were recorded. The AC sinusoidal noise was modeled by generating the sinusoid waveform (50 Hz or 60 Hz)

and implementing the generated sinusoidal discontinuities at one randomly selected data channel [23]. This process is illustrated in Figure 13.

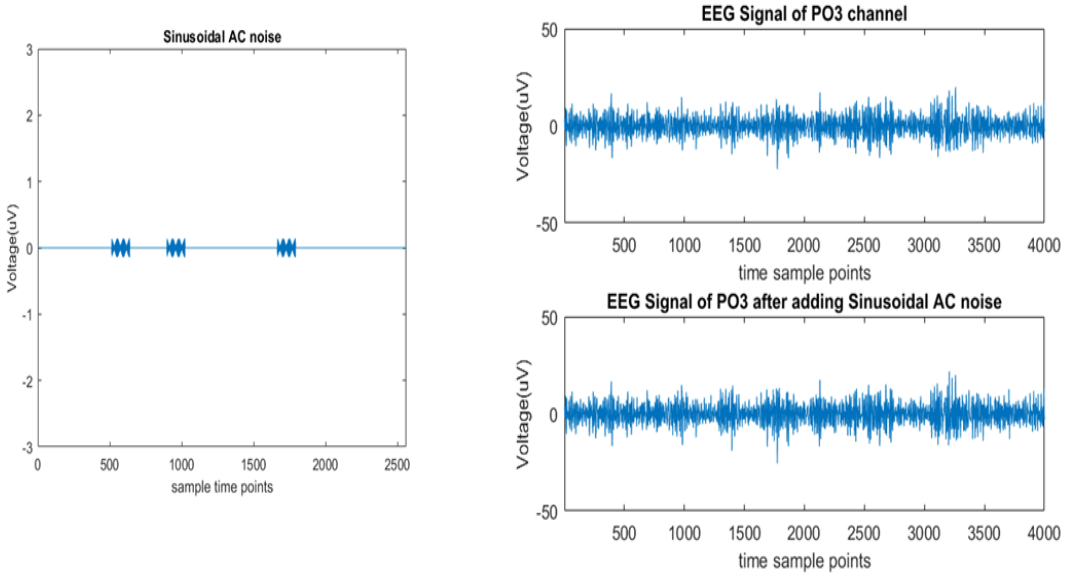

Figure 13: Illustration of Generating an AC Sinusoidal Noise of around 50 Hz and Obtaining the Semi-Simulated EEG Segment from the Modeled AC Noise.

3.1.4 Artifacts from Electrical Equipment and other types of motion

Since the EEG system setup can give out a variety of electrical signals and the long duration of the experiments can cause some restless motions by the participants. These types of noise were modeled using Unfiltered White Gaussian Noise [23]. This modeled artifact was added to the artifact free EEG segment to a randomly selected channel. The

Unfiltered White Gaussian Noise was added with an SNR around -6 dB to - 20 dB. This process of modeling artifact in the form of Unfiltered Gaussian Noise is illustrated in Figure 14.

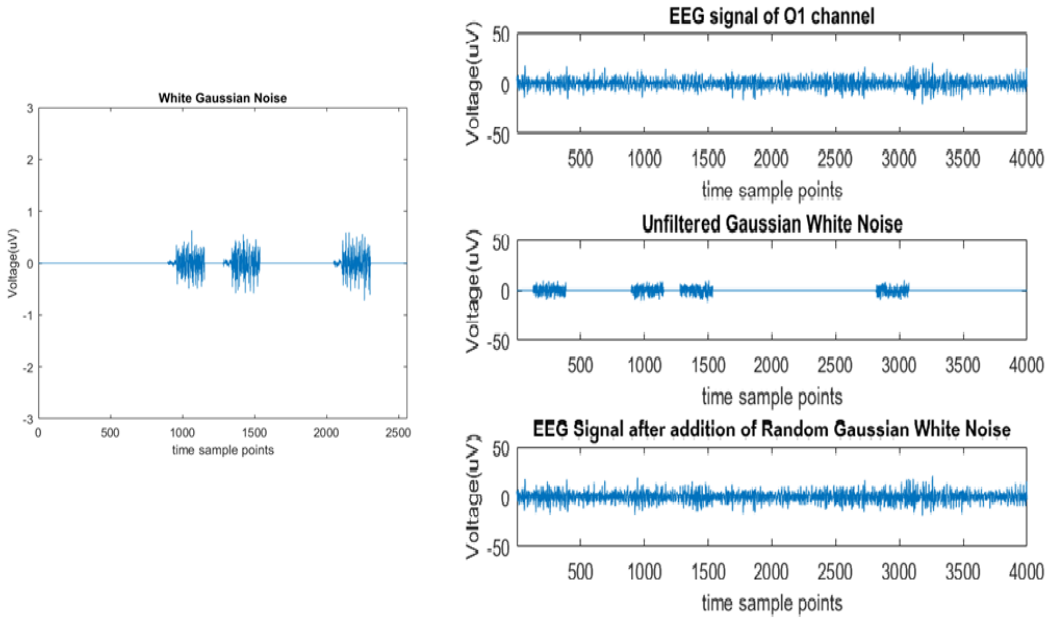

Figure 14: The Process of Obtaining Unfiltered White Gaussian Noise and Computing Semi-Simulated EEG Segment from it.

3.2 Semi-Simulated Electroencephalography Signal Data only with Random Noise

Another set of Semi-Simulated EEG signal database was generated using only simulations of random noise. This set of semi-simulated EEG signals was generated by only modeling random noise and adding these noise artifacts without any biological information to the artifact free EEG signal data. The modeled random noise was added at random time points for a small duration similar to the procedure followed previously. The channels selected were the same as those chosen when computing the Semi-Simulated EEG signal data with biological artifacts for fair comparison. The artifact was added with an SNR in the range of -6 dB to -20 dB. The details about the comparisons between the two sets of Semi-Simulated EEG signal data is given in Chapter V.

CHAPTER IV

FEATURE SELECTION

In pattern recognition tasks, identifying the most relevant features from the observed data is critical for minimizing classification error. The process of achieving this objective is often referred to as Feature Selection. Previously, we extracted features from the multichannel EEG segments corresponding to different feature spaces, primarily the power spectrum and the wavelet coefficients. In this chapter, we will describe the process of selecting features from the combination of the two subspaces that can improve our recognition of emotional affect on an individual. In the first section, we will introduce feature selection and discuss the different types of feature selection. Next, we will describe the Maximum Relevancy and Minimum Redundancy Feature Selection algorithm that was employed to select the candidate features from the two subspaces previously mentioned. Also, a short description of the framework of Max- Relevancy and Min- Redundancy to the concatenated Power Spectrum and the Wavelet features will be provided.

4.1 Introduction to Feature Selection

As mentioned earlier, Feature Selection [20] deals with selecting a subset of features from the defined feature space in order to obtain a better generalization of the attributes characterizing the system. The many advantages of Feature Selection include:

(a) Dimensionality reduction that reduces the computational cost

(b) Obtaining features that are more interpretable and help target the categorical patterns present

There are two types of feature selection techniques: (1) Filters [25] and (2) Wrappers [25]. Filter methods involve filtering out the unnecessary features and selecting only a subset of the candidate features prior to applying a classification task. The process of employing Feature Selection via a filtering technique is shown in Figure 15. Wrapper methods, on the other hand are feature selection techniques embedded around a learning method. Wrapper methods assess features based on the classifier it is "wrapped" around. These techniques have more computational cost and have more chances of overfitting the features with the given dataset. The schematic of a Wrapper technique is illustrated in Figure 16.

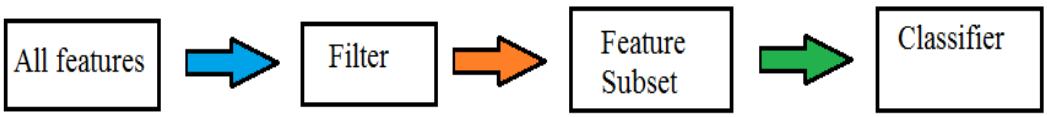

Figure 15: Feature Selection Using the Filtering Method.

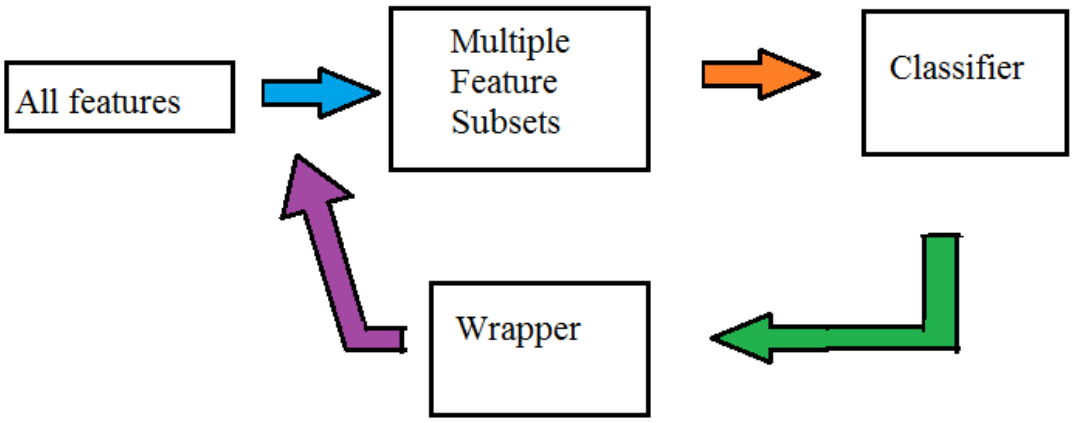

Figure 16: Schematic of Feature Selection Using Wrapper Methods.

4.2 Maximum-Relevancy and Minimum-Redundancy Feature Selection Algorithm

We will now discuss the framework of the feature selection algorithm known as Maximum Relevancy and Minimum Redundancy (mRMR) [15]. This algorithm first ranks the feature variables based on the Mutual Information criteria and later applies the constraint of Max-Relevancy and Min-Redundancy to obtain the candidate features. This approach of feature selection is a Filtering technique. Since mRMR [15] is based on mutual information, we will first define Mutual Information between two random variables *x* and *y* by equation (4.1), where *p(x,y)*, *p(x)* and *p(y)* are the probability density functions.

$$I\,(x,y) = \int\int p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dx\,dy \qquad (4.1)$$

Let x_i represent individual features from the feature set $X = \{x_i, i = 1, 2, \ldots, d\}$ where d is the dimensionality of the feature space feature space considered. Also, let there be m samples from the observed data from which the features were extracted and the class labels be represented by c. In a Max-Relevance setup, the selected features x_i have the largest mutual information $I(x_i, c)$ with the target class individually. In a sequential feature selection algorithm, the N candidate features are selected in the descending order of $I(x_i,c)$. But while performing sequential feature selection, one notices that the combination of individually good features does not guarantee a good classification accuracy. In other words, the N best features selected from sequential feature selection are not the best N features for pattern recognition.

Let's first look into the notion of Max-Dependency and then consider mRMR [15]. Dependency of a feature set is defined as the joint mutual information of a group of feature variables on the identified class label. Max-Dependency feature selection is the process of selecting the feature set S that maximizes the dependency. Max- Dependency can be represented by the equation (4.2) given below.

$$\max D(S, c), \quad D = I(\{x_i, i = 1,2, \ldots, N\}; c) \tag{4.2}$$

It should be noted that when $N = 1$, the solution is the feature variable x_j that maximizes $I(x_j, c)$ where $1 \leq j \leq d$. When $N > 1$, the approach is to increment the candidate feature set S by one feature variable after an iteration is performed to add whichever feature contributes to the largest increase in the value of $I(S; c)$. Computation of $I(S_k; c)$ takes the form given by equations (4.3) and (4.4).

$$I(S_k\,;c) = \iint p(S_k, c)\log\frac{p(S_k,c)}{p(S_k)p(c)}\,dS_k dc \tag{4.3}$$

$$I(S_k\,;\,c) = \iint p(S_{k-1}, x_k, c)\log\frac{p(S_{k-1},x_k,c)}{p(S_{k-1})p(x_k)p(c)}\,dS_{k-1}dx_k dc \tag{4.4}$$

It is difficult to obtain an accurate estimation of the multivariate probability density functions *p (x₁, x₂, …, x_d)* and *p (x₁, x₂, …, x_d, c)*. This is because the number of training samples is insufficient for an accurate estimation of the probability density functions and the estimation also involves computing the inverse of a high dimensional covariance matrix which is an ill-posed problem. Since the Max-Dependency criteria is hard to compute, we approximate it with a Max-Relevance criterion. For estimating the Max-Relevance criterion, we approximate *D (S; c)* as the mean of the mutual information between the individual feature variables being considered x_i and *c*. The mathematical formulation of Max-Relevance is shown by the equation below.

$$\max D(S;c), where\ D = \frac{1}{N}\sum_{x_i \in S} I(x_i;c) \tag{4.5}$$

It is assumed that N is the dimensionality of *S*. It is reasonable to assume that many of the features selected with the Max-Relevance criterion can be redundant as the feature variables selected could be correlated with each other. It could be advantageous to remove these features which are redundant. This task can be achieved by adding a Min-Redundancy condition described by the equation below.

$$\min R(S), where\ R = \frac{1}{N^2}\sum_{x_i,x_j \in S} I\left(x_i, x_j\right) \tag{4.6}$$

The criteria for Max-Relevance and Min-Redundancy form the mRMR [15] Feature Selection approach. The two criteria can be combined in two forms. The first combination

is in the form of a difference as shown by equation (4.7) and is referred to as mRMR with MID. The second combination is in the form of a quotient as shown by equation (4.8) and is known as mRMR with MIQ.

$$\max \Phi\,(D,R), where\ \Phi = D - R \tag{4.7}$$

$$\max \psi\,(D,R),\ where\ \psi = \frac{D}{R} \tag{4.8}$$

In the mRMR algorithm the first candidate feature is selected according to the Max-Relevance criteria of equation (4.5). The remaining candidate features are added in an incremental fashion. Let us assume that *k* candidate features have already been selected by the mRMR algorithm for the set *S* and we want to select additional features from the set **Ω**s = **Ω** - S (the feature space except for those that have been selected). The candidate feature set is incremented based on the scheme chosen and the optimization performed for that scheme. The optimization formula for Mutual Information Difference (MID) criterion and Mutual Information Quotient [16] criterion are shown by equations (4.9) and (4.10).

$$x_i = \arg\max_{i \in \Omega_S}[I\,(x_i, c)\ - \frac{1}{N}\sum_{j \in S} I(x_i, x_j)] \tag{4.9}$$

$$x_i = \arg\max_{i \in \Omega_S} \frac{I\,(x_i, c)}{\left[\frac{1}{N}\sum_{j \in S} I(x_i, x_j)\right]} \tag{4.10}$$

The mRMR [15] algorithm with both the MID and MIQ criterions was applied to the features extracted from the multichannel EEG segments during the training phase of the SVM classifier. The Power Spectral Density features of each channel averaged over six frequency bands were concatenated with the Discrete Wavelet Transform coefficient

features with the assumption that these features contained complementary information which could help in improving the generalization accuracy. The mRMR [15] with MID criterion and MIQ criterion were employed to select the candidate features to be considered as inputs to the SVM.

CHAPTER V

EXPERIMENTS AND INFERENCES

5.1 Experimental Datasets

Recent advances in EEG based BCIs have assisted in the creation of novel databases containing multichannel EEG segments along with labeled emotion classes. This section will give a brief description of the EEG databases used in this thesis work. Three EEG databases containing emotional affect labels and two sets of Semi-Simulated EEG datasets were used to test and evaluate the BCI paradigms developed.

A Database for Emotional Analysis using Physiological Signals (DEAP) [1] is one of the most popular multimodal EEG dataset used for the analysis of emotional affect states. The dataset contains multichannel EEG recordings of 32 participants. Each individual watched a music video for a duration of 1 minute to elicit a desired emotional response. A 3 second baseline was established before and after each video trial. Each individual participated in 40 video trials. The participants gave ratings of their valence, arousal and dominance levels using Self-Assessment Manikins (SAM) [1]. The EEG signals were recorded at a sampling rate of 128 Hz using 32 active AgCl electrodes placed per the international 10-20 system.

Two sets of Semi-Simulated EEG signal Databases were generated using the DEAP dataset. For the first set of Semi-Simulated EEG signal dataset the non-brain artifacts were modeled making use of several biological information and added to the artifact free EEG dataset using the procedure outlined in Chapter III. 40 additional EEG signal data trials were generated for each individual and class labels were replicated accordingly as non-cerebral artifacts and noise should not affect the emotional response of a person. For the second set of Semi-Simulated EEG signal dataset only random noise was modeled and summed with artifact free EEG signals. The time points to which the noise was added was chosen randomly and the channels to which the noise was added was the same as the former Semi-Simulated dataset for fair comparison.

Another dataset considered in this work is the MAHNOB – HCI dataset [2]. The experimental setup for collecting the EEG signals is similar to the DEAP database. MAHNOB- HCI [2] database contains physiological signals and emotion class labels with the intent of creating explicit and implicit video emotion tagging systems. Each participant watched a movie clip lasting for around 90 seconds which was intended to elicit an emotional response. The MAHNOB- HCI database [2] contains EEG signals recorded from 32 channels for a total of 538 sessions. Around 27 participants contributed to this dataset, and each member was involved with around 20 recordings. The multichannel EEG segments were acquired using a Bio-Semi setup similar to the DEAP with 32 channels and a sampling frequency of 256 Hz.

The dataset organized by the University of California, San Diego (UCSD) Imagined Emotion Study Dataset [3] is a well-organized database. It also contains multichannel EEG recordings from 32 individuals. However, the experimental environment for recording the EEG signals is different when compared to the DEAP database [1] and the MAHNOB-HCI database [2]. The participant was first made to relax. The participant was later guided to imagine various emotional scenarios or recall some emotional experiences based on narrations. Each subject was asked to imagine around 15 emotional events. The EEG recording was performed with 256 active AgCl electrodes placed per the international 10-20 system at a sampling frequency of 256 Hz.

5.2 Experimental Results

All the programs for BCI paradigms including the spectral and the spatial filters, feature extraction from EEG datasets, SVM, feature selection were composed and implemented in Matlab (The Mathworks Inc., United States). In this section, we will discuss the experimental results obtained by applying the proposed BCI paradigms with the DEAP dataset [1], the UCSD Imagined Emotions Dataset [3] and the MAHNOB-HCI Database [2]. The proposed BCI paradigms are also compared with the existing prior work on that database.

Tables 2, 3 and 4 give the results of the proposed BCI paradigms on the DEAP dataset for recognizing valence, arousal and dominance respectively. The results have been compared

with existing methodologies used with DEAP dataset. The evaluation of the models was performed using Leave-One Out Cross-Validation (LOOCV). The average of LOOCV for each subject is reported on the table. Same evaluation metric was used by [1], [2]. [13] and [14].

In the Semi-Simulated EEG databases, the number of trials for each subject is higher (80) than the DEAP [1] database. A Semi-Simulated dataset contains an additional 40 trials for each subject which makes a total of 80 trials for a subject. Hence, the evaluation metric we have used was five-Fold Cross Validation (5-Fold CV). The performance of two Semi-Simulated EEG databases have been compared. The Tables 5, 7 and 9 give the affect recognition scores on the dataset generated by modeling non-brain artifacts using biological information and statistics outlined in Chapter III. The Tables 6, 8 and 10 summarize the affect recognition accuracies using the database obtained by summing up random noise with artifact free EEG signals. It should be noted that the testing was performed only on the artifact free EEG signal portion of the dataset.

Reference Method	Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Valence (Average over 32 subjects with LOOCV on each subject)	
				Average Accuracy	# subjects > 65% accuracy
	PSD averaged over seven bands and variance of DWT coefficients with mRMR MID feature selection(d=150)	SVM with linear kernel	150	71.33 %	27 subjects
	PSD averaged over seven bands for each channel and variance of DWT coefficients with mRMR MIQ feature selection (d = 130)	SVM with linear kernel	130	70.55 %	25 subjects
	PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	66.87%	20 subjects
	PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	70.23%	26 subjects
	Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	65.70%	21 subjects
	Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	69.60%	29 subjects
[1]	PSD from five bands with asymmetrical channel differences	Naïve Bayes	216	57.60 %	
[13]	PSD from five bands with asymmetrical channel differences	Hierarchical Bayesian Network	216	58.00 %	
[14]	PSD from five bands with asymmetrical differences and Group information of subjects as privileged information	Triangle shaped Bayesian Network	Around 160 + subject info	60.36%	

Table 2: Results Tabulated for Recognizing Valence from DEAP Dataset

Reference Method	Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Arousal (Average over 32 subjects with LOOCV on each subject)	
				Average Accuracy	# subjects > 65% accuracy
	PSD average over seven frequency bands and variance of DWT coefficients with mRMR MID feature selection (d= 150)	SVM with linear kernel	150	68.92 %	19 subjects
	PSD averaged over seven bands for each channel and variance of DWT coefficients with mRMR MIQ feature selection (d= 150)	SVM with linear kernel	150	67.73 %	18 subjects
	PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	68.59%	22 subjects
	PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	69.37 %	21 subjects
	Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	65	66.94%	15 subjects
	Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	67.96 %	19 subjects
[1]	PSD from five bands with asymmetrical channel differences	Naïve Bayes	216	62.00%	
[13]	PSD from five bands with asymmetrical channel differences	Hierarchical Bayesian Network	216	58.40 %	
[14]	PSD from five bands with asymmetrical differences and Group information of subjects as privileged information	V shaped Bayesian Network	Around 160 + subject info	65. 63%	

Table 3: Results Tabulated for Recognizing Arousal from DEAP Dataset

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Dominance (Average over 32 subjects with LOOCV on each subject)	
			Average Accuracy	# subjects > 65% accuracy
PSD averaged over seven frequency bands and variance of DWT coefficients with mRMR MID feature selection (d = 120)	SVM with linear kernel	120	69.17 %	20 subjects
PSD averaged over seven bands for each channel and variance of DWT coefficients with mRMR MIQ feature selection (d= 130)	SVM with linear kernel	130	67.60 %	18 subjects
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	67.97 %	15 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	69.45 %	18 subjects
Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	65.56%	16 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	69.76%	18 subjects

Table 4: Results Tabulated for Recognizing Dominance from DEAP Dataset

The authors of [1], [13] and [14] do not perform classification on the dominance emotional affect class labels.

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Valence (Average Accuracy of 32 subjects with 5-Fold CV on each subject)	
			Average Accuracy	# subjects > 65% accuracy
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	87.50 %	30 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	84.80 %	30 subjects
Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	75.42 %	24 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	76.76%	30 subjects

Table 5: Recognizing Valence from Biological Artifact Semi- Simulated EEG Dataset

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Valence (Average Accuracy of 32 subjects with 5-Fold CV on each subject)	
			Average Accuracy	# subjects > 65 % accuracy
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	81.72 %	30 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	76.52%	29 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with Gaussian Kernel	64	71.60%	25 subjects

Table 6: Recognizing Valence from Random Noise Artifact Semi- Simulated EEG Dataset

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Arousal (Average Accuracy of 32 subjects with 5-Fold CV on each subject)	
			Average Accuracy	# subjects > 65% accuracy
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	88.90%	30 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	81.58 %	30 subjects
Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	71.59 %	25 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	74.49 %	27 subjects

Table 7: Recognizing Arousal from Biological Artifact Semi- Simulated EEG Dataset

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Arousal (Average Accuracy of 32 subjects with 5-Fold CV on each subject)	
			Average Accuracy	# subjects > 65 % accuracy
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	80.98 %	30 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	73.32 %	28 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	71.13 %	24 subjects

Table 8: Recognizing Arousal from Random Noise Artifact Semi- Simulated EEG Dataset

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Dominance (Average Accuracy of 32 subjects with 5-Fold CV on each subject)	
			Average Accuracy	# subjects > 65 % accuracy
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	88.55 %	29 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	80.34 %	29 subjects
Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	72.71 %	23 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	76.09%	30 subjects

Table 9: Recognizing Dominance from Biological Artifact Semi- Simulated EEG Dataset

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Dominance (Average Accuracy of 32 subjects with 5-Fold CV on each subject)	
			Average Accuracy	# subjects > 65 % accuracy
PSD with respect to frequency bins in the range [0.4 Hz – 64 Hz]	SVM with linear kernel	65	81.75%	30 subjects
PSD averaged over seven frequency bands for each channel	SVM with linear kernel	224	75.82%	28 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	71.99%	23 subjects

Table 10: Recognizing Dominance from Random Noise Semi- Simulated EEG Dataset

Table 11 contains the classification accuracies computed while recognizing Valence tags from the UCSD Imagine Emotions Dataset [3]. The authors of [9] perform valence affect BCI recognition using Filter Bank Common Spatial Patterns on this database. However, they report the average five-fold cross-validation accuracy taken over only 12 subjects of the 32 participants in the dataset. We have reported the average five-fold cross-validation taken over all the 32 participants. In order to be able to compare the techniques properly, the average accuracy over all the subjects that performed better than 65% recognition rate have also been reported.

Reference Method	Feature Extraction Technique	Classifier	Feature dimension	Evaluated using 5 -fold CV on each subject for Valence Label
	PSD with respect to frequency for the range [0.3 Hz – 69 Hz]	SVM with linear kernel	70	Accuracy over 32 subjects = 62.12 % (14 subjects > 65%) Accuracy over the 14 subjects = 75.96%
	Variance of DWT coefficients for 5 levels of decomposition (db 5)	SVM with linear kernel	281	Accuracy over 32 subjects = 62.42 % (18 subjects > 65 %) Accuracy over the 18 subjects = 74.11 %
	Spectral filter of [0.4 Hz – 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	Accuracy over 32 subjects = 59.96 % (12 subjects > 65 %) Accuracy over the 12 subjects = 70.56 %
[9]	Spectral Filter bank of 5 spectral filters and 8 spatial filters	Generalized Linear Model with a logistic link function	40	Accuracy over 12 subjects = 71.3 %

Table 11: Recognizing Valence from UCSD Imagined Emotions Database

Reference Method	Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Valence (Average over 27 subjects with LOOCV on each subject)	
				Average Accuracy	# subjects > 65 % accuracy
	PSD with respect to frequency bins in the range [0.3 Hz – 63 Hz]	SVM with linear kernel	64	67.06%	17 subjects
	PSD averaged over six frequency bands for each channel	SVM with linear kernel	192	68.78%	17 subjects
	Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	62.50 %	15 subjects
	Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	66.94 %	16 subjects
[2]	PSD from five bands with asymmetrical channel differences	SVM with Gaussian kernel	216	57.00 %	
[13]	PSD from five bands with asymmetrical channel differences	Hierarchical Bayesian Network	216	56.90 %	
[14]	PSD from five bands with asymmetrical differences and Group information of subjects as privileged information	V shaped Bayesian Network	Around 160 + subject info	62.85%	

Table 12: Recognizing Valence from MAHNOB-HCI Database

Tables 12, 13 and 14 summarize the results obtained with the MAHNOB-HCI datasets for recognizing valence, arousal and dominance. Comparisons with the existing methods have also been reported.

Reference Method	Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Arousal (Average over 27 subjects with LOOCV on each subject)	
				Average Accuracy	# subjects > 65 % accuracy
	PSD with respect to frequency bins in the range [0.3 Hz – 63 Hz]	SVM with linear kernel	64	73.11%	20 subjects
	PSD averaged over six frequency bands for each channel	SVM with linear kernel	192	72.06 %	20 subjects
	Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	63.91 %	12 subjects
	Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	74.06 %	21 subjects
[2]	PSD from five bands with asymmetrical channel differences	SVM with Gaussian kernel	216	52.40 %	
[13]	PSD from five bands with asymmetrical channel differences	Hierarchical Bayesian Network	216	63.00 %	
[14]	PSD from five bands with asymmetrical differences and Group information of subjects as privileged information	Triangle shaped Bayesian Network	Around 160 + subject info	68.48 %	

Table 13: Recognizing Arousal from MAHNOB-HCI Database

Feature Extraction Technique	Classifier	Feature dimension	Affect Recognized Dominance (Average over 27 subjects with LOOCV on each subject)	
			Average Accuracy	# subjects > 65 % accuracy
PSD with respect to frequency bins in the range [0.3 Hz – 63 Hz]	SVM with linear kernel	64	69.11 %	16 subjects
PSD averaged over six frequency bands for each channel	SVM with linear kernel	192	72.00 %	19 subjects
Common Spatial Patterns with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	SVM with Gaussian kernel	8	69.98 %	16 subjects
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	SVM with linear kernel	64	74.11 %	20 subjects

Table 14: Recognizing Dominance from MAHNOB-HCI Database

The authors of [2], [13] and [14] do not perform recognition on the dominance affect class labels for the MAHNOB-HCI Dataset.

5.3 Experimental Discussions

In this section, we will compare the performance of the Gaussian and linear kernel of SVM on the DEAP dataset and analyze them. Tables 15, 16 and 17 show the evaluation of the SVM models with average LOOCV accuracies of 32 subjects for valence, arousal and dominance respectively. The number of subjects with which we obtain greater than 65% recognition rate is also recorded. This number can give a good indication of how well the model captures the affect label on an individual chosen from a given group of participants. The dimensions of the feature vector used in each BCI paradigm is also noted.

Feature Extraction Technique	Feature dimension	Linear Kernel SVM model evaluation		Gaussian Kernel SVM model evaluation	
Valence Affect Label		Avg. Accuracy with LOOCV	#subjects > 65% accuracy	Avg. Accuracy with LOOCV	#subjects > 65% accuracy
PSD wrt frequency bins in the range [0.4 Hz – 64 Hz]	65	66.873 %	20	63.218 %	17
PSD averaged over 7 frequency bands for each channel	224	70.232 %	26	64.938%	19
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	64	69.602 %	29	64.609 %	17
CSP with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	8	63.840 %	16	65.702 %	21

Table 15: Comparison of Recognition Scores with Linear and Gaussian Kernel SVM for Valence with DEAP Dataset

Feature Extraction Technique	Feature dimension	Linear Kernel SVM model evaluation		Gaussian Kernel SVM model evaluation	
Arousal Affect Label		Avg. Accuracy with LOOCV	#subjects > 65% accuracy	Avg. Accuracy with LOOCV	#subjects > 65% accuracy
PSD wrt frequency bins in the range [0.4 Hz – 64 Hz]	65	68.596 %	22	66.561%	16
PSD averaged over 7 frequency bands for each channel	224	69.374 %	21	66.723%	17
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	64	67.962 %	19	66.953 %	17
CSP with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	8	63.315 %	14	66.942%	15

Table 16: Comparison of Recognition Scores with Linear and Gaussian Kernel SVM for Arousal with DEAP Dataset

Feature Extraction Technique	Feature dimension	Linear Kernel SVM model evaluation		Gaussian Kernel SVM model evaluation	
Dominance Affect Label		Avg. Accuracy with LOOCV	#subjects > 65% accuracy	Avg. Accuracy with LOOCV	#subjects > 65% accuracy
PSD wrt frequency bins in the range [0.4 Hz – 64 Hz]	65	67.972%	15	65.096 %	13
PSD averaged over 7 frequency bands for each channel	224	69.454%	18	68.360%	16
Variance of DWT coefficients along time with 5 levels of decomposition (db 5)	64	69.765%	18	66.403%	13
CSP with spectral filter of [0.4 Hz- 54 Hz] and 8 spatial filters	8	63.062%	14	65.568%	16

Table 17: Comparison of Recognition Scores with Linear and Gaussian Kernel SVM for Dominance with DEAP Dataset

From Tables 15, 16 and 17 we notice that SVM models with linear kernel give better recognition rates than the Gaussian Kernel SVM when using Power Spectral Density and wavelet features. We also notice that SVM models with Gaussian kernel give better recognition accuracy than linear kernel SVMs when dealing with Common Spatial Patterns. Both the average LOOCV accuracies and the number of subjects for which greater than 65% recognition rate was obtained indicate this outcome.

One can observe that the linear kernel SVM is optimum when the dimensions of the feature vectors are large when compared to the number of instances present in the dataset. Since we have used the DEAP [1] database in this section, we have only 40 trials for each subject. The Gaussian kernel SVMs are prone to overfitting in this scenario when the feature dimension is larger than the number of trial examples in the data. In other words, it is simpler to draw a line that separates the two classes when dealing with very high dimensions and less examples. Hence, linear kernel SVMs obtain a better recognition score compared to the Gaussian kernel SVMs in the given scenario (feature dimension > number of instances).

5.4 Inferences

In this thesis, we have presented EEG based BCI-paradigms to recognize emotions from multichannel EEG segments. Principal aspects of emotional affect, namely, valence, arousal and dominance can be classified into positive and negative labels using a few seconds of EEG signals. One of the key results is that the emotional affect labels can be recognized with better than chance levels (50%). It is important to note that these results hold up under the cross-validation evaluation techniques. Furthermore, one should also consider that the segments used during testing need not reflect the exact same emotional scenarios for which the model was trained. For example, love and happiness are distinct emotional experiences, but both have positive valence labels. From this one can infer that the features proposed in this work along with the combination of SVM models we are able

to capture the patterns in multichannel EEG segments present in different discrete emotional scenarios but having the same affect label.

From the Tables in Section 5.2 we see that including high frequency bands and information as features for the SVM models improved the recognition rate of affect labels when compared to the existing techniques [1], [2], [13] and [14] which did not include the high frequency content of EEG signals. Previous works by [3] have also concluded that there is a link between high frequency band EEG activity and emotional valence.

The PSD features averaged over seven frequency bands for each channel computed for all the channels performs better than the PSD features summed over all the channels with respect to frequency for the DEAP dataset and the MAHNOB-HCI dataset. This is mostly because the former represents frequency and spatial information of the EEG signals whereas the latter only represent the frequency content in the EEG signal.

The Common Spatial Patterns (CSP) [12] is considered as the state of the art paradigm for recognition in motor imagery task. The CSP algorithm gives a good generalization accuracy on emotional affect recognition as well. Although, it does not perform as good as the PSD features and the discrete wavelet features, the CSP paradigm can be regarded as a good baseline performance for a given EEG database.

The performance of our proposed Paradigms increased with both types of Semi-Simulated EEG Datasets. This can be inferred due to the higher amount of EEG trial data present for each individual participant. As emotion is more of a personal feeling, it varies from person to person. Hence, classifying the emotional affective states for each person by collecting more data from that individual gives the highest advantage to the SVM classifier model. The increase in the recognition score using biological non-brain artifact Semi-Simulated EEG database was more compared to the random noise Semi-Simulated EEG signal data.

The mRMR feature section algorithm helps to identify the redundant and relevant concatenated features in order to acquire a classifier model which fuses the relevant information from different feature spaces in an efficient manner. The recognition accuracy increased slightly using the mRMR feature selection. The proposed paradigms are simple and practically usable for EEG based BCI emotion recognition systems. The extracted features using power spectral information and the discrete wavelet transformation coefficients with the combination of SVM classifier models capture the key aspects required to recognize emotional affect.

CHAPTER VI

CONCLUSIONS

In conclusion, this work investigated recognizing the principal dimensions of affect – valence, arousal, and dominance from multichannel EEG segments using paradigms comprising of signal processing and machine learning techniques. The paradigms were designed by extracting distinctive features from the EEG signals that represent the Power Spectral Density, Discrete Wavelet Coefficients and Spatial Projections and constructing SVM Classifier models from these features. The paradigms when evaluated using three benchmark databases, namely, the DEAP [1] dataset, UCSD-Imagined Emotions Database [3] and the MAHNOB-HCI [2] Database using cross-validation methods gave consistently good testing accuracy. These evaluations also outperformed the existing paradigms for affect recognition described in [1], [2], [9], [13] and [14]. By employing mRMR feature selection on the combination of Power Spectral Density features and DWT coefficients a subset of complimentary features was selected which slightly improved the generalized recognition score.

The performance of the designed paradigms was also compared with two sets of Semi-Simulated EEG signal data. It was observed that the performance of the SVM models on the set of Semi-Simulated EEG signal data modeled using non-brain bio-artifacts was better compared to the Semi-Simulated EEG signal data generated by applying random noise.

Incorporating features from different modalities like Galvanic Skin Resistance, eye gaze tracking, skin temperature, respiration rate, electrocardiogram and video recording of the participants is one of the challenging future works in this field. Another possible future contribution to this work can include recognizing and predicting the brain source signals relevant to the affect labels using Blind Source Separation techniques like Independent Component Analysis.

REFERENCES

[1] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras. DEAP: A Database for Emotion Analysis using Physiological Signals. IEEE Transaction on Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation. 2012.

[2] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic. A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing. 3: pp. 42 - 55, Issue 1. April 2012.

[3] Onton J, Makeig S. High-frequency broadband modulations of electroencephalographic spectra. Frontiers in Neuroscience 159: 99-120. 2009.

[4] Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4), 261-292. 1996.

[5] Sourina, O., & Liu, Y. A Fractal-based Algorithm of Emotion Recognition from EEG using Arousal-Valence Model. In BIOSIGNALS (pp. 209-214). January 2011.

[6] Shaoda Yu, Peng Li, Honghuang Lin, Ehsan Rohani, Gwan Choi, Botang Shao and Qian Wang. Support vector machine based detection of drowsiness using minimum EEG features. ASE/IEEE Intl. Conf. on Biomedical Computing, pp. 827-835. September 2013.

[7] Murugappan, M., Ramachandran, N., & Sazali, Y. Classification of human emotion from EEG using Discrete Wavelet Transform. Journal of Biomedical Science and Engineering, 3(04), 390. 2010.

[8] Robert Jenke, Angelika Peer, Martin Buss. Feature Extraction and Selection for Emotion Recognition from EEG. IEEE Transactions on Affective Computing. 2014.

[9] C Kothe, Onton JA, S Makeig. Emotion recognition from EEG during self-paced emotional imagery. 3rd Workshop of Affective Brain-Computer Interfaces, Geneva, Switzerland. 2013.

[10] Rosalind W. Picard. Affective Computing. The MIT Press. 2000.

[11] P. D. Welch. The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics, AU-15 (2): 70–73. 1967.

[12] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K. R. Müller. Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. IEEE Signal Processing Magazine, 25(1):41-56. 2008.

[13] Zhen Gao and Shangfei Wang. Emotion recognition from EEG signals using hierarchical Bayesian network with privileged information. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pages 579–582, ACM. 2015.

[14] Wu Shan, Shangfei Wang, Yachen Zhu, Zhen Gao, Lihua Yue and Qiang Ji. Employing Subjects Information as Privileged Information for Emotion Recognition from EEG Signals, International Conference on Pattern Recognition (ICPR). 2016.

[15] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238. 2005.

[16] Chris Ding, and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. Proc. 2nd IEEE Computational Systems Bioinformatics Conference (CSB 2003), pp.523-528, Stanford, CA. August 2003.

[17] Shiyu Chen, Shangfei Wang, Chongliang Wu, Zhen Gao, Xiaoxiao Shi and Qiang Ji. Implicit Hybrid Video Emotion Tagging by Integrating Video Content and Users Multiple Physiological Responses, International Conference on Pattern Recognition (ICPR). 2016.

[18] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. 2011.

[19] H. Ramoser, J. Muller-Gerking and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. in IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 4, pp. 441-446. December 2000.

[20] L. Yu and H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, vol. 5. 2004.

[21] I. Daubechies. Ten lectures on wavelets, CBMS-NSF conference series in applied mathematics. SIAM Ed. 1992.

[22] Kevin Patrick Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press. August 2012.

[23] A. Delorme, T. J. Sejnowski, S. Makeig. Improved rejection of artifacts from EEG data using high-order statistics and independent component analysis. NeuroImage. 2007.

[24] Mark H. Libenson. Practical Approach to Electroencephalography. Saunders. December 2009.

[25] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh. Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing. Springer 2006.

[26] McClellan, James, Parks, Thomas. A Personal History of the Parks–McClellan Algorithm. IEEE Xplore retrieved. 2009.

[27] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Advances in kernel methods, Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.). MIT Press, Cambridge, MA, USA 185-208. 1999.

[28] R. Picard. Affective Computing. MIT Technical Report. 1995.

[29] 10-20 system (EEG) – Wikipedia.

[30] Guido Dornhege, José del R. Millán, Thilo Hinterberger, Dennis J. McFarland, Klaus-Robert Müller. Toward Brain-Computer Interfacing. The MIT Press. July 2007.