Applications of non-invasive brain-computer interfaces for communication and affect recognition

by

Md Rakibul Mowla

B.S., Chittagong University of Engineering & Technology, Bangladesh, 2010

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Mike Wiegers Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Abstract

Various assistive technologies are available for people with communication disorders. While these technologies are quite useful for moderate to severe movement impairments, certain progressive diseases can cause a total locked-in state (TLIS). These conditions include amyotrophic lateral sclerosis (ALS), neuromuscular disease (NMD), and several other disorders that can cause impairment between the neural pathways and the muscles. For people in a locked-in state (LIS), brain-computer interfaces (BCIs) may be the only possible solution. BCIs could help to restore communication to these people, with the help of external devices and neural recordings.

The present dissertation investigates the role of latency jitter on BCIs system performance and, at the same time, the possibility of affect recognition using BCIs. BCIs that can recognize human affect are referred to as affective brain-computer interfaces (aBCIs). These aBCIs are a relatively new area of research in affective computing. Estimation of affective states can improve human-computer interaction as well as improve the care of people with severe disabilities. The present work used a publicly available dataset as well as a dataset collected at the Brain and Body Sensing Lab at K-State to assess the effectiveness of EEG recordings in recognizing affective states.

This work proposed an extended classifier-based latency estimation (CBLE) method using sparse autoencoders (SAE) to investigate the role of latency jitter on BCI system performance. The recent emergence of autoencoders motivated the present work to develop an SAE based CBLE method. Here, the newly-developed SAE-based CBLE method is applied to a newly-collected dataset. Results from our data showed a significant ($p < 0.001$) negative correlation between BCI accuracy and estimated latency jitter. Furthermore, the SAE-based CBLE method is also able to predict BCI accuracy.

In the aBCI-related investigation, this work explored the effectiveness of different features extracted from EEG to identify the affect of a user who was experiencing affective stimuli. Furthermore, this dissertation reviewed articles that used the Database for Emotion Analysis Using Physiological Signals (DEAP) (i.e., a publicly available affective database) and found that a significant number of studies did not consider the presence of the class imbalance in the dataset. Failing to consider class imbalance creates misleading results. Furthermore, ignoring class imbalance makes comparing results between studies impossible, since different datasets will have different class imbalances. Class imbalance also shifts the chance level. Hence, it is vital to consider class bias while determining if the results are above chance. This dissertation suggests the use of balanced accuracy as a performance metric and its posterior distribution for computing confidence intervals to account for the effect of class imbalance.

Applications of non-invasive brain-computer interfaces for communication and affect recognition

by

Md Rakibul Mowla

B.S., Chittagong University of Engineering & Technology, Bangladesh, 2010

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Mike Wiegers Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Approved by:

Major Professor
David E. Thompson

Copyright

Abstract

Various assistive technologies are available for people with communication disorders. While these technologies are quite useful for moderate to severe movement impairments, certain progressive diseases can cause a total locked-in state (TLIS). These conditions include amyotrophic lateral sclerosis (ALS), neuromuscular disease (NMD), and several other disorders that can cause impairment between the neural pathways and the muscles. For people in a locked-in state (LIS), brain-computer interfaces (BCIs) may be the only possible solution. BCIs could help to restore communication to these people, with the help of external devices and neural recordings.

The present dissertation investigates the role of latency jitter on BCIs system performance and, at the same time, the possibility of affect recognition using BCIs. BCIs that can recognize human affect are referred to as affective brain-computer interfaces (aBCIs). These aBCIs are a relatively new area of research in affective computing. Estimation of affective states can improve human-computer interaction as well as improve the care of people with severe disabilities. The present work used a publicly available dataset as well as a dataset collected at the Brain and Body Sensing Lab at K-State to assess the effectiveness of EEG recordings in recognizing affective states.

This work proposed an extended classifier-based latency estimation (CBLE) method using sparse autoencoders (SAE) to investigate the role of latency jitter on BCI system performance. The recent emergence of autoencoders motivated the present work to develop an SAE based CBLE method. Here, the newly-developed SAE-based CBLE method is applied to a newly-collected dataset. Results from our data showed a significant ($p < 0.001$) negative correlation between BCI accuracy and estimated latency jitter. Furthermore, the SAE-based CBLE method is also able to predict BCI accuracy.

In the aBCI-related investigation, this work explored the effectiveness of different features extracted from EEG to identify the affect of a user who was experiencing affective stimuli. Furthermore, this dissertation reviewed articles that used the Database for Emotion Analysis Using Physiological Signals (DEAP) (i.e., a publicly available affective database) and found that a significant number of studies did not consider the presence of the class imbalance in the dataset. Failing to consider class imbalance creates misleading results. Furthermore, ignoring class imbalance makes comparing results between studies impossible, since different datasets will have different class imbalances. Class imbalance also shifts the chance level. Hence, it is vital to consider class bias while determining if the results are above chance. This dissertation suggests the use of balanced accuracy as a performance metric and its posterior distribution for computing confidence intervals to account for the effect of class imbalance.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The work presented in this dissertation is the outcome of five years of research, but the effort and dedication required to reach this stage are immeasurable. During this course of time, many people have had close interaction with me, without whose support I would not have pursued a doctoral degree.

I want to take this opportunity to express my deep gratitude to *Md. Intekhab Alam*, *Homayun Kabir*, *Dr. Syed Moshfeq Salaken*, and *Dr. Md. Tanvir Hasan* for their continuous support and motivation to start my career as a researcher. I would also like to show my gratitude to *Dr. Siew-Cheok Ng* from the University of Malaya, Malaysia, who introduced me the biomedical signal processing methods and electroencephalography (EEG) data.

My parents, especially my mother and sister, encouraged me in every step of my academic endeavor. I am thankful to them. My wife, who supported me during the dissertation writing period, also deserves to be mentioned here.

I feel very fortunate to have *Dr. David E. Thompson* as my major advisor. I want to extend my deepest gratitude to him for allowing me to work with him. His support and mentorship gave me the courage to complete this dissertation. It was always very gratifying to discuss with him about research and, at the same time, out-of-research topics.

Rachael I. Cano and *Katie J. Dhuyvetter* ran most of the EEG sessions for the affective experiment, while *Jesus D Gonzalez-Morales, Jacob Rico-Martinez*, and *Daniel A. Ulichnie* ran most of the EEG sessions for the P3 speller experiment. All of their names deserved to be mentioned here, without a doubt. My labmate, *Dr. Ahmad Suliman*, was also a great support for me here at the BBS lab.

I also want to thank my doctoral committee members, especially *Dr. Bala Natarajan* and *Dr. Dwight D. Day* for their support in the past few years. I am also thankful to the Mike Wiegers Department of Electrical and Computer Engineering for providing financial support through teaching assistantships.

The involvement of human participants with the P3 Speller experiment was approved by the Kansas State University Institutional Review Board under protocol No. 8320. The affective BCI research was approved by the Kansas State University Institutional Review Board under protocol No. 8328.

This material is based upon work supported in part by the National Science Foundation under Award No. 1910526 and in part by Kansas State University faculty startup funds. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the author and do not necessarily reflect the views of the funding agencies.

Dedication

To my parents for their endless love and encouragement.

To my wife for bearing with me during the writing phase.

To my friends, those who supported me during my struggles.

Preface

This dissertation is comprised of manuscripts in various stages of submission and publication except, for the introduction and conclusion chapters. The text and figures in Chapter 2 are re-used with permission, and the permission is documented on the appendix A. Data used in chapter 2 were provided by the Direct Brain Interface lab of the University of Michigan led by Dr. Jane Huggins. Data used in chapter 3 & 4 were all collected at the Brain and Body Sensing lab of Kansas State University.

I have aimed to minimize the recurring and overlapping texts, but some passages have been used verbatim from the published works. While I made every effort to properly fit these published and submitted manuscripts with the story of this dissertation some repetition is unavoidable.

Chapter 1

Introduction

The increasing number of research articles on Brain-Computer Interfaces (BCIs) in the past few years manifests the potential of BCIs both in medical and non-medical applications. BCI technologies are now moving from the lab into industrial and commercial applications. However, these technologies are still in the initial stage, and require effort and skill to achieve accuracy and usability. In this dissertation, an attempt to further improve BCI performance will be made in two different applications. The first application is BCIs as a communication tool, and the second is BCIs as an affective measurement tool. In this chapter, a few basic definitions and terminologies related to BCI, and to this dissertation, are provided. This chapter also presents the motivation of this dissertation work and an introduction to the experimental environment.

1.1 Fundamentals

1.1.1 Electroencephalography (EEG)

The first electrical current variations and spontaneous current variations due to visual stimulation in rabbits and monkeys were reported by Caton (1875). Later these findings were confirmed in other independent studies by Beck and Cybulski (1891) for rabbits and dogs. But the most credit goes to Berger (1929) because of his first-ever report of recording human

EEG.

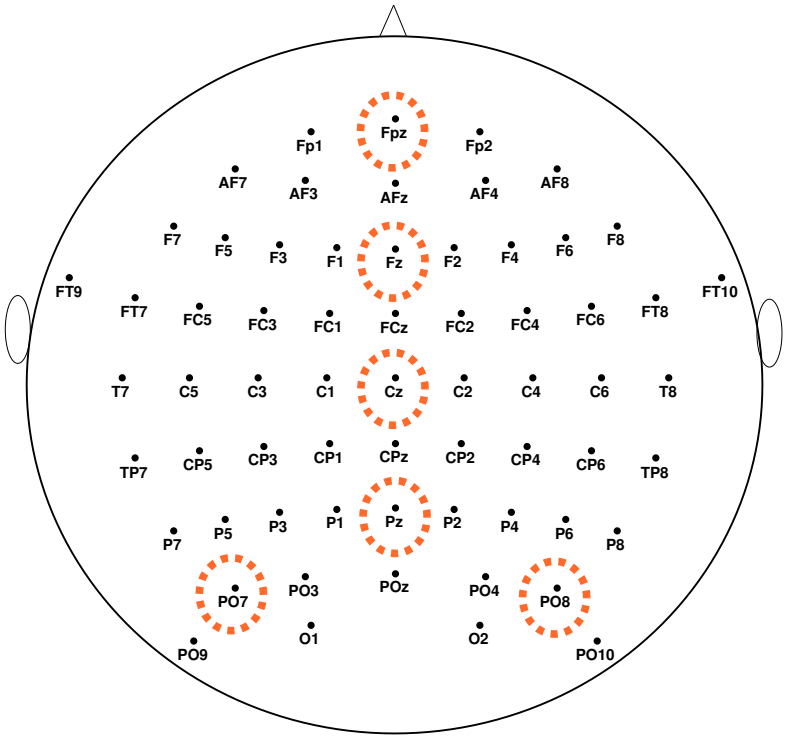


Figure 1.1: *This figure shows EEG electrode placement on the scalp used during EEG data recording. The orange-colored circle is used to indicate electrode positions used in Fig. 1.3.*

EEG records the electrical activity of large, synchronously firing populations of neuron using electrodes placed along the scalp (Niedermeyer and da Silva, 2005). EEG has many applications because of its non-invasive nature. Typically, EEG has a magnitude range of $10 - 50\mu V$. Electrode placement in the EEG cap usually follows the standard international 10/20 or 10/10 systems, approved by the American Clinical Neurophysiology Society (ACNS; former American Electroencephalographic Society). The international 10/20 system provides a method for placing a relatively small number of electrodes (typically 21). An extended 10/20 system and 10/10 system were introduced to facilitate a higher number of electrodes, and has been accepted as standard by ACNS (Society, 1994). The EEG data used in this dissertation were recorded using the extended 10/20 electrode placement system. The exact electrode placement map is shown in Fig. 1.1.

1.1.2 Brain-Computer Interfaces (BCIs)

BCIs use brain signals to provide a direct method of interaction with computers and other devices (Wolpaw et al., 2002) without using peripheral nerves and muscles. BCIs are also sometimes referred to as brain-machine interfaces (BMIs) (Tonet et al., 2008; Lee et al., 2009; Nicolas-Alonso and Gomez-Gil, 2012) in the literature.

Fig. 1.2 shows a general diagram of the BCI system. The first level, called the transducer (TR) level, includes signal processing and classification. The next level, which converts transducer output into a control output, is referred to as the control interface (CI) (Dal Seno et al., 2010; Thompson et al., 2014); for example, in a BCI for spelling, the task of the CI is to combine the classification outputs from the TR level to find the target character. The CI may also incorporate other assistive tools such as a word prediction program in the P300 speller BCI (Ryan et al., 2010). A third level, consisting of the human experience of using the BCI, has been suggested by Thompson et al. (2014).

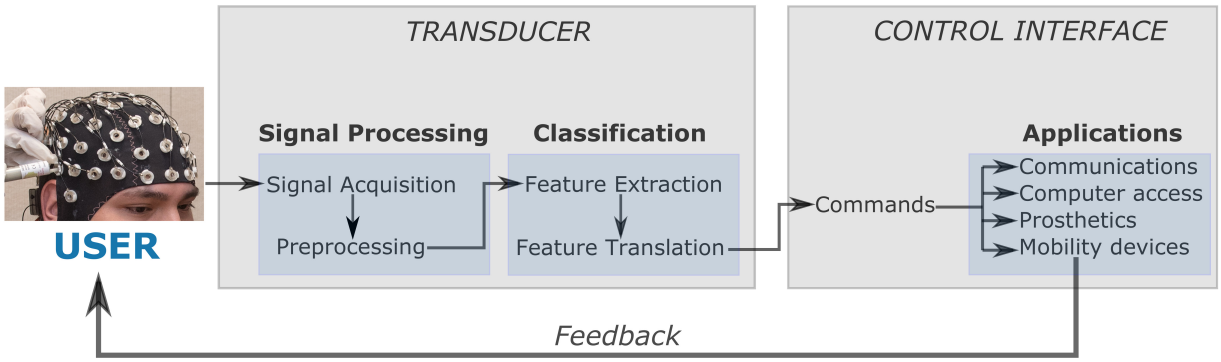


Figure 1.2: *A basic design diagram of a brain-computer interface system. The signals produced by the brain activity are recorded from the scalp using an EEG cap. The signals are preprocessed (e.g., bandpass filtering), and features are extracted (e.g., spectral density, brain rhythms, or raw EEG signals). Classifiers translate these features into certain predefined classes, and the CI uses those classes to translate into commands (e.g., a right-hand motor intent can be translated into "move the cursor right"). At the same time, the CI also sends feedback to the user so that the user can modulate his/her brain activity for better performance.*

P300 Event Related Potentials

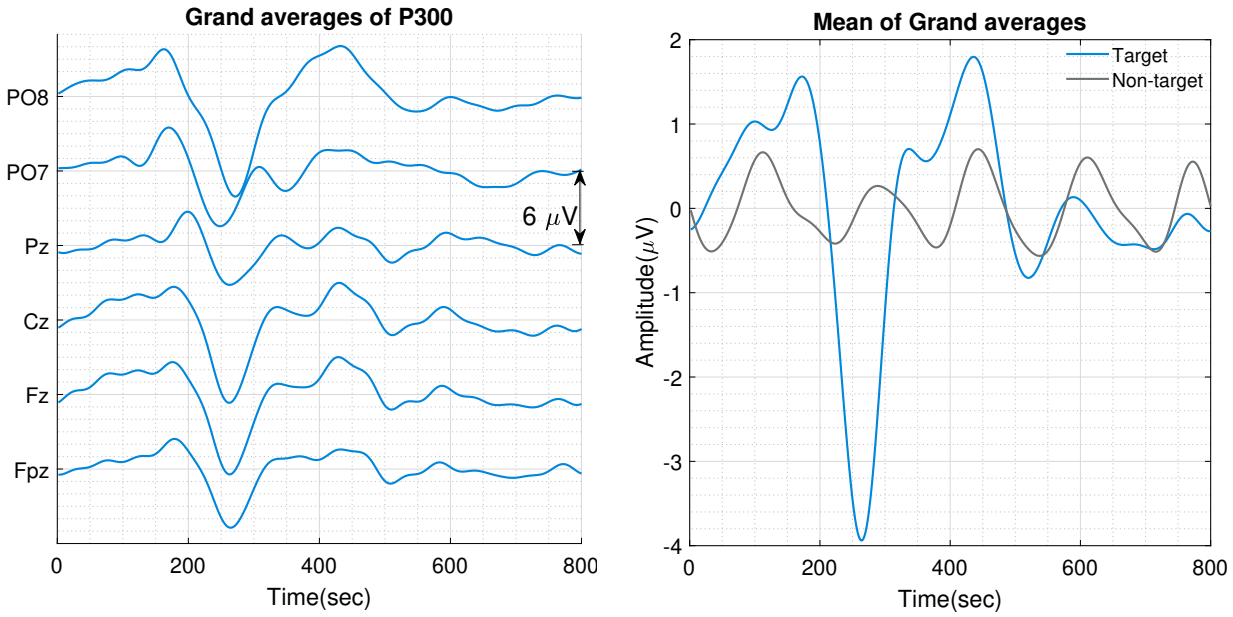
Event-related potentials (ERPs) are tiny voltage fluctuations in the brain's electrical activity in response to specific events (e.g., sensory, motor, or cognitive) or stimuli (Blackwood and Muir, 1990). BCIs for communication mostly uses a particular type of ERPs that are commonly known as P300 or P3 ERPs. P300 ERPs are a type of ERPs that are the brain responses elicited by rare stimuli, with a characteristic positive polarity approximately 300ms post-stimulus (Fabiani et al., 1987). However, the measured ERPs are mixed with unrelated brain activity, as well as interference from non-neuronal sources (eye-blinks, eye-movements, muscle movements) and instrumental noise. These factors lead to the well-known difficulty in recovering ERPs from single trials. The ERPs are buried under the background EEG signals (D'Avanzo et al., 2011) and that background EEG has a much larger amplitude than ERPs.

Therefore, ERPs have a very low signal-to-noise ratio (SNR) and, additionally, may contain stimulus artifacts caused by the repetitive presentation of visual stimuli (Martens et al., 2009). Hence, the P300 is hard to identify in single-trials. The most common P3 ERPs visualization method depends on averaging multiple trials. Averaging helps to obtain the P300 by suppressing the background EEG signals as the P300 is time-locked to the stimulus onset (Romero et al., 2015; Nelson and McCleery, 2008). Fig. 1.3 shows a typical P300 ERP at a few electrode locations averaged over 440 target trials.

P300 Speller

One of the most well-known types of BCI is the P300 or the P3 Speller introduced by Farwell and Donchin (1988), which uses the P300 ERPs.

In classical P3 Speller implementations, a grid matrix of 6×6 or more characters and commands are presented to the user. Columns and rows are highlighted/intensified in random order while the user focuses on the desired character. The probability of the intensified row or column containing the target character in one sequence of 12 flashes, six rows, and six columns, is 1/6. This low probability creates an oddball paradigm (Fabiani et al., 1987),



(a) Grand averages at few electrode locations.

(b) Mean response of grand averages.

Figure 1.3: *Example of typical P300 responses at a few selected electrode locations. The selected electrode locations are shown in Fig. 1.1 as orange colored circle. These responses were constructed using the average of 440 target trials.*

i.e., a rare event that will elicit a P300 response. A classical P300 speller display matrix implemented in BCI2000 (Schalk et al., 2004) is shown in Fig. 1.4.

Figure 1.4: 6×6 *grid matrix of a classical P3 speller display. This screen capture of a P300 speller display was captured in the BCI2000 environment (Schalk et al., 2004).*

1.2 Motivation

BCIs can help to restore communication for people with severe movement impairments such as amyotrophic lateral sclerosis (ALS), neuromuscular disease (NMD), brainstem stroke, cerebral palsy, and spinal cord injury (McFarland and Wolpaw, 2011). This kind of alternative form of communication would help to improve quality of life and may also reduce the cost of intensive care (Nicolas-Alonso and Gomez-Gil, 2012). These factors explain the increasing interest in such technologies. Fig. 1.5 shows the growing number of BCI related research articles from the year 2001 to 2019. The orange line in Fig. 1.5 indicates the number of BCI related articles while the blue line shows the number of EEG-based BCI articles indexed in Google Scholar.

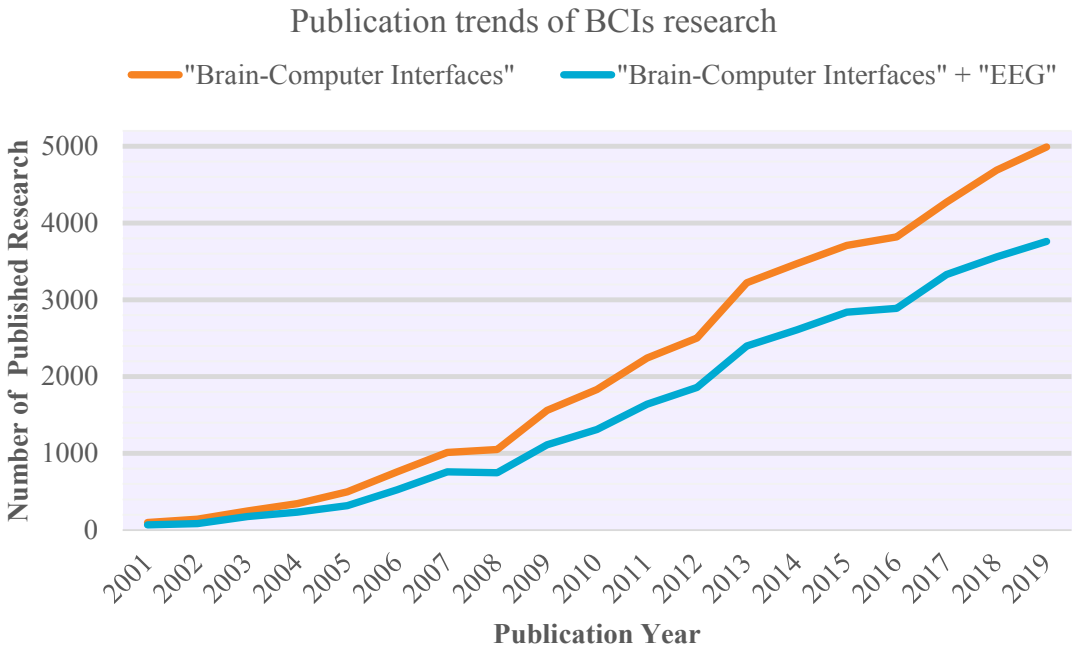


Figure 1.5: *Number of BCIs related publication. The graph shows number of publications from Google Scholar using "Brain-Computer Interfaces" and "Brain-Computer Interfaces" + "EEG" as keywords.*

This dissertation will attempt to explore two different applications of BCIs: (i) P3 speller for communication and (ii) affect (i.e., emotion) recognition. For the P3 speller part, this dissertation work will concentrate on the role of P3 latency variability on BCIs performance.

Spelling accuracies are plotted against added normally-distributed P3 latencies (i.e., simulated jitters) in Fig. 1.6. More details of this simulated latency work can be found in Thompson et al. (2019). Fig. 1.6 shows evidence of the relationship between P3 latency variations and spelling accuracy and thus motivated to further investigate the relationship.

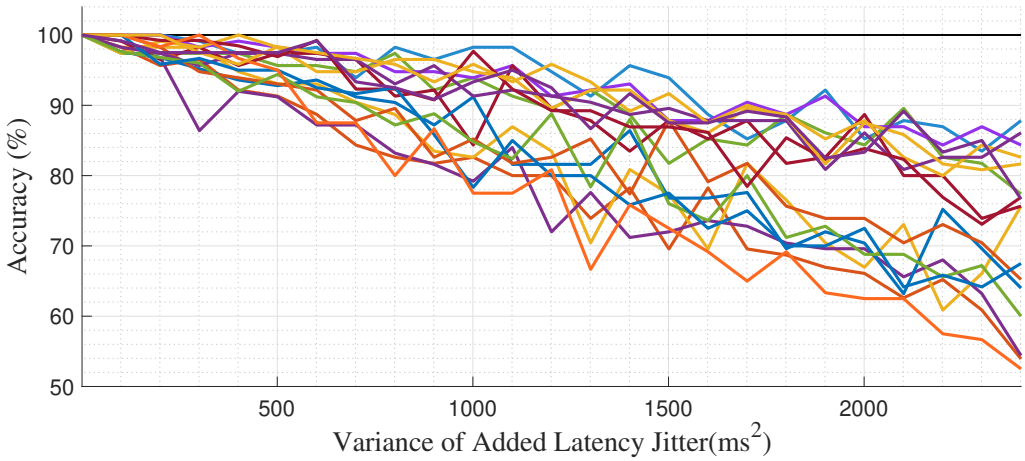


Figure 1.6: *Normally-distributed added latency jitter versus P3 spelling accuracy.*

1.2.1 P300 Speller Experiment

The P300 Speller (Farwell and Donchin, 1988) is one of the prevalent applications of BCIs, and can serve as a communication tool for patients with the diseases as mentioned earlier (Sellers and Donchin, 2006). Due to the very low SNR of P300 ERPs, a common technique of extracting P300 ERPs is based on signal averaging. However, the ERPs can vary in terms of latency and amplitude due to mental fatigue, stress, attention, and several other medical conditions (Boksem et al., 2005; Mowla et al., 2018a). In some cases averaging may be useful, but the averaging does not allow us to do single-trial analysis. Research has shown that P300 latency variability is strongly related to the cognitive function and provides a measure of cognitive health (Polich and Herbst, 2000). Hence, averaging causes the loss of important information related to P300 variability.

In this dissertation, the goal of the P300 speller experiment is to develop and validate a method of estimating single-trial P300 latency in order to calculate variability.

Single-trial estimation of the P300 will help to understand the underlying cognitive process of ERPs and also to improve the speed of BCI systems. In this experiment, data were collected while participants performed the copy-spelling task using BCI2000's (Schalk et al., 2004) row-column P300 speller paradigm. Each participant completed nine copy-spelling tasks in three sessions on different days. In each session, participants copy-spelled three sentences. Chapter 3 consists of further details on this experiment, including the sentences used and participants' demographics. Fig. 1.7 shows a participant performing the copy-spelling task of the P3 experiment.

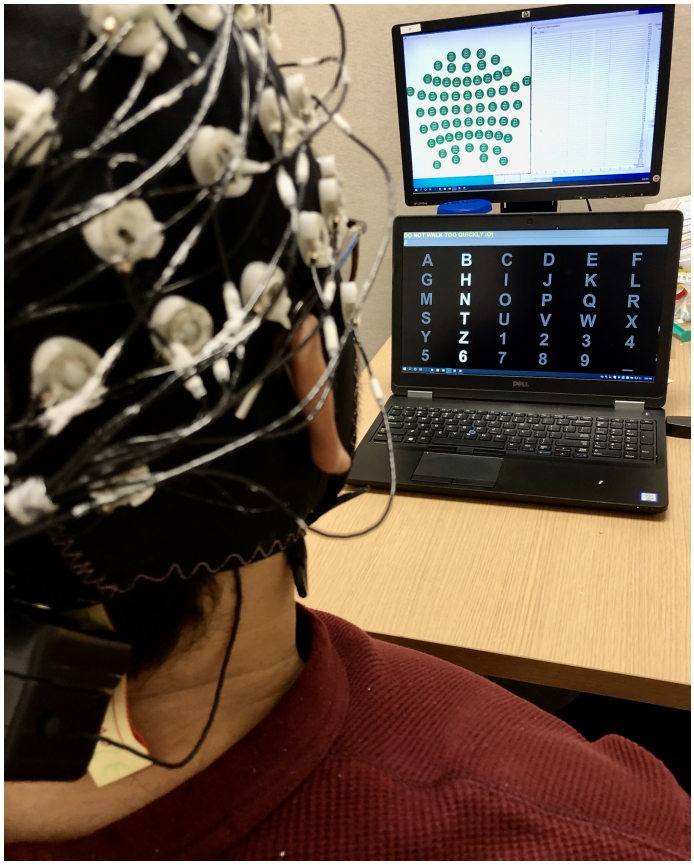


Figure 1.7: *Visualization of the experimental setup for P300 speller.*

1.2.2 Affective Computing Experiment

Understanding the human mind is a capability people would like to have, and is an active field of research. Studies have shown that affect (emotion), cognition, and decision-making

processes are interrelated with a complex network (Schwarz, 2000). Hence affective state has an essential role in human decision making process (Forgas, 1995). The ability to manage affective states is related to the skills of logical reasoning, learning, and extracting critical information (Salovey and Mayer, 1990). However, to achieve the ability to manage affective states requires an understanding of affective responses.

The first question that needs to be answered is "how can we understand affective responses?" The answer to this question is not straightforward, because affect involves several internal and external processes of the human body and organs. Rather, it is easier to answer a question like, "what are the physiological indicators that can be used to detect affective states?" Numerous studies have attempted to answer this question because it was a major topic of interest for affective computing researchers (Cacioppo and Tassinary, 1990). Several indicators can be used for affect recognition, such as pupil diameter (Oliva and Anikin, 2018), heart rate variability (Quintana et al., 2012), skin conductance (Nakasone et al., 2005), temperature (Levenson et al., 1992), voice tone (Petrushin, 1999), muscle tension, facial expressions (Pantic and Rothkrantz, 2000) and many others. But these indicators will fail to recognize true affect if someone can disguise his or her emotions (Picard et al., 2001). Hence, estimation of affective states is a challenging task and requires other sources of physiological signals which are hard to alter.

The obvious choice is neural signals, because the brain is the center of processing all these emotions and feelings. Specifically, the thalamus, hypothalamus, hippocampus, cingulate cortex, and the amygdala are considered to be responsible for emotion processing (Dalgleish, 2004). Hence, it should be possible to detect affect from neurophysiological signals. Systems that do this are named affective brain-computer interfaces (aBCIs) (Mühl et al., 2014). These aBCIs can be based on invasive or non-invasive technologies. However, to estimate affective state, one might not choose to implant electrodes on his or her brain. Thankfully, with the advent of few completely non-invasive brain imaging techniques, it became possible to record brain signals without any surgical procedure. The available methods are magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG) and magnetoencephalography (MEG). Among them, fMRI and PET have a high

spatial resolution but low temporal resolution because of their dependence on metabolic changes. On the other hand, EEG and MEG have a high temporal resolution but low spatial resolution.

EEG-based systems are preferable for multiple reasons. The first is the unobtrusive nature of the EEG based systems. Second, the ability to be recorded using wearable devices which makes the system mobile. Finally, EEG acquisition systems are cheaper and less complicated than other techniques (Mühl et al., 2015). Because of these advantages, EEG based systems are a practical choice for aBCIs. The next task, before using aBCIs for detecting affect is to express different affective states on a measurable scale. For this purpose, Russell (1980) proposed an emotional state model known as the *circumplex model* based on the dimensions of valence and arousal dimension.

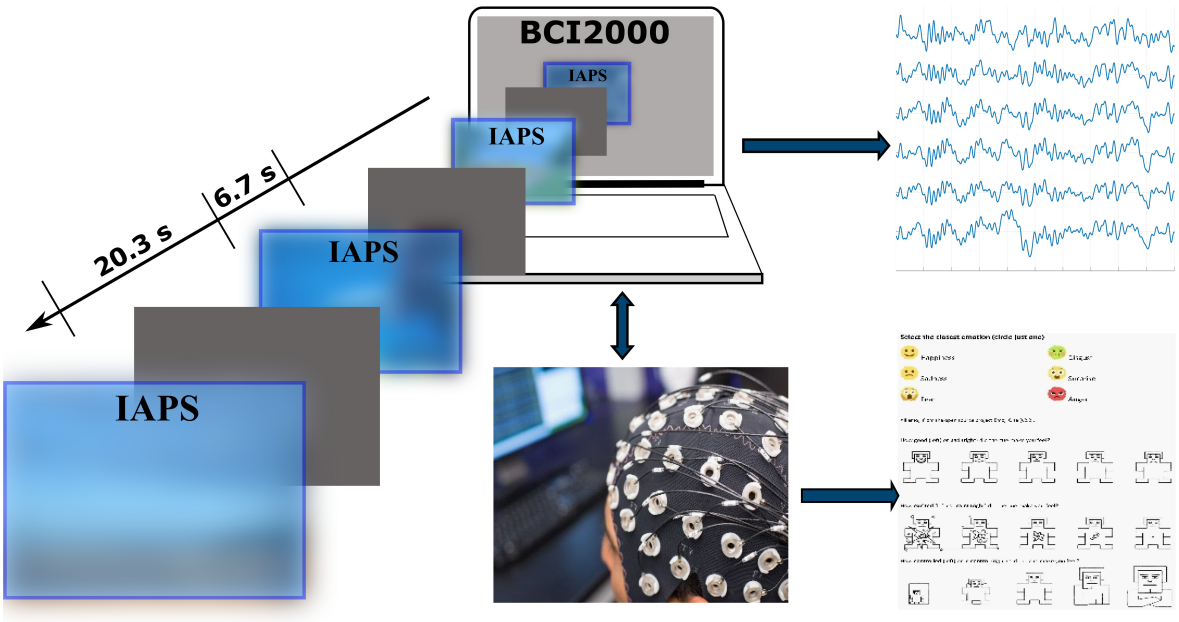


Figure 1.8: *Visualization of the experimental setup for affective computing.*

In the affective computing experiment, a set of visual stimuli was presented to the participants and simultaneously EEG data was recorded. This experiment used the international affective picture system (IAPS) (Lang et al., 2008) images as the stimuli. For the interfacing purpose, the BCI2000 (Schalk et al., 2004) was used to present picture stimuli to the sub-

jects. Each picture was displayed for 6.7 seconds and a blank display followed for 20 more seconds for participants' self report on a printed self-assessment manikins (SAM) (Bradley and Lang, 1994) for each stimulus to rate them in valence, arousal, and dominance on a discrete 5-point scale. Fig. 1.8 shows a simplified flow diagram of the experimental setup.

1.3 Outline

The first goal of this dissertation was to investigate the relationship of the P3 latency variations with BCI performance and utilize that relationship to improve BCI performance. A method using classifier-Based Latency Estimation (CBLE) and wavelet transform to enhance the P3 speller performance is introduced in Chapter 2 to accomplish the first goal. Further, a state-of-art classification technique using sparse autoencoders (SAE) is used and compared with least squares (LS) and step-wise linear discriminant analysis (SWLDA) methods for CBLE. The comparison is presented in Chapter 3. The second goal of this dissertation was to explore the potentials of the BCIs system to recognize emotion. Chapter 4 explores the usability of BCIs in the area of affective computing and emotion recognition. Finally, Chapter 5 includes the concluding remarks, future directions based on the presented work, limitations of the proposed methodology in various chapters, and contributions.

Chapter 2

Enhancing P300-BCI Performance Using Latency Estimation

In this chapter, the Classifier-Based Latency Estimation (CBLE) and wavelet transform was used to enhance the P3 speller performance. The CBLE method uses a classifier to estimate the latency variance. Hence a second-level classifier was used to classify the target characters. Least squares (LS), step-wise linear discriminant analysis (SWLDA), and support vector machine (SVM) classifiers were used in this chapter as the second-level classifier.

2.1 Introduction

Brain-Computer Interfaces (BCIs) use brain signals to provide a direct method of interaction with computers and other devices (Wolpaw et al., 2002). BCIs can help to restore communication for people with severe movement impairments such as amyotrophic lateral sclerosis (ALS), neuromuscular disease (NMD), brainstem stroke, cerebral palsy, and spinal

cord injury (McFarland and Wolpaw, 2011). One of the most common BCI applications is the P300 or P3 Speller indroduced by Farwell and Donchin (Farwell and Donchin, 1988), which uses event-related potentials (ERPs), including the P300 – a positive deflection approximately 300ms post-stimulus. In classical P3 Speller implementations, a grid matrix of 6×6 or more characters and commands are presented to the user. Subsets of the matrix, usually rows and columns, are flashed in a random order (c.f. (Townsend et al., 2010)). The probability of the flashed row/column containing the target character is 1/6, which creates a rare event that will elicit a P300 response. A classifier can detect those elicited P300 responses and identify target characters (Fabiani et al., 1987). With a few exceptions (e.g. (Kindermans et al., 2014a,b)), classifiers must be trained on data from each participant, as all event-related potentials (ERPs) including the P300 are participant-specific.

Researchers have tried different feature extraction and classification methods for the P3 Speller in search of better performance (see e.g. (Krusienski et al., 2006)). Early research used step-wise linear discriminant analysis (SWLDA) and showed that SWLDA performed well as a P3 Speller classifier (Sellers and Donchin, 2006; Donchin et al., 2000; Krusienski et al., 2008). For the 2003 BCI competition data, support vector machines (SVM) (Kaper et al., 2004; Rakotomamonjy and Guigue, 2008) outperformed other classifiers, though the performance was dependent on proper tuning parameters. Other recent works used Bayesian Linear Discriminant Analysis (BLDA) and Fisher's Linear Discriminant Analysis (FLDA) (Hoffmann et al., 2008) and Convolutional Neural Network (CNN) (Cecotti and Graser, 2011) for classification. All these above-mentioned works reported some improvement on performance compared with prior studies.

Early research on the P300 found that P300 latency and reaction time varies between people (McCarthy and Donchin, 1981; Kutas et al., 1977). Magliero found that the latency of the P300 depends on the stimulus evaluation process (Magliero et al., 1984). These and other studies have shown that P300 latency varies, and that this variation is related to age, cognitive disabilities and other factors (Picton, 1992; Polich, 2007). Latency varies within-user, within the same session (Fjell et al., 2009; Thompson et al., 2012) and even trial to trial (Blankertz et al., 2011). Hence, P300 latency can affect classifier performance

and BCI speed (Polich and Herbst, 2000). Though P300 latency is an important factor for the P3 Speller, only a few very recent studies have attempted to explicitly calculate or correct for P300 latency. Researchers used Bayesian methods (D'Avanzo et al., 2011) and spatiotemporal filtering methods (Li et al., 2009) to estimate properties of single-trial event-related potentials (ERPs), including latency estimates. But, surprisingly only one study has been found in the literature which attempted to correct latency jitter (Walhovd et al., 2008), using a maximum-likelihood estimation (MLE) method. In Thompson et al. (2012), we proposed a classifier-based latency estimation (CBLE) method to estimate the P300 latency. In that work, the latency estimates were primarily used to predict BCI performance from small datasets.

In this study, we used a wavelet transform of the CBLE scores as input features to another classifier, improving overall BCI performance. As for the CBLE method it relies upon, this new technique should be helpful regardless of the classifier used. The new technique should dynamically account for latency variation on a per-flash basis, unlike previous work such as Iturrate et al. (2014), which showed improved BCI performance from a static correction for the average latency in different tasks.

2.2 Experimental Data and Methods

2.2.1 Data Description

An earlier study by Thompson et al. (2012) demonstrated a classifier-based latency estimation technique to estimate and predict BCI accuracy from small datasets, which will be discussed later in section 2.2.3. Some of the data used here were previously reported in (Thompson et al., 2012, 2009), and all other data were taken using the same protocol. This protocol involved three separate visits (sessions) for each participant. There are three data files per session, with an additional training file in the first session. This study includes data from all files from sessions one, two and three. Results are shown separately for the average for files from session one and the average for files from sessions two and three combined. The

participants included 9 people with ALS, 4 people with NMD, and 20 control participants with no motor impairments. Only people who completed the study are included.

EEG data were collected using a 16-electrode cap from ElectroCap International, with mastoid reference and ground. The electrodes were fixed in the cap at F3, Fz, F4, T7, T8, C3, Cz, C4, Cp3, Cp4, P3, Pz, P4, PO7, PO8, and Oz according to the 10-20 electrode placement system. The data were amplified and digitized at 256 Hz using a g.USBamp (Guger Technologies). Stimulus presentation and recording was controlled through the BCI2000 software platform.

Online classification was performed using least squares (LS). The training file was used to create a participant-specific classifier that was used in all three sessions. A heuristic based on training accuracy was used to set the number of times each row and column flashed (sequences). Each data file contains at least 23 characters of BCI typing; users corrected mistakes using a backspace selection within the BCI, so the number of characters varies between files. For additional details, see Thompson et al. (2012).

2.2.2 Classifier Basics and Terminology

Perhaps because classifiers and machine learning techniques have broad application domains, their terminology is not yet perfectly standardized. In this work, we will be discussing three classifiers - Least Squares (LS), Step-Wise Linear Discriminant Analysis (SWLDA), and Support Vector Machine (SVM). The earlier CBLE work (Thompson et al., 2012) demonstrated better performance on the dataset used here using LS classification method. A comparison study (Krusienski et al., 2006) showed that SWLDA provides the best overall performance characteristics for practical P300 Speller classification. Another related study showed that the linear SVM classifier performed better in identifying P300 ERPs compared to Fisher linear discriminant analysis (FLDA) (Combaz et al., 2012). Hence this study will examine these three (LS, SWLDA, Linear-SVMs) classification methods as the second-level classifier. Each method is a linear classifier, meaning that it works by taking a weighted sum of the inputs (features). This weighted sum will be called a "score." This process is done once per

"observation" or measurement, in our case once per "flash." In typical binary classification tasks, the sign function is applied to the score for each observation, in order to estimate the class "label" - whether the observation in question belongs to the positive or negative class.

The classifiers we use differ primarily in how the weights (which then are used to calculate the score) are chosen. The score, $\hat{y}$, is calculated using the following equations (Murphy, 2012) :

$$\texttt{LS:}\ \hat{y}(\mathbf{x}) = \mathbf{X}\hat{\mathbf{W}}_{LS} \text{ where, } \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix} \text{ and } \hat{\mathbf{W}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.1}$$

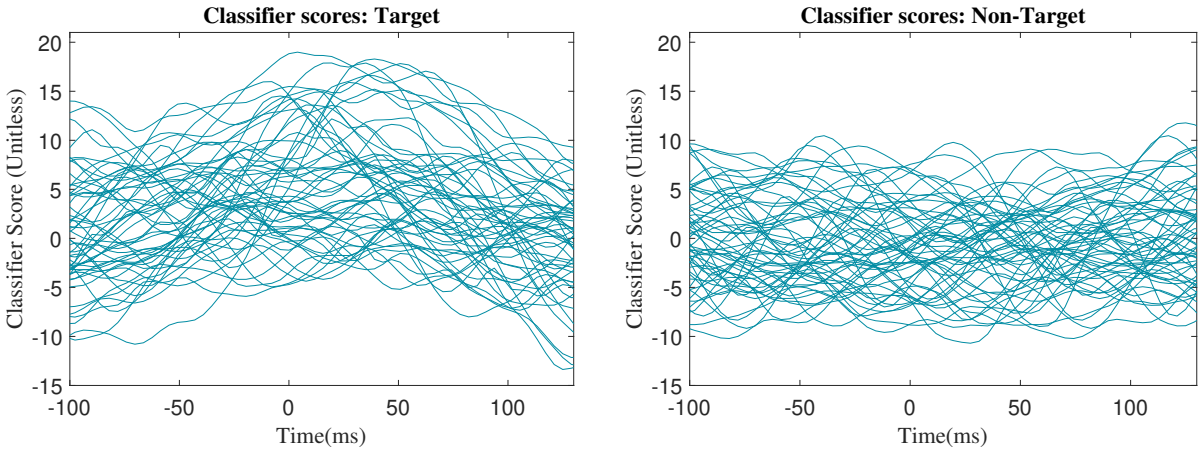
$$\begin{aligned} \texttt{LDA:}\ \hat{y}(\mathbf{x}) &= \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) \text{ where, } \mathbf{w} = \textstyle\sum^{-1}(\mu_1 - \mu_0) \text{ and} \\ \mathbf{x}_0 &= \frac{1}{2}(\mu_1 + \mu_0) - (\mu_1 - \mu_0)\frac{\log(\pi_1/\pi_0)}{(\mu_1 - \mu_0)^T \sum^{-1}(\mu_1 - \mu_0)} \\ &\pi \text{ is the prior probability of membership in each class} \end{aligned} \tag{2.2}$$

$$\texttt{SVM:}\ \hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \text{ Where, } \alpha_i = \lambda_i y_i,\ \lambda \text{ is the } \ell_1 \text{ regularization term.} \tag{2.3}$$

P3 Spellers are unusual among binary classification tasks, because each row and column is flashed multiple times while the user is trying to produce a single output character. The sign function is therefore not used, and instead the scores ($\hat{y}$) for the multiple observations of each row and column are averaged. Then the maximum-scoring row and column are chosen. Note that as these three classifiers are all linear (we used a linear kernel for the SVM), this process is equivalent to averaging the features from multiple observations prior to classification.

2.2.3 CBLE

Traditional P300 classification uses a single time window locked to the stimulus presentation, for example, the EEG signal 0 to 800 ms post-stimulus (Krusienski et al., 2006). Classifier



(a) Target characters.

(b) Non-target characters.

Figure 2.1: *Sample classifier scores as function of time shift of participant K143 (participant with ALS).*

based latency estimation (CBLE) creates many copies of that time window, each offset by an integral number of amplifier samples; for example, if the sampling rate was 1kHz, one window might be -1 to 799 ms, and another 1 to 801 ms. The "first-level classifier" (here, LS) is applied to the data in each window, producing a score as described above. Thus, the method produces a vector of scores, with one element per time shift used.

In Thompson et al. (2012), the time shift that produced the maximum score was used as an estimate for the latency difference between the new P300 response and the average P300 response from training data. The variance of that latency difference estimate on target characters was used to predict BCI performance. The vector of scores was not used directly, although we did note that there are strong differences in the shape of the scores for target and non-target characters. In this work, by contrast, we wanted to use the full vector of scores directly, to aid in detection of the P300 response.

Fig. 2.1 shows the CBLE scores as a function of time shift from a participant with ALS, for several representative flashes. Fig. 2.2(a) shows the average across all flashes, which reflects the overall shape of the responses to target flashes. The CBLE scores from most target flashes show a peak near 0 time shift; different flashes produce different peak times. This is an indication of latency jitter and also shows how latency jitter can affect the P300

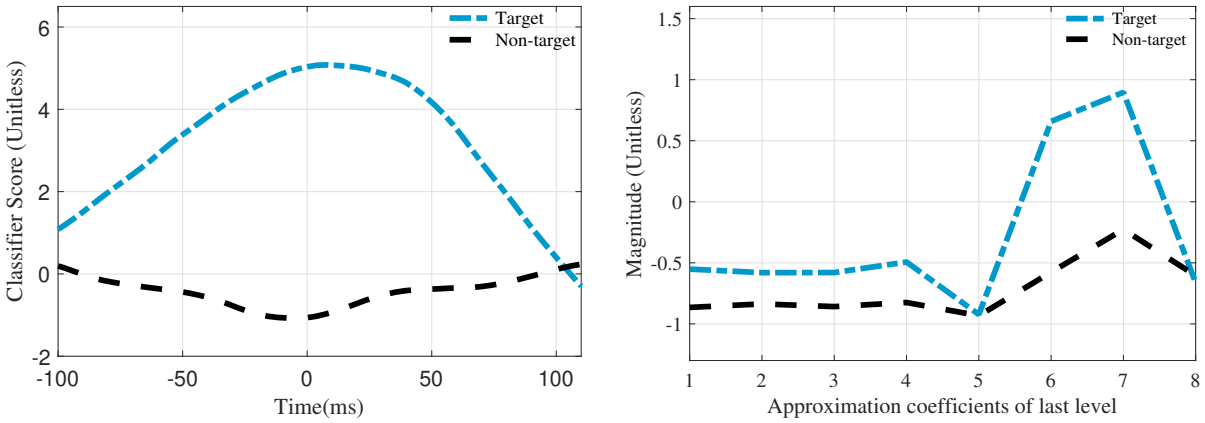
classifier performance. For non-target characters there are no visible peaks which is also expected behavior in this paradigm. A few naive approaches were to (i) align all the single-trials based on CBLE and use the aligned trials to train a second-level classifier, (ii) use the CBLE outputs along with the non-aligned trials as extra features for a second-level classifier and (iii) use the CBLE scores alone to train a second-level classifier. We tested each of these approaches on pilot data and found the third approach more useful than the others. However, latency jitter is still visible in the fact that the CBLE scores peak at non-zero time shift; we wanted to reduce the number of features for the end classifier and also reduce the latency jitter. Given the characteristic shape of the CBLE scores for target flashes, we thought a frequency domain transform such as wavelets would be valuable.

2.2.4 Wavelet Transforms

Wavelet transforms are generally used for decomposing signals into multiple time-frequency domains. However, they also can be used for feature reduction. We accomplished both purposes by computing the wavelet approximation coefficients of CBLE scores. For a signal with $N-1$ samples, $x(t) = \{x(1), \ldots, x(N-1)\}$, the approximation coefficients can be calculated from equation 2.4 (Chun-Lin, 2010):

$$W_{\phi}[j_0, k] = \frac{1}{\sqrt{N}} \sum_{n} x(n) \Phi_{j_0,k}(n) \tag{2.4}$$

There are many wavelet families; we applied different mother wavelet transformations on session 1 data for several participants, and that found the Daubechies-4 mother wavelet, particularly the 5 level wavelet decomposition, produced good results while significantly reducing the number of features. To find the approximation coefficients of last level we used MATLAB default `appcoef.m` fuction. Fig. 2.2(b) shows the averages of approximation coefficients of CBLE score vectors, for target and non-target flashes.



(a) Averages of classifier scores.

(b) Averages of approximation coefficients.

Figure 2.2: *Averages of classifier scores which are shown in Fig. 2.1(a) & Fig. 2.1(b) and wavelet approximation coefficient for target and non-target characters.*

2.2.5 Second-level Classifier

Wavelets reduced the dimensionality of the CBLE scores while still showing a difference between target and non-target characters, but a classifier is still needed to make decisions based on the wavelet coefficients. We investigated three classifiers (LS, SWLDA, and SVM) as "second-level" classifiers, which were given only the wavelet coefficients as input features. The scores from these second-level classifiers were used in the typical P3 Speller fashion - each flash was scored by the second-level classifier, and the scores for each row and column were averaged individually. The row and column with the highest average score was designated as the selected output character. Both first- and second-level classifiers were trained using only the training data file.

2.2.6 Performance Measurement

Because the goal of this work is improving communication accuracy and speed, a performance metric capturing *throughput* was chosen. Although Information Transfer Rate is often used for the P3 Speller, we have chosen BCI utility (Dal Seno et al., 2010), in line with the suggestions in (Thompson et al., 2014, 2013). BCI utility (U) is calculated using the formula of equation 2.5:

Table 2.1: *Performance in different sessions for participants with ALS. Bolded participants show consistent improvement.*

participants		**K143**	**K145**	K146	K147	**K152**	**K155**	K156	K158	**K160**
Average Accuracy	Online	91.61	58.89	96.3	95.15	90.26	59.9	93.05	88	70.57
in Session 1 (%)	Wavelet	92.85	63.33	95.37	93.86	91.34	65.2	93.05	87.97	72.5
Utility change(%)		2.97	30.77	-2.00	-2.84	2.67	62.69	0.00	-0.08	9.4
Average Accuracy	Online	92.15	77.94	89.7	88.65	50	59.09	86.98	61.91	30.08
in Session 2 and 3 (%)	Wavelet	93.05	80.1	88.35	90.43	53.4	62.99	88.16	62.86	36.37
Utility change(%)		2.14	7.76	-3.40	4.62	157.1	18.19	3.17	6.97	NE[1]

[1] NE: Utility does not exists.

$$U = \frac{2p-1}{c}, \text{ this is only valid for } (2p-1) > 0, \text{ i.e, } p > 0.5 \tag{2.5}$$

where c is the time per selection, and p is the probability of correctly selecting a symbol or character in the interface. We calculated this probability by assuming it was constant for all characters and within the duration of each file. Backspaces, if required to produce correct text, were counted as correct selections for calculating accuracy. These accuracies were then averaged together if multiple files were used (as an example, if we report average session 1 accuracy).

BCI utility (U) is a useful metric in that it correctly calculates the rate of corrected characters per unit time. In other words, BCI utility (U) is a measure of "corrected typing speed," or how quickly a person can produce corrected text.

2.3 Results

Table 2.1 shows the online accuracies and accuracies after the proposed method for a subset of participants to demonstrate how BCI Utility changes with the change of accuracies. For readability, we limited the table to only participants with ALS as they come from a potential end-user population. Bolded participant identifiers indicate consistent improvement across sessions, which was found for the five of the six participants with online accuracies at or

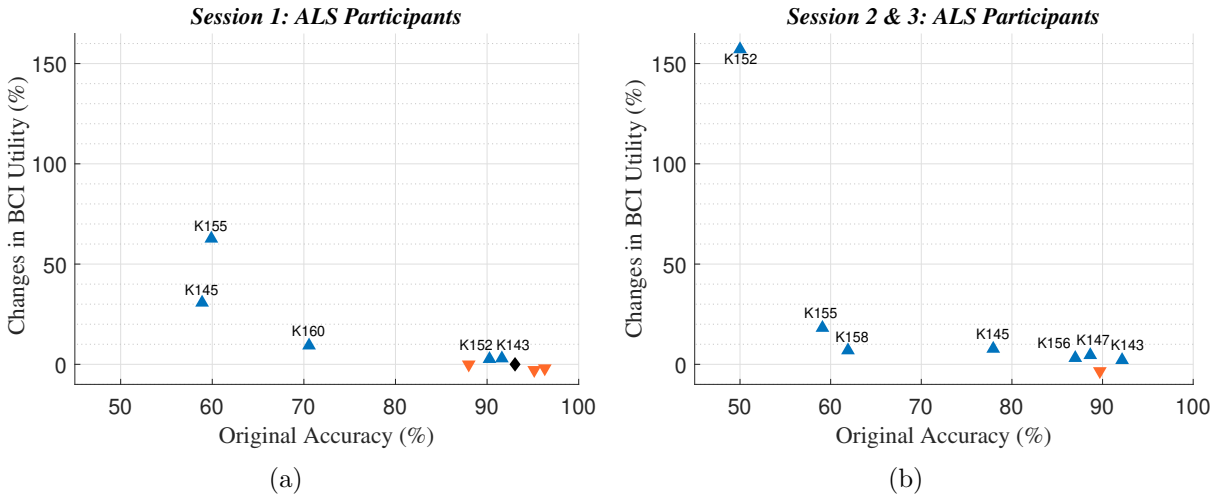


Figure 2.3: *Changes in BCI Utility for participants with ALS versus online test accuracy in different sessions. Upper-triangle and lower-triangle indicates the BCI Utility increased and decreased, respectively. The diamond indicates no change. Participant IDs are shown only for improved performance, to allow the reader to assess consistency of improvement.*

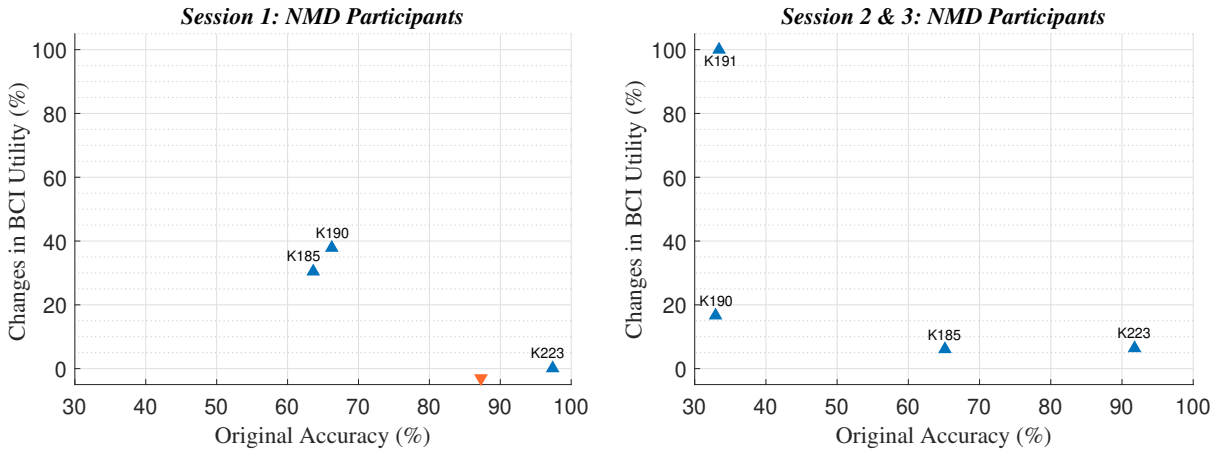
below 90%.

Fig. 2.3 shows the improvement of BCI Utility for participants with amyotrophic lateral sclerosis (ALS) after using the proposed technique. The technique shows greater improvement for participants with lower online accuracy. Improvement of performance is also consistent in other sessions as shown in Fig. 2.3(b). Note participant K160 is not plotted in Fig. 2.3(b) despite the improvement shown in Table 2.1 because this participant's online utility was zero and the percentage change become mathematically undefined. Mean change in BCI Utility in session 1 is 11.5% and in session 2 and 3 is 24.57%.

Fig. 2.4 shows the improvement for participants with neuromuscular disease (NMD). Again, larger benefit is shown for individuals with lower online accuracy. In session 2 and 3, the performance improved for all four participants, but are larger in value for lower online accuracies. Mean BCI Utility change for session 1 is 16.35% and for session 2 and 3 is 32.29%.

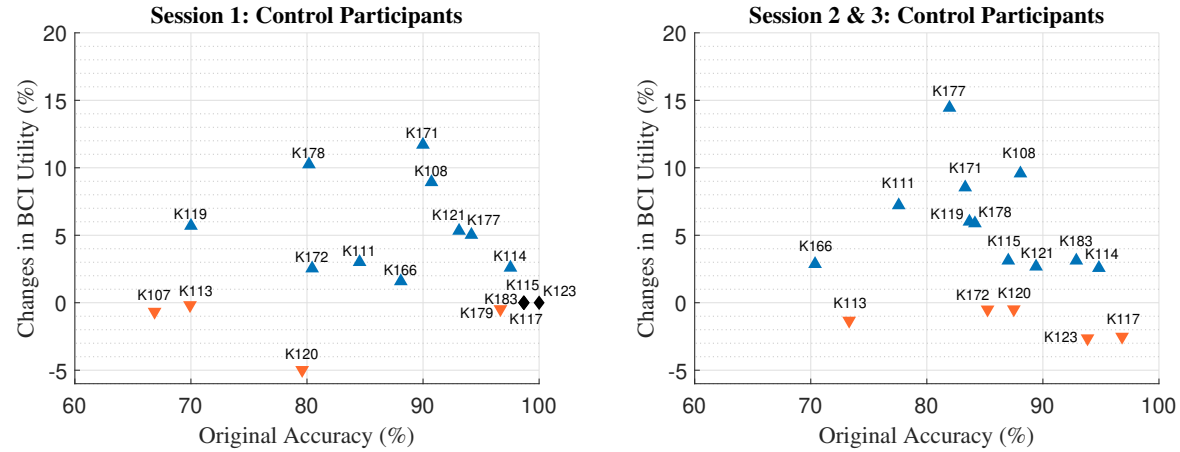
Fig. 2.5 shows the effect of the proposed technique on 21 control participants. Unlike participants with ALS and NMD, there is no obvious pattern. Mean BCI Utility change in session 1 is 2.03% and in session 2 and 3 is 16.59%.

Overall we had data from 33 participants. In session 1, 18 participants showed improved



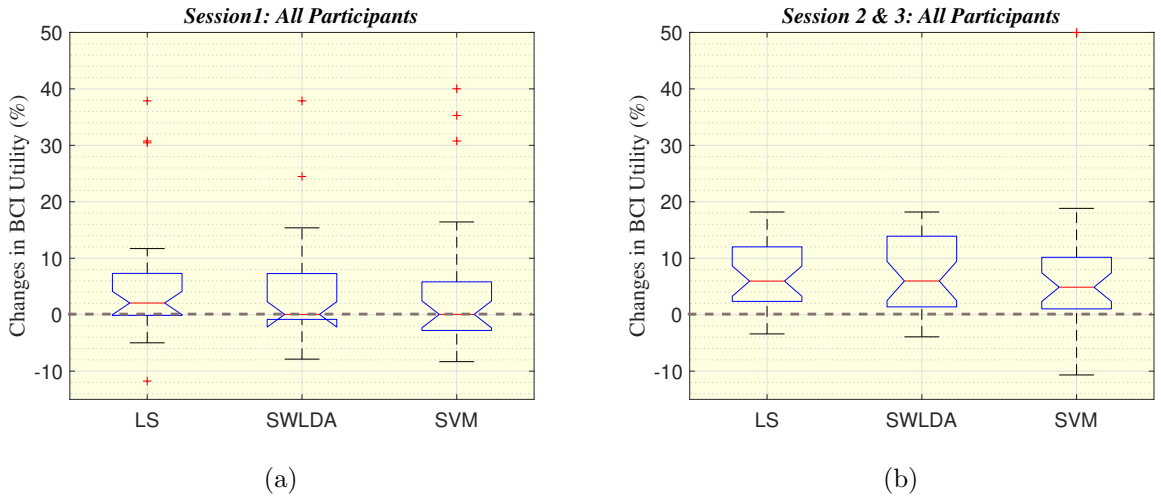
(a) (b)

Figure 2.4: *Changes in BCI Utility for participants with NMD versus online test accuracy in different sessions. Upper-triangle and lower-triangle indicates the BCI Utility increased and decreased, respectively. Participant IDs are shown only for improved performance, to allow the reader to assess consistency of improvement.*



(a) (b)

Figure 2.5: *Changes in BCI Utility for control participants versus online test accuracy in different sessions. Upper-triangle and lower-triangle indicates the BCI Utility increased and decreased, respectively. Diamonds indicate unchanged performance. Participant IDs are shown only for improved performance, to allow the reader to assess consistency of improvement.*



(a) (b)

Figure 2.6: *Box plots of the BCI Utility changes for LS, SWLDA and SVM on different sessions.*

performance with the new method, with a mean of 13% increase in BCI Utility. Nine participants showed decreased performance, with a mean decrease of -2.89% in BCI Utility. Five participants had no change in performance. Among these five, one participant had very low online accuracy and the other four had online accuracies around 98%. Overall mean change in BCI Utility for session 1 for all participants is 6.5%.

In session 2 and 3, two participants (K118, K160) had original accuracies of 32% and 30%. While their accuracy improved by 2 and 6 percentage points, neither showed non-zero BCI Utility with or without the new method. Twenty-three additional participants showed increased performance, with mean BCI Utility changes of 27.5%. Seven participants' performance worsened with a mean of -1.6% utility change. Overall mean of utility changes in session 2 and 3 is 20.75%.

We have also compared the performance of LS, SWLDA and SVM binary classifiers as the second-level classifier. The results are shown in Fig. 2.6. In session 1, both LS and SWLDA classifiers have median change above 0, and first quartile at or very near 0, indicating that approximately 75% of the subjects experienced improvement or at least no decrease. For SVM on session 1, the median is nearer to zero but the quartile is below zero. On session 2 and 3, the box plots are more similar between classifiers.

Finally, Fig. 2.7 shows the changes in BCI Utility versus the standard deviation of CBLE-

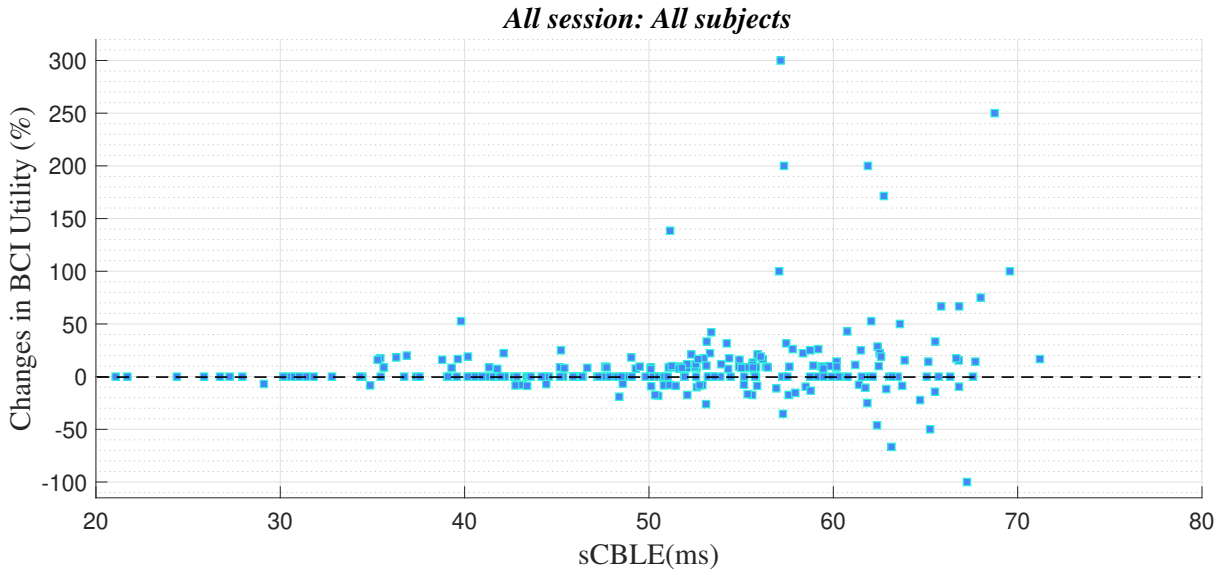


Figure 2.7: *Changes in BCI Utility versus sCBLE (the standard deviation of CBLE-estimated P300 latency from target flashes).*

estimated latency for all 288 files. The Fig. demonstrates that the changes in performance are located at areas with higher estimated latency jitter.

Here, we have reported data for 32 participants overall. Modeling performance change as a binomial random variable with "success" being increased performance, the maximum likelihood estimate (MLE) of increased performance probability is 0.67 with 95% confidence interval of [0.54 0.78]. This can be interpreted as the technique being more likely to help than do nothing or decrease performance. Using the two-sided p-value gives us $p_2 = \sum_{s=43}^{64} Bin(s|64, 0.5) + \sum_{s=0}^{21} Bin(s|64, 0.5) = 0.0081 < 0.01$.

If we define success more generously, as improving or at least not changing performance, the MLE is 0.75 with 95% confidence interval of [0.63 0.85]. Two-sided p-value in this case is given by $p_2 = \sum_{s=48}^{64} Bin(s|64, 0.5) + \sum_{s=0}^{16} Bin(s|64, 0.5) = 0.000077 < 0.001$.

2.4 Discussion

From previous studies, it is obvious that P300 latency varies between individuals, between sessions for the same individual, and most importantly between trials even for the same

individual and session (McCarthy and Donchin, 1981; Kutas et al., 1977; Fjell et al., 2009; Thompson et al., 2012; Blankertz et al., 2011). The effect of latency variations between individuals can be compensated by using subject-specific classification, and training on the same day can address some of the between-session variations. But trial-to-trial variations in latency make the classification task difficult and also affect BCI performance. This motivated us to find a technique to correct latency variation and thus minimize the effect of latency jitter on performance. Previously, the classifier-based latency estimation (CBLE) method has been used to predict BCI performance (Thompson et al., 2012). Here, CBLE-estimated latency has been used to improve BCI performance.

At the beginning of the investigation, we used CBLE-estimated latency to "correct" for latency jitter on a trial-by-trial basis, and used the corrected trials as the "second-level" classifier's features. However, the improvement in performance was not significant enough to merit reporting - without knowledge of the class labels, correcting for latency had the unfortunate effect of maximizing the classifier score for examples that did not contain P300's, leading in many cases to less separable score distributions. Further investigations using feature reduction techniques, such as wavelet transforms, provided better results. We found that wavelet approximation coefficients of CBLE scores are also different for target and non-target characters (Fig. 2.2(b)). That findings motivated us to use wavelet approximation coefficients as features for our "second-level" classifier. For that "second-level" classifier, we have compared LS, SWLDA and SVM. Though the performance for all three classifiers was almost equal, it is notable that a comparitively simple classification technique, LS, was found to be equally or more effective than SWLDA and SVM. This provides an insight that using better feature transformation methods may give better results even while using simple classification techniques.

The proposed technique appears to be helpful only for participants with lower accuracies and higher estimated latency jitter, which are strongly correlated (Thompson et al., 2012). The largest improvements were found among participants with ALS and NMD. Part of this is because the BCI Utility metric highlights the importance of relatively small absolute changes in accuracy when the accuracy is low. For example, one participant with ALS had an increase

of accuracy of 4.4%, resulting in a 30.76% improvement in BCI Utility. However, BCI Utility is measuring the capability of a person to produce corrected text - in other words, it is an ecologically valid measure of communication throughput. While the accuracy changes are small, for users who struggle with the BCI (accuracies in the 50-70% range), even small changes can show large improvements in usability.

The two-sided p-value we have found demonstrates that the proposed technique statistically improves BCI performance. It should be noted that for some subjects, the performance was already good enough that there was no need or room for improvement in accuracy. In this case, it is desirable that our performance improvement technique would not decrease performance for these individuals. Hence, we have also computed the two-sided p-value for the performance improving or at least not changing the performance. That p-value was also statistically significant and demonstrates that this method is more likely to help or do nothing than to hurt performance.

While the improvements here are not large in magnitude (and not the orders of magnitude of improvement that are necessary to restore natural speech, for example), it is notable that the improvements are much larger in our pool of participants from potential user populations. The method does not completely compensate for the effects of latency jitter found in (Thompson et al., 2012), and significant improvement is still required beyond this work to bring all users to equal performance.

We believe the power of this method lies in its ability to correct for latency variation. Our previous work has shown that latency variation as measured by CBLE is strongly inversely correlated to BCI performance (Thompson et al., 2012). This has a compounding effect for individuals with high online accuracies. If the online accuracy was near 100%, not only is there little room for improvement in an absolute sense, but the participant almost certainly demonstrated little latency variation. Since this method provides improvement by removing latency variation, these individuals will see little benefit from this method. However, it should be noted that target populations for BCI often experience lower performance than controls, so this is not a critical weakness of this method.

Finally, it is noted that on this dataset LS appears to perform better than SWLDA, even

without CBLE. This is in contrast with (Krusienski et al., 2006), and may be due to the fact that the number of stimulus presentations for each participant was chosen based on LS performance.

2.5 Limitations

CBLE itself has been demonstrated to be at least partially classifier independent (Thompson et al., 2012). Therefore, it is possible that this boosting method could be applied with other first-level classifiers being used to estimate the latency. However, we have not tested this claim here. We did use a SWLDA-based CBLE with this approach, but the results were not different enough to merit inclusion.

This is an offline analysis of existing data, and the method is not yet ready for online implementation.

2.6 Conclusion

This work demonstrates an improvement in information throughput using a technique that can be used with many classifiers, including the relatively simple LS classifier used here. Interestingly, the improvement is the largest for participants with marginal accuracies, those for whom the typical techniques produced some communication but not ideal performance. This suggests that the technique helps to offset, but does not eliminate, the negative effect of latency jitter on classification. Further work on removing latency jitter should continue providing improved performance for individuals for whom current-generation BCIs do not perform well.

Acknowledgement

The authors would like to thank all individuals involved in early stages of this investigation, especially Carmela Lee and Robert Trotter who performed the bulk of the data collection.

Funding

This work was supported in part by the National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH) under grant R21HD054697 and by the National Institute on Disability and Rehabilitation Research (NIDRR) in the Department of Education under grant H133G090005. The opinions and conclusions are those of the authors, not the respective funding agencies.

Chapter 3

Comparison of Classification Techniques to Predict BCI Accuracy Using CBLE

Copyright notice: The following text is reformatted from "A Comparison of Classification Techniques to Predict BCI Accuracy Using Classifier-Based Latency Estimation." as submitted for publication (Mowla et al., 2020a).

In this chapter, an extended CBLE method using sparse autoencoders (SAE) is proposed and compared with LS- and SWLDA-based CBLE. The objective of this study is to compare different classification techniques to predict BCI accuracy using the variance of CBLE estimates. Here, the newly-developed SAE-based CBLE and previously used methods are applied to a newly-collected dataset. Results showed a significant ($p < 0.001$) negative correlation between BCI accuracy and estimated latency jitter. This study showed that whole CBLE worked regardless of the method and electrode count; the effect of the number of electrodes on BCI performance was classifier-dependent.

3.1 Introduction

Brain-computer interfaces (BCIs) are an alternative communication technology for people with severe neuromuscular disorders such as amyotrophic lateral sclerosis, cerebral palsy, stroke, or spinal cord injury. BCIs are defined as systems that record brain signals, interpret and translate those signals into an output device to perform user-desired actions (Shih et al., 2012). One type of BCI is the P300 speller, first introduced by Farwell and Donchin (Farwell and Donchin, 1988), which gained significant attention from BCIs researchers due to its short training period and good performance (Bianchi et al., 2019). As the name suggests, the P300 speller uses the P300 event-related potential (ERP), which is elicited by rare and task-relevant stimuli (Donchin et al., 2000). In the standard P300 speller system, the user observes different characters and commands in a matrix format and the columns and rows are flashed in a random order. The user will count the number of times the target character is flashed. An oddball paradigm is created due to the low probability of a flashed row/column containing the target, which therefore elicits P300 ERPs.

However, the P300 is not a perfectly stereotypical waveform. Its amplitude and latency vary widely for different users (Guger et al., 2009), and even for the same user in different sessions (Fjell et al., 2009). These variations are influenced by many factors, such as age, gender (Polich and Kok, 1995), fatigue, exercise (Yagi et al., 1999) and attention (Polich, 2007). One major effect of P300 latency variation is decreased system performance (Thompson et al., 2012; Aricò et al., 2014).

Because of such variations in P300 amplitude and latency, several studies have proposed methods to estimate characteristics of the P300 potential including latency (e.g., D'Avanzo et al. 2011; Li et al. 2009). But, fewer studies have examined the effect of this jitter on P300 Speller performance; to our knowledge, the first was our paper on classifier-based latency estimation (CBLE) (Thompson et al., 2012). A later study by another group also independently confirmed a negative link between latency jitter and BCI performance (Aricò et al., 2014). Later, the CBLE estimates and a wavelet transform were used to provide the latency jitter information to a second-level classifier (Mowla et al., 2017). The combination

resulted in an enhanced BCI performance. The potential of the CBLE method to predict BCI performance made us interested in investigating a non-linear classifier based CBLE method and the prediction accuracy of CBLE on a different dataset.

CBLE uses the classifier's sensitivity to latency variability to estimate P300 latency. In our previous work, it was claimed that i) CBLE is classifier independent and ii) CBLE can be used to predict BCI accuracy. A comparison of least-squares (LS) and stepwise linear discriminant analysis (SWLDA) was used to support the first statement. However, both LS and SWLDA are linear classifiers, and SWLDA has the same solution subspace with LS for binary classification problems (Ye, 2007; Lee and Kim, 2015). Hence classifier independence was indicated, but not verified, particularly for non-linear classifiers.

In this work, we will extend our previous CBLE investigations using a sparse autoencoder (SAE), and will examine if classifier independence holds for this non-linear classifier. Both previously-used classification methods (LS, SWLDA) as well as the new non-linear method (SAE) will be used with a new P300 dataset to further verify the ability of CBLE to predict BCI accuracy. The motivation behind choosing these three classification methods are:

i) LS provided the best overall performance on the dataset used in CBLE's original article (Thompson et al., 2012),

ii) In a classifier comparison study (Krusienski et al., 2006) SWLDA provided the overall best performance, and

iii) A recent study (Vařeka and Mautner, 2017) showed that SAE provided the best overall performance on their dataset for P300 speller. But SAE has never been used to estimate latency jitter to our knowledge.

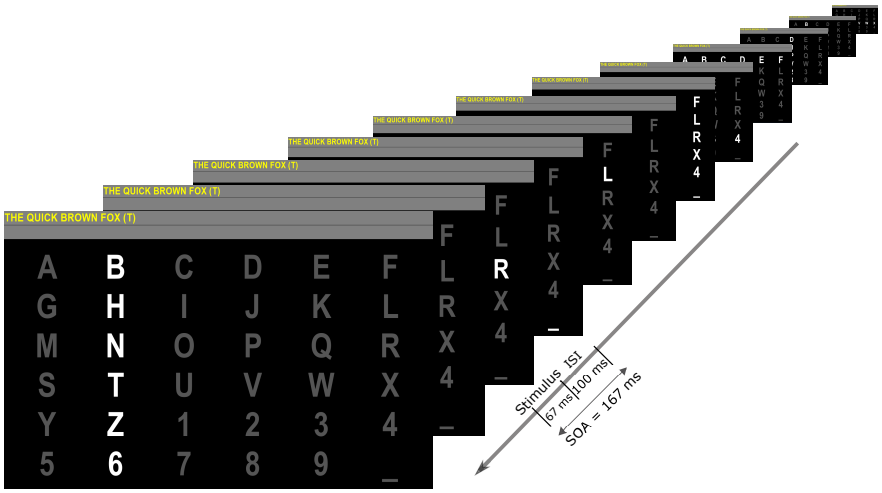


Figure 3.1: *A Visual interface of the* 6×6 *matrix used in this study. A row or column intensifies for 67 ms, followed by a 100 ms pause. The front-most image shows an intensification of the column containing the character "T". This is the current target, so a P300 is expected to be elicited by this intensification.*

3.2 Methods

3.2.1 Experimental Setup

Data were collected from each participant in three sessions, i.e., on three different days, using BCI2000's (Schalk et al., 2004) row-column P300 speller paradigm. Each session was comprised of copying three sentences. For each sentence, each row/column was either intensified or replaced with Einstein's face for 67 ms (stimulus duration) with an inter-stimulus interval of 100 ms. The stimulus onset asynchrony (SOA) was therefore 167 ms. A complete set of 12 intensification or replacements is called a sequence. For each character, we recorded data for 10 sequences. The copied sentences are shown in Table 3.1. The data from the first sentence in session 01 was used as training data to train the online classifiers and the data for remaining sentences were used as test data. The bolded sentences (one for each session) used Albert Einstein's iconic tongue face image instead of flashing.

EEG data were recorded using a Cognionics Mobile-72 EEG system with a sampling frequency of 600Hz. The Mobile-72 EEG system is a high-density mobile EEG system with active Ag/AgCl electrodes placed according to the modified 10-20 system. Reference and

Table 3.1: *Sentences copied by the participants.*

Session	Sentence to spell
01	THE QUICK BROWN FOX
	THANK YOU FOR YOUR HELP
	THE DOG BURIED THE BONE
02	MY BIKE HAS A FLAT TIRE
	I WILL MEET YOU AT NOON
	DO NOT WALK TOO QUICKLY
03	**YES. YOU ARE VERY SMART**
	HE IS STILL ON OUR TEAM
	IT IS QUITE WINDY TODAY

ground were on the right and left mastoids, respectively.

3.2.2 Participants

Nine healthy volunteers participated in this study. Data from two participants have been excluded due to their poor online and offline performance. Among the remaining participants, six were male and one female, with an average age of 20.86 ± 4.56 years. Two participants had previous brain-computer interface experience. Participants were provided informed consent and the recording process was performed in accordance with Kansas State University's Institution Review Board (IRB) protocol No. 8320.

3.2.3 EEG Pre-processing

Data were filtered using a finite impulse response (FIR) bandpass filter with corner frequencies at $(0.5 - 70.0)$ Hz, then split into epochs of 750 ms post-stimulus. The epochs were then downsampled by a factor of 30 using a moving average and downsample operation.

Two different sets of electrodes were used for classification. The first set was all 64 electrodes, while the other set was composed of 32 electrodes selected based on data from

each participant. To select the electrodes, the average P300 ERPs was produced by taking the difference of the average responses to target and non-target epochs on the training data. The power spectral density (PSD) of the resulting average ERP was used to select the 32 channels with the largest 3 Hz signal power (which should include the P300 response).

3.2.4 Classification Strategy

Detecting the presence of the P300 ERP is a binary classification problem, and most classifiers use the following general equation:

$$\hat{y}(\mathbf{x}) = \hat{\mathbf{w}}^T.f(\mathbf{x}) + b \tag{3.1}$$

where $\mathbf{x}$ is the feature vector, $\mathbf{w}$ is the weight vector and $f(.)$ is the transformation function. This transformation function $f(.)$ can be a nonlinear function, linear function, or simple identity function. For example, the sparse autoencoder classifier uses a logistic sigmoid function. $\hat{y}(\mathbf{x})$ is called the classifier's "score", and is used to decide the class of each "observation" or measurement. Since we expect the presence of P300 for one row and one column in each sequence, the target character is selected by

$$\hat{R} = \arg\max_{r} \sum_{r=1}^{6} \sum_{s=1}^{S} \hat{y}(\mathbf{x_{row}}) \tag{3.2}$$

$$\hat{C} = \arg\max_{c} \sum_{c=1}^{6} \sum_{s=1}^{S} \hat{y}(\mathbf{x_{col}}) \tag{3.3}$$

Here $\hat{R}$ and $\hat{C}$ are the predicted row and column, respectively. S is the number of sequences for each character. This classification strategy prevails in the P300 classification literature and is used in numerous studies (e.g., Krusienski et al. 2006; Rakotomamonjy and Guigue 2008).

Classifier-Based Latency Estimation (CBLE)

Standard P300 classification uses a single time window (e.g., 0 ms to 800 ms post-stimulus (Krusienski et al., 2006)) time-locked to each stimulus presentation. The Classifier-Based Latency Estimation (CBLE) method (Thompson et al., 2012) uses many time-shifted copies of the post-stimulus epochs, and finds the time shift that corresponds to the maximum score. The statistical variance of the CBLE is denoted vCBLE and is used as the predictor of the BCI's performance. In this study, BCI accuracy is predicted for each participant using the vCBLE estimates of that participant and the regression coefficients of the relationship between vCBLE and accuracy. The regression coefficients are obtained from the relationship between vCBLE and accuracy from all other participants (i.e., equivalent to leave-one-participant-out cross validation).

Least squares (LS)

LS is a linear classifier, meaning that it works by taking a weighted sum of the inputs (features).

$$\hat{y}(\mathbf{x}) = \hat{\mathbf{w}}_{LS}^{T}[\mathbf{x} \quad \mathbf{1}] \tag{3.4}$$

where $\hat{\mathbf{W}}_{LS}$ is estimated from the training data and corresponding class labels ($\mathbf{y}$) using the following equation:

$$\hat{\mathbf{W}}_{LS} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y} \tag{3.5}$$

Step-Wise Linear Discriminant Analysis (SWLDA)

Step-Wise Linear Discriminant Analysis (SWLDA) is an extension of Fisher's linear discriminant (Fisher, 1936) and was found very effective for P300 classification (Krusienski et al., 2008). SWLDA trains a linear discriminant analysis (LDA) classifier using a stepwise forward and backward regression method. Based on the F-test statistic, the step-wise method progressively adds the most correlated features in the discriminant model and removes the least correlated features during the forward and backward regression, respectively. LDA

finds the optimal features using the following equations:

$$\hat{y}(\mathbf{x}) = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) \tag{3.6}$$

where

$$\begin{aligned}\mathbf{w} =& \Sigma^{-1}(\mu_1 - \mu_0), \text{ and} \\ \mathbf{x}_0 =& \frac{1}{2}(\mu_1 + \mu_0) - (\mu_1 - \mu_0)\frac{\log(\pi_1/\pi_0)}{(\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0)}\end{aligned}$$

where Σ is the covariance matrix, π is the prior probability of membership in each class, and μ is the mean vector. In our case, we used $p< 0.05$ as a threshold to consider a feature statistically significant, and $p> 0.10$ to remove the least significant features. Also, the maximum number of features to be included was restricted to 60 features according to (Krusienski et al., 2008).

Sparse autoencoder

A single autoencoder (AE) is a fully-connected, two-layer neural network model which consists of one encoding layer and one decoding layer. The dimension of the encoding layer is the same as the dimension of the input features. The dimension of the decoding layer is, in general, less than the dimension of the encoding layer. The task of an AE is to encode the input features (x) to a hidden representation (z) with the aim to later reconstruct the input features (x) from z by minimizing the reconstruction error. For an input vector x, the encoder layer maps the vector x to another vector u such that

$$\bar{u} = f^{(1)}(W^{(1)}\bar{x} + \bar{b}^{(1)}) \tag{3.7}$$

Here, f is the transfer function of the encoder, W is the weight matrix, b is the bias vector and the superscript $*^{(1)}$ denotes layer 1. In our work, we will use a modified version of AE which is commonly known as sparse autoencoders (SAE). In SAE, sparsity is induced by adding

a regularizer term to the cost function to limit over-fitting. The sparsity regularization (Olshausen and Field, 1997) term, $\Omega_{sparsity}$ is defined by the using the Kullback-Leibler divergence of the average activation value, $\hat{\rho}_i$ of a neuron i and its desired value, ρ,

$$\begin{aligned}\Omega_{sparsity} &= \sum_{j=1}^{L} KL(\rho \| \hat{\rho}_i) \\ &= \sum_{j=1}^{L} \rho \log \frac{\rho}{\hat{\rho}_i} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_i}\end{aligned} \tag{3.8}$$

Here, L is the number of the neuron in the hidden layer. Kullback-Leibler divergence (Kullback and Leibler, 1951) is a measure of how similar or different two distributions are. Adding the sparsity regularization term requires ρ and $\hat{\rho}_i$ to be very similar to minimize the cost function. Another regularization, known as L_2 regularization, is also used to prevent $\Omega_{sparsity}$ from becoming small due only to higher values of weights. L_2 regularization, $\Omega_{weights}$ is defined as:

$$\Omega_{weights} = \sum_{i=1}^{L} \sum_{j=1}^{N} \sum_{k=1}^{D} w_{jk}^2 \tag{3.9}$$

Here, N is the number of observations and D is the dimension of the input (number of variables). Then the sparse autoencoder method uses the following cost function to estimate the parameters:

$$J(w,b) = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{D} (x_{dn} - \hat{x}_{dn})^2 + \lambda \Omega_{weights} + \beta \Omega_{sparsity} \tag{3.10}$$

where λ is the L_2 regularization coefficient and β is the sparsity regularization coefficient. The SAE decoding layer reconstructs the input features and attempts to minimize the cost function shown in eq (3.10). Once the SAE is trained, the decoding layer is removed and the encoded features are used as input to a softmax classifier. Softmax classifiers are a generalized version of the logistic classifier, and provide the probability that input features

belong to certain class.

$$\hat{y}(\mathbf{x}) = p(y = 1|\mathbf{z}) = \frac{e^{\mathbf{z}^T \mathbf{w}_1}}{\sum_{i=1}^{2} e^{\mathbf{z}^T \mathbf{w}_i}} \tag{3.11}$$

These probabilities are treated as the classifier scores as mentioned in the equation 3.1.

Parameter selection

LS has no parameters to optimize, and SWLDA parameters were selected from the literature (Krusienski et al., 2008). This work used 200 hidden units with $\lambda = 0.004$ and $\beta = 4$. We empirically chose the number of hidden units and the values of regularization coefficients. We also investigated the performance of stacked-SAEs (i.e., multiple layers of sparse autoencoders) and found negligible or no improvement in spelling performance. During the investigation of stacked-SAEs, we used data from all participants. Given the significant increase in computational complexity with stacked-SAEs, and the corresponding negligible or no improvement in performance, we used single-layer SAEs in this investigation.

3.2.5 Performance Evaluation

To evaluate the classifier performance we have computed the system spelling accuracy on each test sentence. Though the information transfer rate (ITR) (Wolpaw et al., 1998) or BCI utility metric (Dal Seno et al., 2010) are commonly used metrics for system performance evaluation, these metrics will only differ in the number of sequences are different for different participants or methods. Since we have used a fixed number of sequences (10 sequences) per character for all participants, a comparison using spelling accuracy will reflect the equivalent comparison using ITR or Utility metric. Comparing ITR or Utility metric for a fixed number of sequences for all participants is redundant if spelling accuracy is reported.

The accuracy for each method will be compared using multiple statistical tests. Firstly, accuracy for each method is compared using the Friedman test (Friedman, 1937) to find the difference between accuracy for different methods. The Friedman test (Friedman, 1937, 1940) is the non-parametric alternative to repeated-measures Analysis of Variance (ANOVA) that uses a group ranking method. The Friedman test is recommended method for comparisons

between classifiers (Demšar, 2006) because of its robustness to outliers and the fact that it does not assume normality of the sample means. If the Friedman test detects a significant difference between the obtained accuracy for different methods, a post hoc analysis is required to find which pairs in the group have significant differences.

For the post hoc analysis, we used mean rank based multiple comparison methods (Hochberg and Tamhane, 1987). Mean ranks post-test is recommended as post hoc Friedman test in many articles (e.g., Demšar 2006; Marascuilo and McSweeney 1967) and books (Gibbons and Chakraborti, 2011; Kvam and Vidakovic, 2007). However, alternative tests are also suggested in the literature (Benavoli et al., 2016). In (Benavoli et al., 2016), they discussed several drawbacks of mean ranks-based post hoc analysis and suggested to use a sign-test or the Wilcoxon signed-rank test (Wilcoxon, 1945) to overcome the identified drawbacks. The Wilcoxon signed-rank test is also suggested as an alternative for comparing two classifiers in (Demšar, 2006). Based on the results of the Friedman test, besides mean ranks based comparison, a post hoc analysis using the Wilcoxon signed ranks test (Wilcoxon, 1945) also performed as suggested in (Demšar, 2006) for multiple accuracy comparison. In our study, we used the Wilcoxon signed-rank test for multiple comparisons post hoc analyses, adjusting the p-value with the conservative Bonferroni correction method.

For the above statistical analysis, we used MATLAB as the primary analysis platform. For the Friedman test, `friedman.m` function of the Statistical toolbox was used. For the multiple comparison method, `multcompare.m` function was used. In case of the Wilcoxon signed-rank test based multiple comparison post hoc analysis, `signrank.m` function and a custom MATLAB implementation following the procedure described in (Benavoli et al., 2016) were used.

3.3 Results

As explained in section 3.2.3, we have assessed BCI performance using two different sets of electrodes. LS(64), SWLDA(64), and SAE(64) will denote the classification results using data from all 64 electrodes, while LS(32), SWLDA(32), and SAE(32) will denote the

classification results using data from 32 electrodes.

3.3.1 Friedman Test with Post Hoc Analysis

In our case, the null hypothesis of the Friedman test is "no significant difference between the accuracies of each method." The Friedman test yielded a $p-$value of $< 10^{-17}$, which allowed us to reject the null hypothesis.

Fig. 3.2 shows a graphical representation of the results from the post hoc analysis. It shows the mean ranks for each method from the Friedman test and the confidence intervals of the ranks from the post hoc analysis. This figure illustrates the significant or non-significant differences between each method. For instance, the rank of the method $LS(64)$ is significantly lower than the ranks of all other methods. The mean rank of SWLDA(64) is significantly better than the rank of LS(64), LS(32), and SAE(32).

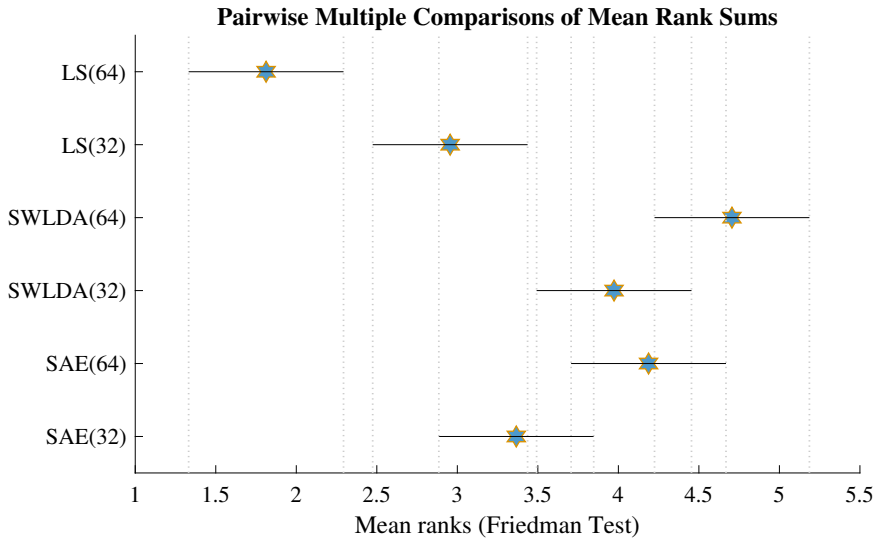


Figure 3.2: *Post hoc analysis: Mean ranks of BCI accuracy with confidence intervals for each methods using multiple comparison method (Hochberg and Tamhane, 1987). Higher numerical rank indicates better performance.*

3.3.2 Wilcoxon Signed-Ranks Test

Table 3.2 shows the p-values of pairwise multiple comparisons using the Wilcoxon signed-ranks test. The effect of the number of electrodes and the classification methods are reported in section 3.3.3 and 3.3.4, respectively, based on the results showed in Fig. 3.2 and Table 3.2.

Table 3.2: *Adjusted (Bonferroni correction (Hochberg and Tamhane, 1987)) P-values of pairwise multiple comparisons using Wilcoxon signed-ranks test.*

Methods	LS(64)	LS(32)	SWLDA(64)	SWLDA(32)	SAE(64)
LS(32)	$1.55e^{-04***}$	-	-	-	-
SWLDA(64)	$1.33e^{-08***}$	$8.29e^{-05***}$	-	-	-
SWLDA(32)	$3.67e^{-06***}$	$3.34e^{-04***}$	0.543	-	-
SAE(64)	$2.09e^{-08***}$	0.0047^{**}	1	1	-
SAE(32)	$1.77e^{-04***}$	1	0.0013^{**}	0.149	0.068

* Adjusted $p < 0.05$; ** Adjusted $p < 0.01$; *** Adjusted $p < 0.001$.

3.3.3 Effect of Number of Electrodes

All three classification methods were examined using EEG recordings from all electrodes and a reduced number of electrodes. Here, we will report the statistical test results for all channels vs the reduced number of channels. From the Table 3.2,

1. LS: The accuracy using all channels is significantly worse than using a reduced set of channels.

2. SWLDA: The set of all channels performed better than the reduced channel set, but the difference was not significant.

3. SAE: The set of all channels performed better than the reduced channel set, with the difference close to but above the usual significance threshold (adjusted $p = 0.068$, below

0.05 without Bonferroni correction).

3.3.4 Effect of Classification method

Here we will focus on the differences between different classification methods from Fig. 3.2 and Table 3.2. We compared the best-performing channel set for each method to ensure a fair comparison. Therefore, results for LS(32) were compared to SWLDA(64) and SAE(64).

1. LS vs SWLDA: SWLDA significantly outperformed LS (adjusted p-value $8.29e^{-05}$). The results from Table 3.2 and Fig. 3.2 are congruent in this case.

2. SWLDA vs SAE: SWLDA slightly outperformed SAE, but the different was highly non-significant (p-value 1).

3. SAE vs LS: SAE significantly outperformed LS (adjusted p-value 0.0047). The significant difference is also observed in Fig. 3.2.

3.3.5 Relation Between BCI Accuracy and P300 Latency Variations

Fig. 3.3 shows the relationship between BCI accuracy and the variance of CBLE using LS, SWLDA and SAE classifiers. To prevent over-cluttering, Fig. 3.3 includes only results using all electrodes. From this figure, it is evident that BCI performance is highly negatively correlated with the variance of CBLE. The negative correlation is consistent for all three classification methods. For LS, the correlation coefficient is -0.85 ($p < 10^{-15}$), for LDA correlation coefficient is -0.90 ($p < 10^{-20}$), and for SAE correlation coefficient is -0.87 ($p < 10^{-17}$).

3.3.6 Predicting BCI Accuracy Using CBLE

Fig. 3.4 shows the predicted accuracy using variances of CBLE (vCBLE) for LS, LDA, and SAE classifiers, respectively. BCI accuracy is predicted for each participant using the

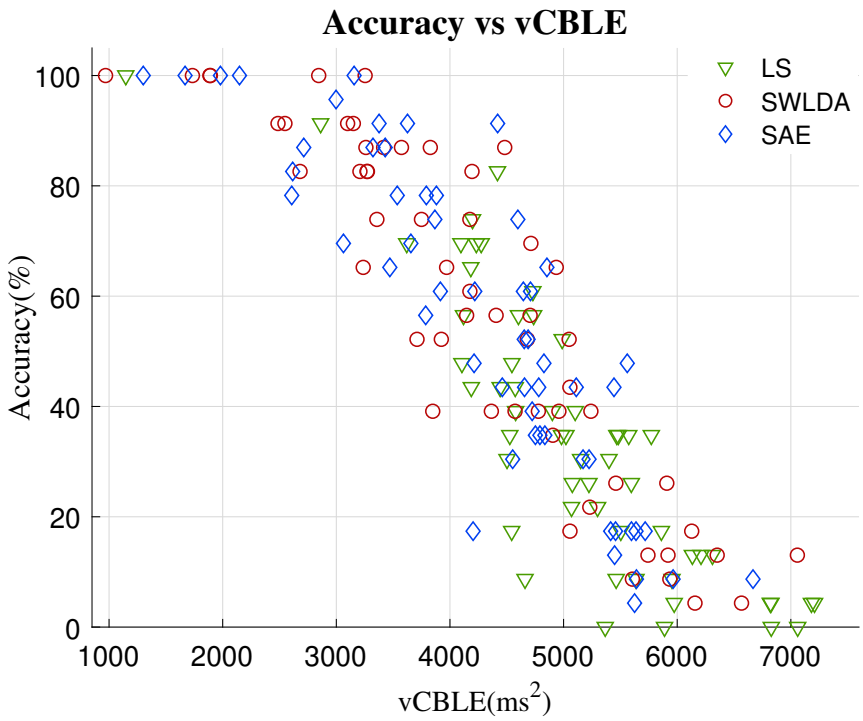
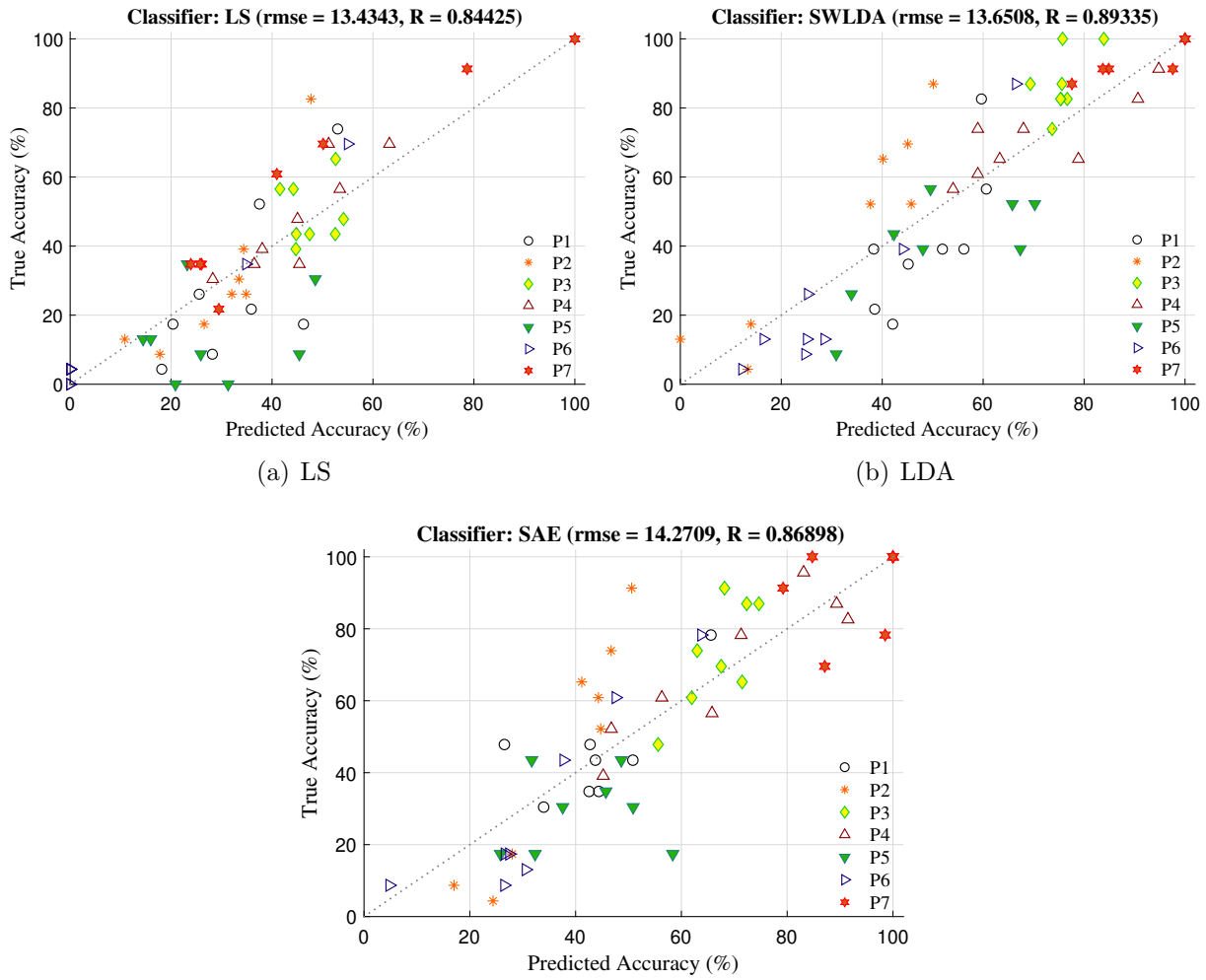


Figure 3.3: *Accuracy plotted against the variance of classifier-based latency jitter estimates (vCBLE) using LS, SWLDA and SAE classifiers.*

relationship between vCBLE and accuracy from all other users. Predicted accuracy using vCBLE for all the classifiers are significantly correlated with the actual accuracy. The root mean square errors (rmse) for three classifiers are $rmse_{LS} = 13.43$, $rmse_{LDA} = 13.65$, and $rmse_{SAE} = 14.27$, the coefficients of determination are $R^2_{LS} = 0.713$, $R^2_{LDA} = 0.798$, and $R^2_{SAE} = 0.755$. While these metrics leave some room for improvement, the randomness inherent in observing accuracy from a small number of characters prevents reaching perfect prediction. Even for "ideal" prediction (where the system correctly guesses the exact binomial parameter for each dataset), the resulting error would be expected to be $rmse_{ideal} = 8.0 - 8.4$ and $R^2 = 0.9 - 0.93$ based on our simulations.

3.4 Discussion

From the results shown in section 3.3.3, we observed that the effect of the number of electrodes is classifier-dependent. LS performed better with features from fewer electrodes



(a) LS

(b) LDA

(c) SAE

Figure 3.4: *Predicted BCI accuracy from vCBLE are plotted against true accuracy for three different classifiers. P1, P2, P3, P4, P5, P6, and P7 are indicating each participant.*

whereas both SWLDA and SAE performed better with features from all available electrodes (though the SWLDA and SAE effects were not statistically significant). This is consistent with theory - both SWLDA and SAE use inherent feature reduction techniques and should be less prone to the curse of dimensionality.

On our current dataset, the performance of SWLDA is significantly better than the performance of LS classification, which is congruent with the reported findings in (Krusienski et al., 2006). But SAE failed to prove better than the performance of SWLDA which is in contrast with the results in (Vařeka and Mautner, 2017). Furthermore, the required training

time for SAEs is often outweighing their performance (Vařeka and Mautner, 2017). Overall, SWLDA may be a better choice for P300 speller BCIs in terms of combined performance and practicability.

For our P300 speller dataset, we have observed a high negative correlation between P300 latency jitter and classification accuracy. This finding is consistent with our previously reported results in the earlier CBLE study, as well as the findings reported in another independent study (Aricò et al., 2014).

3.4.1 Limitations

CBLE is based on an assumption that the ERP complex shifts with a single latency which is estimated on a single-trial basis. This prevents any study of latency variation between different ERP components such as P3a and P3b. The same assumption prevents the study of single-trial spatial latency variations, if such variations exist.

3.5 Summary

From the results presented in section 3.3.3, we can conclude that the effect of the number of electrodes on performance is relative to the classification methods. LS classification works well with less features (data from fewer electrodes); SWLDA and SAE work well with a higher number of features (data from all available electrodes). Overall, SWLDA was the best classifier on our dataset, and also had the strongest correlation between BCI performance and vCBLE.

The similitude of the results from this dataset and the results reported in the CBLE original work strongly establishes that i) the P300 BCI system performance is negatively correlated with latency variations, ii) CBLE can be used to predict BCI accuracy. Moreover, the similar vCBLE and accuracy correlation supports the claim that CBLE is classifier independent.

Acknowledgement

Opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. The involvement of human participants with this research was approved by the Kansas State University Institutional Review Board under protocol No. 8320.

Funding

This work was supported in part by the National Science Foundation under Grant No. 1910526 and in part by the NIH Bridges and Developing Scholars Program (DSP) at the Kansas State University.

Chapter 4

Affective Brain-Computer Interfaces: A Tutorial to Choose Performance Measuring Metric

Copyright notice: The following text is reformatted from "Affective Brain-Computer Interfaces: A Tutorial to Choose Performance Measuring Metric." as submitted for publication Mowla et al. (2020b).

4.1 Introduction

The term *affective* (Picard et al., 1995) is a psychological concept referring to the experience of human emotion or feeling. Brain-computer interfaces (BCIs) are usually defined as a direct means of communication between the brain and external devices or systems which enable the brain signal to control some external activity (Wolpaw et al., 2002). Yet BCIs also allow investigation of brain activity and analysis of brain state. Affective Brain-Computer Interfaces (aBCIs) can be defined as a human affect estimation system from brain signals using BCIs. The interest in automatic detection of people's affective states has increased over the last few decades. Studies have shown that affective states play an important role

in human decision making (Forgas, 1995). The ability to manage one's affective states is also related to the ability of logical reasoning, learning and extracting important information (Salovey and Mayer, 1990). According to Goleman's model of emotional intelligence, having knowledge of your own affective states is a key factor behind personal and professional success (Goleman, 1996).

However, estimation of the affective state is a difficult task for several reasons. Human subjects do not always reveal their true emotions, and often inflate their degree of happiness or satisfaction in self-reports (Strack et al., 1990). Additionally, there is some ambiguity in understanding and defining affective states (Picard et al., 2001).

Facial expression analysis is one of the most popular methods (Pantic and Rothkrantz, 2000) for estimating affective states, but it is possible to deliberately fake facial expressions unrelated to one's true inner affective state. Therefore, as Picard argued, the estimation may have a high error rate if someone has the ability to disguise his or her emotion (Picard et al., 2001).

With the improvements in brain imaging techniques, there is a growing interest in relationships between affective states and brain activities. Investigating affective states using electroencephalogram (EEG) is becoming popular among researchers because EEG is one of the most convenient, non-invasive forms of recording brain activity. EEG also has high temporal resolution, which makes it a preferable candidate for fast affective state estimation (Niemic and Warren, 2002). Before using EEG-based BCIs to estimate affective states, one major challenge is to model affective states in a measurable and understandable scale. A current, widely accepted affective state model is the circumplex model of affect (Fig. 4.1), which was initially proposed by J. A. Russel (Russell, 1980). Finding distinct physiological patterns for each affective state has also always been a major topic of interest for affective computing researchers (Cacioppo and Tassinary, 1990). Picard argued that emotion consists of more complex, underlying processes rather than outward physiological expression (Picard et al., 2001).

Interest in EEG-based emotion recognition has increased over time and is still growing. Searching "EEG emotion recognition" in Google scholar gives 115000 results in March 2020.

Among them, there are 2100 just in the first quarter of 2020. Because these projects rely on individuals' emotional responses, the distribution of affective states (classes) is often uneven. However, most of these articles do not mention the class imbalance percentage but instead only report classification accuracy as a performance measuring metric. This creates a serious ambiguity and makes the results incomparable between works. For example, a publicly available database for emotion recognition known as the DEAP database (Koelstra et al., 2012) has been cited over 1600 times on March 2020, and using "EEG emotion recognition" search keywords within the DEAP-citing articles gives more than 1330 results. Out of those 1330 articles, at least 170 articles have included the DEAP dataset in their analysis. Out of those 170 articles, only approximately 33 articles mentioned or considered class imbalance. Classification accuracy, without considering class imbalance, is misleading for reasons we will present in this paper. Additionally, out of those 170 articles, only approximately 30 articles discussed statistical significance. This raised a few serious research questions:

1. Are those classification accuracies better than unskilled classifiers?
2. If so, are those accuracies significantly better than chance?
3. In the presence of class imbalance, what is the correct chance level?
4. What performance evaluation metric should be used in affect classification?

The main goal of this work is to investigate these questions. As a case study, we will use our investigations into EEG-based detection of binary (high/low) valence, arousal, and dominance in response to different sets of stimuli. For this investigation, we use both our own data as well as the previously mentioned, publicly available DEAP database (Koelstra et al., 2012).

Affective states can be elicited through visual (Lang et al., 2008), auditory (Lang and Bradley, 1999), and audio-visual stimuli (Baveye et al., 2015), among other methods. The emotional experience is more profound when visual presentations are combined with auditory stimuli, intermediate under visual stimuli and minimal during auditory stimuli (Güntekin and Başar, 2014). In our experiment, we used visual stimuli, the International Affective

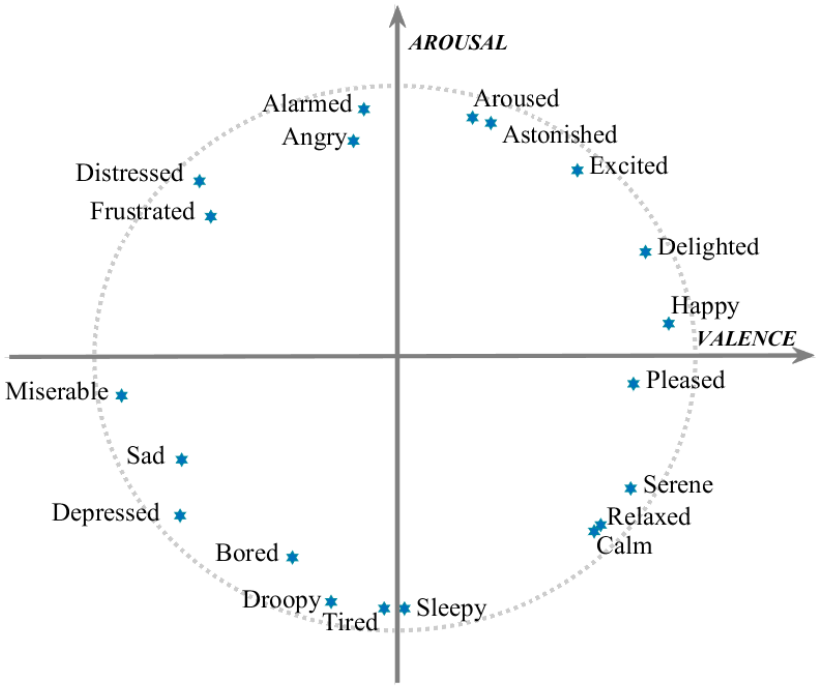


Figure 4.1: *An example of the circumplex model where emotions are expressed in the valence and arousal dimensions. Valence refers to how pleasant or unpleasant an emotion is, and arousal refers to how exciting or boring it is. Words are placed according to direct circular scaling coordinates for 28 affect words from Russel's article (Russell, 1980).*

Picture System (IAPS) (Lang et al., 2008), to evoke emotions. The DEAP database used audio-visual stimuli.

4.2 Related Work

In the field of affect recognition, a huge number of studies have been conducted on emotion recognition using EEG signals. With the improvement of dry electrodes, EEG is nearing or at the point of being a practical, out of the lab solution for affect recognition. More detailed EEG-based emotion recognition reviews can be found in (Wagh and Vasanth, 2019; García-Martínez et al., 2019). One major problem in EEG-based emotion recognition research is the lack of publicly available datasets. Consequently, researchers use their own data and as a result studies become more difficult to compare. To solve this problem, a few researchers developed publicly available datasets including the DEAP (Koelstra et al., 2012), USTC-

ERVS (Wang et al., 2014) and MAHNOB-HCI datasets (Soleymani et al., 2011). Among these datasets, the DEAP is the most cited and used for emotion recognition. Thus, we were motivated to use the DEAP dataset in this work.

Studies where DEAP was used as the benchmark dataset mostly used support vector machine (SVM) (Piho and Tjahjadi, 2018; Li et al., 2018; Soleymani et al., 2017; Zheng et al., 2017; Wang et al., 2017; Özerdem and Polat, 2017; Verma and Tiwary, 2017) for classification. The second most-used classification technique was the k-nearest neighbor (kNN) classifier (Piho and Tjahjadi, 2018; Zheng et al., 2017; Özerdem and Polat, 2017). Other classification techniques, such as deep convolutional neural network (Li et al., 2017), decision tree (García-Martínez et al., 2016), linear discriminate analysis (LDA) (Al Zoubi et al., 2018), logistic regression (Zheng et al., 2017), discriminative graph regularized extreme learning machine (GELM) (Zheng et al., 2017), back-propagation neural networks (BPNN) (Purnamasari et al., 2017), probabilistic neural networks (PNN) (Purnamasari et al., 2017), and multilayer perceptron (MLP) (Verma and Tiwary, 2017) have also been used to classify emotion on the DEAP dataset. Features used in these studies are statistical features: mean, standard deviation, variance, zero crossing rate (Verma and Tiwary, 2017; Liu et al., 2018; Torres-Valencia et al., 2017; Menezes et al., 2017), Hjorth parameters (Li et al., 2018; Mert and Akan, 2018), fractal dimension (Liu et al., 2018; Nakisa et al., 2018), Shannon entropy (Liu et al., 2018), spectral entropy (Verma and Tiwary, 2017; Liu et al., 2018), kurtosis (Hemanth et al., 2018), skewness (Yin et al., 2017a), different EEG band powers (Torres-Valencia et al., 2017; Yoon and Chung, 2013), relative power spectral density (PSD) for delta, theta, alpha, beta and gamma frequency bands (Wang et al., 2015), differential entropy (DE), differential asymmetry (DASM), rational asymmetry (RASM), asymmetry (ASM) (Zheng et al., 2017), wavelet coefficients (Özerdem and Polat, 2017), and higher order crossings (HOC) (Piho and Tjahjadi, 2018).

In the DEAP dataset, emotions are expressed in valence, arousal, and dominance dimensions on discrete 9-point scales. To design the classification model those scales need to be labeled. Here also, inconsistencies exist between different studies. Not only are different numbers of classes chosen by different groups, but even within the same number of classes

the thresholds are different. In these previously mentioned studies on the DEAP, classification labels were created by splitting the ratings into 3-class (1-3:negative, 4-6:neutral, and 7-9:positive) (Jirayucharoensak et al., 2014), 3-class (1-4.5:negative, 4.5-5.5:neutral, 5.5-9:positive) (Verma and Tiwary, 2017), 2-class (High/low, 4.5-9: high) (Daimi and Saha, 2014), 2-class (negative $\leq 5 <$ positive) (Wang et al., 2015), 2-class (negative $< 5 \geq$ positive) (Wang et al., 2017; Padilla-Buritica et al., 2016; Gupta et al., 2016), and 2-class (1-3: low and 7-9: high) (Menezes et al., 2017). Hence, the class imbalance in all these studies are different based on their individual approach when generating class labels.

Even though all these above-mentioned studies used the DEAP dataset, where significant class imbalance exists, very few studies have considered it while reporting results. Studies where class imbalance was considered mainly reported the F1 score (Koelstra et al., 2012; Soleymani et al., 2017; García-Martínez et al., 2016; Padilla-Buritica et al., 2016; Yin et al., 2017b) and a few other studies used receiver operating characteristic (ROC) (Piho and Tjahjadi, 2018; Menezes et al., 2017), area under ROC (AUC) (Li et al., 2018) and balanced accuracy (Clerico et al., 2018) along with accuracy metric. But AUC can be a misleading metric for a comparative study especially in the presence of variable class imbalance (Lobo et al., 2008) and computing the F1 score for multiclass classification is also not straightforward. For multiclass problems, F1 can be computed using macro-averaging or micro-averaging (Van Asch, 2013). The difference between macro- and micro-averaged F1 can be large; if studies do not report which was used then comparing results is impossible. For example, (Gupta et al., 2016) reported classification accuracies of 67% and 69% and F1 scores of 0.67 and 0.69 for valence and arousal, respectively. It is not clear how these F1 scores were calculated. F1 scores for both classes were not considered in that study which makes the study incomparable and provides misleading results.

To eliminate those above-mentioned problems we are suggesting to use balanced accuracy as the classification performance evaluation metric in high/low valence, arousal and dominance classification. To our knowledge, this has only been used in (Clerico et al., 2018). However, that study did not consider the lower bound of the credible intervals for balanced accuracy; here in this study we will further discuss using the posterior distribution of bal-

anced accuracy to compute credible intervals and perform statistical significance testing.

4.3 Data Description

In this work, we have used data from the publicly available DEAP dataset and EEG recordings from our lab.

4.3.1 Database for Emotion Analysis Using Physiological Signals (DEAP)

The DEAP is a publicly available, multimodal dataset consisting of 32-channel EEG, electrooculography (EOG), electromyography (EMG), galvanic skin response, respiration, plethysmograph, and temperature data (Koelstra et al., 2012). These signals were collected from thirty-two healthy participants, with an equal male-female ratio and an average age of 24.9 years. Data were recorded at a sampling rate of 512Hz and then pre-processed.

Minute-long music videos were used as emotional stimuli. After each video, participants were provided enough time to rate those videos for valence, arousal, and dominance on a discrete 9-point scale using self-assessment manikins (SAM) (Bradley and Lang, 1994). Each participant viewed forty videos.

4.3.2 Data collected at Brain and Body Sensing (BBS) lab

The BCI2000 (Schalk et al., 2004) system was used to present picture stimuli to the participants. Each picture was displayed for 6.7 seconds, followed by a 20.8s pause for participants' self-report. A total of 244 pictures were selected from IAPS (Lang et al., 2008) images; the average valence and arousal ratings reported in the IAPS manual of the selected pictures are shown in Fig. 4.2. Pictures were presented in six blocks, with breaks for participant comfort. EEG data were recorded using a Cognionics Mobile-72 EEG system with a sampling frequency of 600Hz. The Mobile-72 EEG system is a high-density mobile EEG system with

active Ag/AgCl electrodes placed according to the modified 10-20 system. Reference and ground were on the right and left mastoids, respectively.

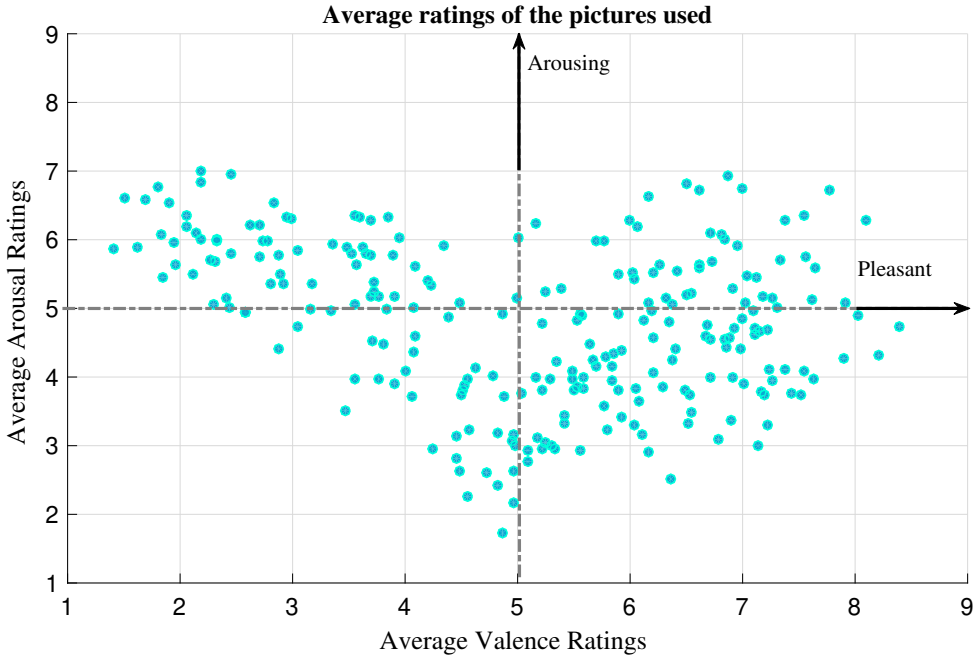


Figure 4.2: *Visualization of the average valence and arousal ratings (from the IAPS manual) (Lang et al., 2008) of picture sets used to collect data at the BBS lab.*

In total, we had nine participants. Data from two participants have been excluded due to one data entry error and one battery failure. All participants were healthy college students with an age range of 21 to 22 years. Each participant was shown 244 pictures through two or three different sessions. Each participant rated each stimulus for valence, arousal, and dominance on a discrete 5-point scale using self-assessment manikins (SAM) (Bradley and Lang, 1994).

4.3.3 Pre-processing

For the DEAP, both raw and pre-processed data are available for use. In this work, we will use this MATLAB-ready preprocessed version of the data. Pre-processing includes common-average referencing, down-sampling to 128Hz, band-pass filtering with the cut-off frequency at $(4.0 - 45.0)$ Hz, and eye blink artifact removal via independent component analysis. The

data contain 32 channels of EEG plus an additional eight channels of other physiological signals and the length of the time segment for each trial is 60 seconds. We have only used EEG recordings for classification. Data were then transformed into scalp surface Laplacian or current-source density (CSD) because it has been argued that CSD transformation gives a more sensitive index of individual variations in frontal asymmetry than other EEG recording montages and also helps to reduce non-frontal contributions to the frontal asymmetry (Velo et al., 2012; Allen and Reznik, 2015).

The data collected at the BBS lab was filtered using a finite impulse response (FIR) bandpass filter at $(4.0 - 45.0)$ Hz. Data were then transformed into scalp surface Laplacian or current-source density (CSD). To transform the EEG recordings into surface Laplacian, we used the CSD toolbox (Kayser and Tenke, 2006) which provides a MATLAB implementation and uses the spherical spline algorithm (Perrin et al., 1989) to estimate the surface Laplacian.

4.4 Methods

In our study, we will use $x(t) \in \mathcal{R}^T$ as the time series of a recording from a single electrode with N samples. The first and second derivatives of $x(t)$ with respect to time are $x'(t)$ and $x''(t)$, respectively. Standard deviation of $x(t)$, $x'(t)$ and $x''(t)$ are denoted as σ_x, σ_d, and σ_{dd}, respectively. Class labels are denoted by $c \in \{1, 2, \ldots, C\}$ and predicted class labels are denoted by y when classifying. $\mathbb{H}$ denotes entropy.

4.4.1 Feature Sets

Frequency domain features

Power spectral density and signal power at different frequency bands are popular for EEG-based affective state classification and have been used as features in several studies (Lin et al., 2010; Jenke et al., 2014). Spectral density and band powers can be computed using various algorithms, including Fast Fourier Transform, short-Time Fourier Transform, or Welch's power spectral density estimations algorithm. Here, we have used Welch's power spectral

density (PSD) estimation method (Welch, 1967) and then computed power in each band powers from the resulting PSD. The frequency ranges used for EEG bands varies slightly between different studies. In our analysis, the frequency ranges we have used are theta: (4-8) Hz, alpha: (8-12) Hz, beta-1: (12-18) Hz, beta-2: (18-30) Hz, gamma: (31- 63) Hz.

In a few studies, it has been argued that frontal EEG asymmetry can be a moderator and mediator of affective states (Coan and Allen, 2004; Allen et al., 2004). Frontal alpha asymmetry is mostly used as a discriminator between depressed and healthy individuals (van der Vinne et al., 2017). However, it also can be used for affective state classification. Here, we will use both frontal EEG asymmetry (1-50 Hz) and frontal alpha asymmetry (8-12 Hz) as features for classifying affective states. If R_p represents the signal power of electrodes located at the right frontal lobe and L_p represents the signal power of electrodes located at the left frontal lobe then frontal EEG asymmetry can be calculated from

$$\text{Frontal asymmetry} = \ln\left(\frac{R_p}{L_p}\right) \tag{4.1}$$

Another form of the frontal asymmetry is the normalized version of equation (4.1) and is written as

$$\text{Frontal asymmetry} = \ln\left(\frac{R_p - L_p}{R_p + L_p}\right) \tag{4.2}$$

Here, we have used equation (4.1) to find the frontal asymmetry. We have computed separately the frontal asymmetry index (FAI) and frontal alpha asymmetry index (FAAI). The frequency range of $0 - 64$Hz is used to compute FAI and the alpha band is used for FAAI.

We also used frontal theta beta ratios (TBR) as frequency domain features even though TBR has not been used previously for affective classification. But it has been reported to be related to affective traits (Putman et al., 2010). Because of their relation with affective traits, this study will examine the capability of frontal TBR in affect recognition. To compute the frontal TBR we used equation (4.3)

$$\text{TBR} = \ln\left(\frac{\theta_p}{\beta_p}\right) \tag{4.3}$$

here θ_p represents the theta band power and β_p represents the beta band power of electrodes located at the frontal lobe. Frequency ranges for beta-1 and beta-2 are used in β_p to compute TBR1 and TBR2, respectively.

Hjorth parameters

Hjorth parameters are time-domain features of EEG recording, proposed by Hjorth (1970). Hjorth parameters have been recently used in several studies (Mert and Akan, 2018; Jenke et al., 2014) as features for affective state estimation. The parameters are Activity, Mobility, and Complexity. Activity is simply the variance of the time signal. If the signal is denoted as $x(t)$, then Activity $= \sigma_x^2$ and is the measure of the squared standard deviation of amplitudes. Mobility measures the standard deviation of the slope with respect to the standard deviation of the amplitude. Mobility is defined as the square root of the ratio between the variances of the first derivative and the time signal. Complexity is a measure of how much the time signal deviates from a pure sine shape and is defined as the ratio between the mobility of the first derivative of the time signal and the mobility of the time signal.

$$\text{Mobility} = \frac{\sigma_d}{\sigma_x}$$
$$\text{Complexity} = \frac{\sigma_{dd}/\sigma_d}{\sigma_d/\sigma_x}$$

Here, we have used mobility and complexity as features. For each trial, there will be an equal number of mobility and complexity values and the number equals the EEG electrode number.

Entropy

Entropy is a measure of disorder in a system. In the case of EEG, entropy measures the irregularity in the signal. Spectral entropy of EEG recordings has been used to discriminate different affective states in other studies (Vakkuri et al., 2004) and it recently has been used in recognition of emotional states (Zheng et al., 2017). In this work, we will use spectral

entropy (SE), which is the normalized Shannon entropy of the power spectrum.

$$\text{Spectral Entropy} = -\frac{\sum_{i=1}^{N} p(X=i)\log_2 p(X=i)}{\log_2 N} \tag{4.4}$$

where X is denoting the power spectrum of the time series $x(t)$, $p(X)$ is the spectral distribution such that $\sum_{i=1}^{N} p(X=i) = 1$, and N is the number of frequency bins.

4.4.2 Classification

The ultimate goal for emotion estimation is a many-class classification or continuous-output regression. However, for this initial investigation, we focused on the easier binary classification problem, following multiple literature examples (Wang et al., 2017; Menezes et al., 2017; Wang et al., 2015; Daimi and Saha, 2014; Padilla-Buritica et al., 2016; Gupta et al., 2016). Thus, we use a two-class classification system for valence, arousal, and dominance. Participants in our experiments rated each axis from 1 to 5, we have labeled $ratings < 3$ as low valence, arousal, and dominance and $ratings \geq 3$ as high valence, arousal, and dominance. One participant never rated arousal less than 3, so for this participant (number 6) we shifted the split point from 3 to 4. In the DEAP database, participants rated each axis from 1 to 9; we have labeled $ratings < 5$ as low and $ratings \geq 5$ as high following the original work (Koelstra et al., 2012) and some other related studies (Liu et al., 2018; Clerico et al., 2018; Mohammadi et al., 2017).

In this study, support vector machine (SVM) and K-nearest neighbor (kNN) classifiers were used to test the affect recognition from EEG data. For our data, we will use 10-fold cross-validation. In case of DEAP data, we will use "Leave-One-Out" cross-validation technique. Which means at each step of the cross-validation, one sample was used as the test set and the rest were used as training set. The reason of using "Leave-One-Out" cross-validation in lieu of "K-fold" cross-validation is to maintain the congruity with other studies (Koelstra et al., 2012; Soleymani et al., 2017; Daimi and Saha, 2014; Clerico et al., 2018). These classifiers are the most commonly used techniques among published reports using

the DEAP dataset (e.g., Piho and Tjahjadi, 2018; Özerdem and Polat, 2017; Verma and Tiwary, 2017; Liu et al., 2018; Menezes et al., 2017; Wang et al., 2015; Clerico et al., 2018; Mohammadi et al., 2017).

Support vector machines (SVMs)

SVM uses a kernel trick and a separating hyperplane to create the support vectors. SVMs can be used for both regression and classification. In SVMs, with the observation vector $\mathbf{x}$ the predicted class label can be found using (Murphy, 2012)

$$\hat{f}(\mathbf{x}) = \text{sgn}\Big(\hat{w}_0 + \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x})\Big) \tag{4.5}$$

Where, $\alpha_i = \lambda_i y_i$, λ is the ℓ_1 regularization term and $k(\mathbf{x}_i, \mathbf{x})$ is the kernel function. For Gaussian kernel or radial basis function (RBF) kernel -SVM, the kernel function is defined by

$$k(\mathbf{x}_i, \mathbf{x}) = \exp\Big(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x})^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x})\Big) \tag{4.6}$$

Here we have used the MATLAB built-in function `fitcsvm` for SVM classifier with a medium Gaussian/RBF kernel. In `fitcsvm` 'Gaussian' and 'RBF' kernel are used interchangeably.

K-Nearest Neighbours (KNN)

kNN is a simple classification algorithm where an example is classified based on the plurality vote of its k number of nearest neighbors. The nearest neighbours are chosen by a distance metric. Distance metrics can be City block distance, Chebychev distance, Minkowski distance, Euclidean distance or Mahalanobis distance. Here we have used the built-in MATLAB function `knnsearch` using Euclidean distance with $k = 9$ using Euclidean distance. The kernel and hyperparameters for both classifiers are chosen empirically using a 15% test set partition strategy.

4.5 Performance Metrics

The most commonly used classification performance measurement metric is accuracy. Nevertheless, accuracy can be misleading, especially with the presence of class imbalance. In these situations, classifiers can learn from class label proportion rather than the features, a property sometimes known as "unskilled classification". In biased datasets, the unskilled performance is equal to the class imbalance. Thus, the same reported accuracy should be interpreted differently based on class bias. For example, consider a study reporting 80% accuracy in a two-class classification. This may be good performance on a balanced dataset but is at or below unskilled classification levels for biases $\geq 80\%$.

Comparing the performance of a similar classification task with different proportions of class labels is difficult. To make this kind of comparison meaningful, researchers suggest using other performance measuring metrics such as the Kappa statistic or area under ROC curve (AUC) for imbalanced data. But since the multiclass ROC curve analysis is not well developed (Lachiche and Flach, 2003), AUC is not recommended for multiclass problems (Sokolova and Lapalme, 2009). Moreover, the accuracy metric is the most widely used, and the most intuitive solution would be to make the accuracy metric meaningful by scaling down the baseline to be the performance of an unskilled classifier. One way to scale the baseline is to compute the balanced accuracy (Velez et al., 2007) where the accuracy in each class is considered separately.

4.5.1 Balanced Accuracy

If there are m number of classes, the balanced accuracy (Velez et al., 2007) is defined as

$$Balanced\ Accuracy = \frac{1}{m}\sum_{k=1}^{m}\frac{C_{kk}}{n_k} \tag{4.7}$$

Here, n_k is the total number of observations in class k and C_{kk} is the number of correctly classified observations in that same class label.

Since our focus is on two-class classification, here, k=2. If the classifier performs equally

well on both classes then the balanced accuracy will be exactly equal to the conventional accuracy (Velez et al., 2007; Brodersen et al., 2010). Since balanced accuracy is the average accuracy of each class, it is unaffected by the class imbalance and is more meaningful than the traditional accuracy metric. Further, it has the convenient property that an unskilled classifier always achieves $1/k$ accuracy regardless of class imbalance.

Although the traditional accuracy metric is a scaled binomial random variable, researchers often use a normal posterior distribution to compute credible intervals. The assumption behind the posterior normal distribution comes from the central limit theorem, where for a sufficiently large number of observations ($n \geq 30$), a binomial distribution can be approximated using the normal distribution. Nonetheless, this approximation becomes unreliable for small n. Particularly in the case of imbalanced data, the number of observations for the minority class can be smaller than the required number for the normal approximation. Therefore, finding chance performance and the credible interval of the misclassification rate for balanced accuracy is not as straightforward as it is in the case of traditional accuracy. For the two-class classification case, it is a combination of two separate distributions. In a multi-class scenario, accuracy in each class will have a separate distribution.

Credible Intervals of Balanced Accuracy

If the probability of predicting correct classes of a classifier denoted by $\mathcal{A}$ with a prior distribution $p(\mathcal{A})$, then the posterior is expressed as $p(\mathcal{A}|\mathcal{D})$ on observed data $\mathcal{D}$. Lets assume $y = 1$ and $y = 0$ for correct and incorrect predictions, respectively. Now the classification predictions can be written as $y_1, y_2, \ldots, y_n$ which resembles the results of a Bernoulli experiment. So we can write

$$\begin{aligned} p(y_k|\mathcal{A}) &= \text{Bern}(y_k|p(\mathcal{A}) \\ &= \mathcal{A}^{y_k}(1-\mathcal{A})^{1-y_k} \end{aligned} \tag{4.8}$$

If the total number of success (correct predictions) of a Bernoulli trial $y_1, y_2, \ldots, y_n$ is c, then it follows a Binomial distribution.

$$\begin{aligned} p(c|\mathcal{A}, n) &= B(c|\mathcal{A}, n) \\ &= \binom{n}{c} \mathcal{A}^n (1-\mathcal{A})^{n-c} \end{aligned} \tag{4.9}$$

This suggests choosing Beta density as the prior of $\mathcal{A}$ since it is the conjugate prior of the Binomial distribution. This implies

$$\begin{aligned} p(\mathcal{A}) &= \text{Beta}(\mathcal{A}|a, b) \\ &= \text{Beta}(\mathcal{A}|1, 1) \end{aligned} \tag{4.10}$$

Now the posterior can be written using Bayes theorem as

$$\begin{aligned} p(\mathcal{A}|c, n) &= \frac{p(c|\mathcal{A}, n)p(\mathcal{A})}{p(c)} \\ &= \frac{B(c|\mathcal{A}, n) \times \text{Beta}(\mathcal{A}|1, 1)}{p(c)} \end{aligned} \tag{4.11}$$

From equation 4.11, we obtain the posterior $p(\mathcal{A}|c, n) = \text{Beta}(\mathcal{A}|c+1, n-c+1)$ and the posterior $(1-\alpha)100\%$ credible interval is (Carrillo et al., 2014)

$$\left[F^{-1}_{\text{Beta}(c+1, n-c+1)}(\alpha/2); F^{-1}_{\text{Beta}(c+1, n-c+1)}(1-\alpha/2) \right] \tag{4.12}$$

where $F^{-1}_{\text{Beta}(\cdot)}(\cdot)$ is the inverse density function of the Beta distribution and for 95% credible interval, $\alpha = 0.05$. In a multiclass scenario, each class has the distribution shown in equation (4.11). To find the posterior of the balanced accuracy $m-$fold convolution is used for m classes. Numerical approximations are used to compute the posterior since analytical forms are not available for the $m-$fold convolution. In this work we have used a MATLAB routine to compute the credible intervals of balanced accuracy provided in (Brodersen et al., 2010).

4.5.2 F1 Measure

Another alternative performance evaluation metric is the F1-measure which has been used in some papers using the DEAP dataset (Koelstra et al., 2012; Soleymani et al., 2017; Daimi and Saha, 2014). The F-measure was originally proposed by Van Rijsbergen (Van Rijsbergen, 1979) and is defined as (Chinchor, 1992)

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{4.13}$$

where P and R denotes precision and recall and are defined as $P = tp/(tp + fp)$, $R = tp/(tp + fn)$ ($tp \rightarrow$ true positive, $fp \rightarrow$ false positive, $fn \rightarrow$ false negative). β is a parameter to control balance between P and R. When $\beta = 1$, F_1 becomes the harmonic mean of precision and recall. Hence the F_1 measure is

$$F_1 = \frac{2PR}{P + R} \tag{4.14}$$

Since P and R are calculated considering one class as a positive class, P and R have to be calculated per class and hence the F1 measure as well. P and R per class can be calculated in two ways: microaveraging and macroaveraging. Microaveraging aggregates the individual true positives, false positives, and false negatives of each classes to calculate the P and R.

$$\begin{aligned}
miP &= \frac{\sum_{k=1}^{m} C_{kk}}{\sum_{k=1}^{m} C_{kk} + \sum_{k=1}^{m} \sum_{\substack{j=1 \\ j \neq k}}^{m} C_{jk}} \\
miR &= \frac{\sum_{k=1}^{m} C_{kk}}{\sum_{k=1}^{m} C_{kk} + \sum_{k=1}^{m} \sum_{\substack{j=1 \\ j \neq k}}^{m} C_{kj}} \\
miF_1 &= \frac{2 \cdot miP \cdot miR}{miP + miR}
\end{aligned} \tag{4.15}$$

An alternative technique is known as macroaveraging. In macroaveraging, P and R are calculated for each classes and then F_1 for each class is computed using P and R of individual classes, and finally the macroaverage is the simple average of individual class F_1 scores.

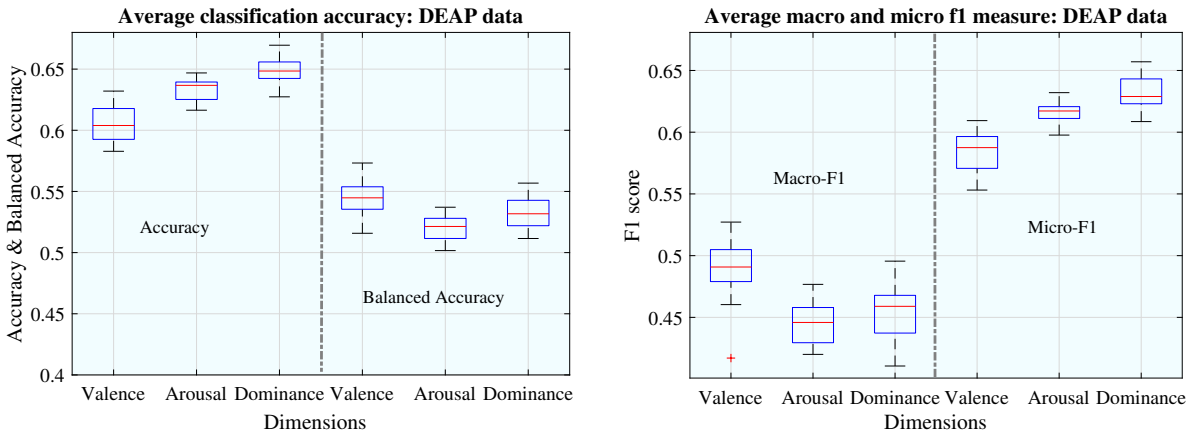
$$
\begin{aligned}
P_k &= \frac{C_{kk}}{C_{kk} + \sum\limits_{\substack{j=0 \\ j \neq k}}^{m} C_{jk}} = \frac{C_{kk}}{\sum\limits_{j=1}^{m} C_{jk}} \\
R_k &= \frac{C_{kk}}{C_{kk} + \sum\limits_{\substack{j=0 \\ j \neq k}}^{m} C_{kj}} = \frac{C_{kk}}{\sum\limits_{j=1}^{m} C_{kj}} \\
maF_1 &= \frac{1}{m} \sum_{k=1}^{m} \frac{2 \cdot P_k \cdot R_k}{P_k + R_k}
\end{aligned}
\tag{4.16}
$$

The difference between miF_1 and maF_1 can be significant. Macro-averaging gives equal weight to each class, whereas micro-averaging gives equal weight to each per-class classification decision. Since F_1 measure ignores true negatives, the influence of large classes is higher than small classes in micro-averaging (Manning et al., 2010). However, the F_1 measure's harmonic means suggest that the averaging should be over the per-class classification decision of each instances. And in that case macro-averaging is not consistent with the original definition of the F_1 measure (Powers, 2015). Hence we yet do not have a convincing argument for choosing between miF_1 and maF_1 for multiclass classification.

4.6 Results

Because we have used seventeen different feature sets, it is not feasible to show all results here. To summarize the results, the classification results are averaged over all participants for each feature set. Those average classification accuracies, and other performance metrics for different feature sets, are presented in Fig. 4.3, Fig. 4.4 and Table 4.1. All the results presented here are for the SVM classifier since it performed better than the kNN approach.

4.6.1 DEAP Dataset



(a) Accuracy and balanced accuracy

(b) Macro and micro f1 measure

Figure 4.3: *Average classification rate of all participants in valence, arousal and dominance recognition for different features using DEAP dataset.*

Fig. 4.3(a) shows the average classification accuracies and balanced accuracies for different feature sets using SVM. The mean classification rates for all features are 0.604, 0.637, and 0.648 for valence, arousal, and dominance, respectively. These results are very comparable with the results reported in DEAP original work (Koelstra et al., 2012) and other related studies (Soleymani et al., 2017; Daimi and Saha, 2014). But then if we check the balanced accuracies on the right side of the Fig. 4.3(a), we will observe very different results. The mean classification rate in balanced accuracies for all feature are 0.544, 0.521, and 0.531 for valence, arousal and dominance respectively. These results are very different than the results with the simple accuracy metric except for valence recognition. The average class bias rate in these three dimensions are 0.59, 0.64 and 0.66 for valence, arousal, and dominance.

Fig. 4.3(b) shows the average macro- and micro-averaged f1 measure for different feature sets using SVM. The mean macro-f1 for all feature are 0.49, 0.45 and 0.46 for valence, arousal, and dominance, respectively. On the contrary, the mean micro-f1 for all feature are 0.59, 0.62 and 0.63 for valence, arousal, and dominance, respectively. The best classification rate in the valence dimension is achieved using beta band power as a feature, as we found using balanced accuracy. For valence, the average across all participants macro-f1 for BetaP

feature is 0.53 and the micro-f1 is 0.61. For arousal, the average across all participants macro-f1 for ThetaP feature is 0.48 and the micro-f1 is 0.63. For dominance, the average across all participants' macro-f1 for the TBR1 feature is 0.50 and the micro-f1 is 0.65.

Table 4.1: *The average for all participants classification rate in terms of balanced accuracy and the lower bound of the 95% credible intervals of balanced accuracies for different feature sets.*

Features	**Valence**		**Arousal**		**Dominance**	
	Balanced Accuracy (BAcc)	Lower bound of BAcc	Balanced Accuracy (BAcc)	Lower bound of BAcc	Balanced Accuracy (BAcc)	Lower bound of BAcc
PASI	0.5448	0.4297	0.5279	0.4227	0.5317	0.4262
FAI	0.5222	0.4089	0.5118	0.4130	0.5219	0.4178
TBR1	0.5479	0.4267	0.5247	0.4235	**0.5568**	**0.4435**
TBR2	0.5381	0.4198	0.5109	0.4090	0.5220	0.4142
ThetaP	0.5388	0.4211	**0.5371**	**0.4336**	0.5302	0.4206
AlphaP	0.5432	0.4286	0.5238	0.4281	0.5492	0.4432
BetaP	**0.5732**	**0.4531**	0.5303	0.4263	0.5370	0.4247
GammaP	0.5585	0.4381	0.5282	0.4265	0.5409	0.4323
TBR-C	0.5663	0.4482	0.5318	0.4263	0.5550	0.4439
TABG	0.5578	0.4401	0.5090	0.4122	0.5349	0.4301
Hjorth	0.5323	0.4159	0.5268	0.4268	0.5204	0.4104
PASI+FASI	0.5473	0.4355	0.5214	0.4207	0.5338	0.4307
Avg-Entropy	0.5158	0.4177	0.5200	0.4312	0.5176	0.4269
PSD	0.5525	0.4451	0.5148	0.4259	0.5292	0.4361
BARatio	0.5230	0.4077	0.5016	0.4054	0.5115	0.4059
All	0.5517	0.4447	0.5069	0.4178	0.5229	0.4290
All-PCA	0.5365	0.4160	0.5166	0.4086	0.5484	0.4329

Table 4.1 shows the average balanced accuracies and lower bound of the 95% credible intervals of balanced accuracies for different feature sets using equation (4.12). All results are for the SVM classifier. The highest obtained balanced accuracy across all dimensions is 0.5732, achieved for valence recognition using beta band power. Unfortunately, the average lower limit of the credible intervals, in this case, is not above 0.5 (random chance). Though the average provides an overall recognition rate, it does not reflect the performance of individual participants. Explaining results for all features would be cumbersome; here we will explain classification results for each participant for only the best feature in each dimension. For valence, beta band power worked best. Using this feature, the balanced

accuracy obtained for a participant (s10) with 0.75 and the lower bound of the credible interval is 0.622, which means that the valence classification rate is significantly above chance for this participant. Out of 32 participants, balanced accuracy is greater than 0.5 for 23 participants. For 8 of these participants, the lower bound of the credible interval is greater than 0.5. For arousal, theta band power worked best. Using the thetaP feature, the highest balanced accuracy obtained for a participant (s17) is 0.73 and the lower bound of the credible interval is 0.60, which means the arousal classification rate is significantly above chance for this participant. For 21 participants, observed balanced accuracy is greater than 0.5. However, only 4 participants were the lower bound of the credible interval greater than 0.5. For dominance, theta beta-1 ratio worked best. Using TBR1, the highest balanced accuracy obtained for a participant (s17) was 0.74 with a lower bound of 0.61, which means the dominance classification rate is significantly above chance for this participant. For 24 participants, balanced accuracy is greater than 0.5. Yet again, only for 4 participants was the lower bound of the credible interval greater than 0.5.

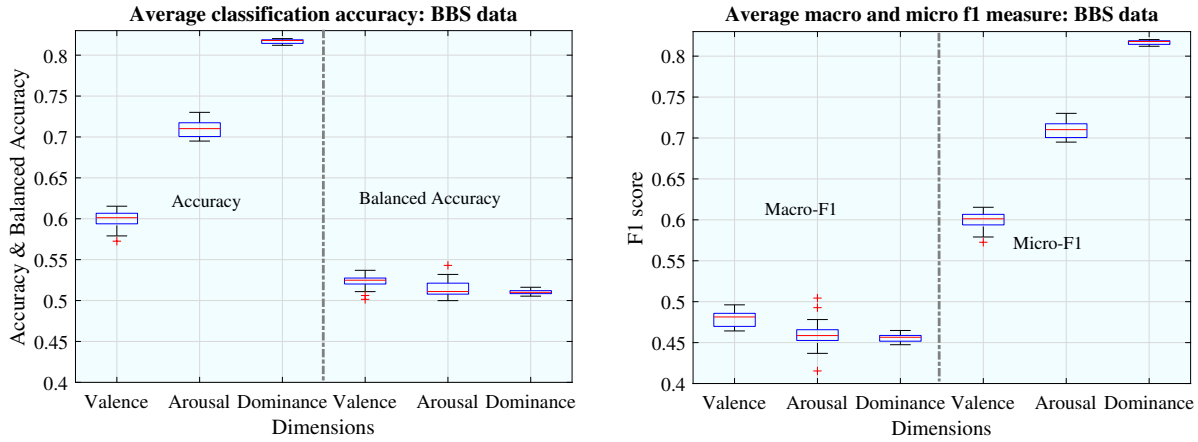
Table 4.2 shows the affect recognition rate in terms of balanced accuracy, micro and macro averaged F1 score and also compared with the original work (Koelstra et al., 2012) and some other related studies. These compared studies used Gaussian naive Bayes classifier (Koelstra et al., 2012) and Gaussian/RBF kernel SVM (Soleymani et al., 2017; Daimi and Saha, 2014; Clerico et al., 2018) for affective classification. Rather than presenting the best results in

Table 4.2: *The classification rate in terms of balanced accuracy and micro and macro F1 scores of affect recognition compared to the DEAP dataset original work and related studies. The results shown here are average of all participants for beta band power (BetaP) features.*

	Valence			**Arousal**			**Dominance**		
	bAcc	miF1	maF1	bAcc	miF1	maF1	bAcc	miF1	maF1
Koelstra et al., 2012	–	–	0.563	–	–	0.583	–	–	–
Daimi and Saha, 2014	–	–	0.550	–	–	0.570	–	–	0.552
Soleymani et al., 2017	–	–	0.645	–	–	0.570	–	–	0.533
Clerico et al., 2018	0.604	–	–	0.583	–	–	0.564	–	–
Current study	0.573	0.610	0.530	0.530	0.620	0.460	0.537	0.630	0.460

each dimension, we chose to present results for one specific feature set for consistency. The results presented under the current study are for beta band power (BetaP) feature using an SVM classifier. Note that our comparison studies seem to have picked the best result in each dimension for their reported results (only Clerico et al., 2018 unambiguously stated this).

4.6.2 Data from BBS lab



(a) Accuracy and balanced accuracy

(b) Macro and micro f1 measure

Figure 4.4: *Average classification rate of all participants in valence, arousal and dominance recognition for different features using BBS data.*

The data collected at the BBS lab using IAPS came from seven participants. For 2-class classification, the average class-bias were 0.60, 0.72, and 0.82 for valence, arousal, and dominance, respectively. For valence with SVM, the best 2-class classification results were obtained using gamma-band power considering the average of all participants. The obtained accuracy was 0.62 and the balanced accuracy was 0.54. The macro and micro averaged f1 scores were 0.49 and 0.60, respectively.

For arousal with SVM, the best 2-class classification results were obtained using the power asymmetry index (PASI) considering the average of all participants. The obtained accuracy was 0.73 and the balanced accuracy was 0.54. The macro and micro averaged f1 scores were 0.50 and 0.71, respectively.

For dominance with SVM, the best 2-class classification results were obtained using beta

band power considering the average of all participants. The obtained accuracy was 0.82 and the balanced accuracy was 0.52. The macro and micro averaged f1 scores were 0.46 and 0.82, respectively.

4.7 Discussion

For the DEAP, the average class bias or majority class percentage in a 2-class classification scenario for valence, arousal and dominance are 0.59, 0.64 and 0.66 respectively. We have argued that class imbalance is important to understand the results of the classifier and should be reported. Performance metrics that include or account the class-biases are thus preferred to use. Any metric that ignores class imbalance will mislead readers. To illustrate this, consider the results from Table 4.1 where balanced accuracies and its lower bound of the 95% credible interval were presented for different feature sets for DEAP data using SVM. The best average classification accuracy for all participants in the valence dimension was 0.602 using beta band power as a feature, whereas the balanced accuracy, in this case, was 0.573. Without knowing the class bias and considering the accuracy metric, one might think the result is promising. But the lower bound of the 95% credible interval of balanced accuracy shows that the classification rate can not be claimed as statistically significant.

However, class imbalance for each participant for all three-dimension (valence, arousal, dominance) would be cumbersome and impractical to report. The biases mentioned earlier were averaged across all participants. Since affective state estimation is a participant-specific task, averaged results do not reflect individual performances. So comparisons using average results are not meaningful. Hence, we need something else which can address both the class imbalance problem and make the average performance meaningful. Considering those above-mentioned problems, balanced accuracy is a promising candidate since the baseline performance for balance accuracy is the same (50%) across all dimensions(valence, arousal, dominance) for all participants. Thus, balanced accuracy will make results easier to understand and compare. For example, just looking at the results in Table 4.1, we can easily conclude that the valence recognition rate is better than arousal and dominance recognition.

Statistical comparison between the balanced accuracies for valence, arousal and dominance presented in Table 4.1 is done by using MATLAB inbuilt function `ttest2`. Two-sample t-test resulted in the rejection of the null hypothesis (two groups are equal) when comparing valence and arousal. The valence recognition rate is significantly better than the arousal and dominance recognition rate with p-values 0.035 and $7.44e^{-06}$. The dominance recognition rate is also significantly better than arousal with p-value of 0.031. These three two-sample t-tests suggested that valence has the highest recognition rate and arousal has the lowest for the DEAP dataset.

Averages for all participants of the balanced accuracies, macro, and micro f1 measure were compared with other related studies in Table 4.2. Since they have not discussed the methods of statistical analysis, here we will use our obtained results shown in Table 4.1 for discussion. Our average balanced accuracies are very similar to the highest balanced accuracy reported in (Clerico et al., 2018). They claimed that all the reported balanced accuracies were better than random voting classifiers with $p < 0.05$. This statement is true if we perform statistical analysis considering results from all participants as a group rather than individual participants. The number of participants with balanced accuracy above 0.5 is 25 for valence using all frequency band powers, 21 for arousal and 20 for dominance. In this case the probability that overall balanced accuracy is above chance are 0.66, 0.66 and 0.63 with intervals $(0.47 - 0.82)$, $(0.47 - 0.82)$, and $(0.44 - 0.79)$ for valence, arousal and dominance, respectively. But the significance of the experiment as a whole does not capture the significance of each participant's performance. Hence, just based on these statistics we are not comfortable to claim the accuracies are above chance. Rather we suggest using the probability of individual participants' performances being above chance to claim the results are significant. Using the number of participants that are significantly above chance, we have 6 for valence, 3 for arousal and 4 for dominance out of 32 participants. That tells us that the probabilities of a participant's classification accuracy being significantly above chance for valence, arousal and dominance are 0.19, 0.09 and 0.13 bounded by $(0.07 - 0.36)$, $(0.02 - 0.25)$ and $(0.04 - .29)$, respectively. These are not very encouraging, as valence is only above the typical 0.05 threshold. This low rate of significant performance may be of

concern for the EEG based affective computing community, and as a community, we need to be more careful while reporting results.

4.8 Conclusion

In this work, experimental results for affective state estimation performance were shown using the publicly available DEAP database and our own data. Our results with DEAP data were also compared with the results reported in a few related studies (Soleymani et al., 2017; Daimi and Saha, 2014; Clerico et al., 2018). Among these studies, Clerico et al. (2018) used balanced accuracy as the performance measuring metric and others used macro averaged f1 score. Some other studies (Yin et al., 2017a,b) also reported f1 score for the low valence/arousal/dominance class. But since their f1-score was computed using only one class, the results are not comparable. In most of the related studies, only classification accuracy has been reported (e.g. (Mohammadi et al., 2017; Atkinson and Campos, 2016)), which makes the results hard to interpret in the presence of class bias and also incomparable with other studies.

In conclusion, we suggest using balanced accuracy and its posterior distribution as the performance evaluation metric for emotion estimation. Though F1 measure is a popular choice, it is not yet well established which F1 measure (macro/micro) we should use for multiclass classification. As our results demonstrate, that choice is important. Further, if macro-averaging is chosen, the statistical significance of the metric is not well understood.

In contrast to the F1 measure, balanced accuracy has several advantages. First, balanced accuracy does not have a "preferred class" and is thus comparable between groups. Second, the credible bounds can be calculated using known formulas. Third, the extension to large numbers of classes is straightforward. Fourth and finally, balanced accuracy is insensitive to class bias and always has the intuitive 1/k chance performance for unskilled classifiers.

We note that traditional accuracy metrics would have classified the performance of many more of our participants as statistically significant, relative to the number classified this way by balanced accuracy. Nevertheless, we maintain that balanced accuracy is far less mis-

leading, and that the traditional accuracy metric substantially over-estimates performance is these unbalanced datasets.

Acknowledgement

Opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. The involvement of human participants with this research was approved by the Kansas State University Institutional Review Board under protocol No. 8328. The authors would like to thank our participants for enduring long EEG sessions.

Funding

This work was supported in part by Kansas State University faculty startup funds and in part by the National Science Foundation under Award No. 1910526.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Brain-Computer Interfaces (BCIs) have immense potential as an alternative tool for communication and control. It has already been long established that this kind of technology has considerable medical implications for patients with severe motor disabilities. The recent trend of integrating BCIs with augmented reality (AR) and virtual reality (VR) environment, and video-game environment (Lécuyer et al., 2008; Van Erp et al., 2012) manifests the higher potentials of BCIs in non-medical applications. In my opinion, this is "the" technology which will make science fiction into reality and allow us to control our environment by just thinking about it. To reach that far, however, we, the researchers, need to improve the BCIs from their current stage. In this dissertation, a small effort was made to enhance the BCIs further and make them usable.

This dissertation also attempted to incorporate BCIs with affective computing. Affective computing is also an emerging field of research that will enable computers to understand human affect. The area is increasing by combining various multidisciplinary approaches such as facial image analysis, heart rate detection, skin temperatures, and similar other measures. But BCIs also can be very useful in this area.

5.2 Contributions

Major contributions of this dissertation are summarized below and the outcomes (i.e., publications) of this dissertation-related research are listed in section 5.2.4.

5.2.1 Performance Enhancement of P300 Speller

To enhance the P3 BCI's performance, classifier-based latency estimation (CBLE) and a wavelet transform were used to provide information about latency jitter to a second-level classifier. Three second-level classifiers were tested: least squares (LS), step-wise linear discriminant analysis (SWLDA), and support vector machine (SVM). Of these three, LS and SWLDA performed better than the original online classifier. The resulting combination demonstrated improved detection of brain responses for many participants, resulting in better BCI performance.

5.2.2 Comparison of Classification Techniques to Predict BCI Accuracy Using CBLE

To investigate the role of latency jitter on BCI system performance, Thompson et al. (2012) proposed the classifier-based latency estimation (CBLE) method. In (Thompson et al., 2012), CBLE was based on the least-squares (LS) and stepwise linear discriminant analysis (SWLDA) classifiers. This dissertation extends the CBLE method using sparse autoencoders (SAE). Here, the newly-developed sparse autoencoder-based CBLE method was applied to a newly-collected dataset. Findings include a significant ($p < 0.001$) negative correlation between BCI accuracy and estimated latency jitter. Furthermore, the SAE-based CBLE method was used to predict BCI accuracy and the resulting coefficient of determination (R^2) was 0.755. In contrast to Vařeka and Mautner (2017), we have not observed an improvement in P300 classification accuracy using sparse autoencoders. This confirms the findings of Krusienski et al. (2006), that SWLDA provides the best overall performance for P300 classification.

5.2.3 Performance Assessment of Affective BCIs

Automatic recognition of affective states has a wide variety of applications, such as human-computer interaction scenarios where users' affective state is important, emotion sensitive automatic tutoring systems, and entertainment and gaming programs where the program can respond based on the users affective state. But the current research on EEG-based affect recognition lacks a proper evaluation metric, which made the results from different studies incomparable. To remove that hindrance, an evaluation method based on using balanced accuracy was proposed with a detailed analysis using other candidate metrics (e.g., F1 measure, area under curve). The proposed methodology will serve as a tutorial guideline for EEG based affective computing research.

5.2.4 Publications

1. Mowla et al., 2016. "Boosting BCI accuracy using wavelet enhanced CBLE scores as a classifier feature." *Proceedings of the 6th International Brain-Computer Interface Meeting.*

2. Mowla et al. (2017). "Enhancing p300-BCI performance using latency estimation", *Brain-Computer Interfaces*, pages 1–9.

3. Mowla et al. (2018b). "Evaluation and Performance Assessment of the Brain-Computer Interface System", chapter 33, pages 634–649. CRC Press, 2018.

4. Mowla et al. (2020a). "Estimation of inter-trial p300 latency variability using an autoencoder-based CBLE method". *Manuscript submitted for publication.*

5. Mowla et al. (2020b). "Affective brain-computer interfaces: Balanced accuracy as the performance measuring metric". *Manuscript submitted for publication.*

5.3 Limitations

Limitations are provided in each chapter. Overall the main limitations are: *(i)* BCIs for communication are intended for users with neuromuscular disorders, but the data collected at the BBS lab are from healthy college students. *(ii)* Another unavoidable limitation of the affective experimental design is to stimulate consistent affective state across subjects. Since emotional feelings are very subjective, it is very difficult to generalize ratings for each stimulus (Coan and Allen, 2007). For example, a snake picture can be arousing to one person whereas someone else might not be aroused at all. Another person may experience a feeling of fear looking at a snake picture whereas someone else may feel affectionate.

5.4 Future Work

The experiments described in Chapter 3 and 4 are still being conducted. The P3 experiment consists of two stimulus presentation paradigms. One is standard row/column intensification, and another one is replacing the row/column with Albert Einstein's iconic tongue face picture. Due to the small number of participants, this dissertation work did not attempt to compare the differences in BCI performance between these two paradigms. One of the major future tasks is to compare the variances of estimated latency using the proposed SAE based CBLE method for these two paradigms.

The affective experiment is also ongoing using sound stimuli. The primary future task of this part will be to compare affect recognition using picture and sound stimuli for the same participant. But before analyzing the effectiveness of sound and image as stimuli, the first task will be to explore other EEG features further. An alternative approach to using balanced accuracy would be to use random or selective oversampling to produce synthetically balanced datasets. A future research project could be comparison of the effectiveness of these approaches.

The affective classification results reported in this dissertation demonstrate that the traditional feature sets are not very useful in identifying human affect from EEG. That suggest

investigating the effectiveness of non-traditional EEG features in identifying affect. A potential analysis method could be using brain connectivity patterns for different affective stimuli. The effective connectivity refers to the directional influence that one neural system exerts over another (Friston, 2009). Effective connectivity involves inferred causality and describes the directional effects of one neural system over another. Since the connectivity measure depicts the directional effect from one location to another, a 64-channel EEG data will produce $64^2 - 64 = 4032$ directional interactions in each time-frequency combination, which makes such analysis very computationally expensive. A significant challenge in this future direction will be to determine how to use the connectivity features for affect recognition.

Bibliography

Richard Caton. Electrical currents of the brain. *The Journal of Nervous and Mental Disease*, 2(4):610, 1875.

Adolf Beck and Napoleon Cybulski. Further research on the electrical phenomena of the cerebral cortex in monkeys and dogs. *Rozprawy Wydzialu Matematycvno-przyrodniczych Polska Akademia*, 32:369, 1891.

Hans Berger. Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

American Electroencephalographic Society. Guideline thirteen: Guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology*, 11(1):111–3, 1994.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain-Computer Interfaces for Communication and Control. *Clinical neurophysiology*, 113(6):767–791, 2002.

Oliver Tonet, Martina Marinelli, Luca Citi, Paolo Maria Rossini, Luca Rossini, Giuseppe Megali, and Paolo Dario. Defining brain–machine interface applications by matching interface performance with device requirements. *Journal of neuroscience methods*, 167(1): 91–104, 2008.

Jong-Hwan Lee, Jeongwon Ryu, Ferenc A Jolesz, Zang-Hee Cho, and Seung-Schik Yoo. Brain–machine interface via real-time fmri: preliminary study on thought-controlled robotic arm. *Neuroscience letters*, 450(1):1–6, 2009.

Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.

Bernardo Dal Seno, Matteo Matteucci, and Luca T Mainardi. The utility metric: a novel method to assess the overall performance of discrete brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(1):20–28, 2010.

David E Thompson, Lucia R Quitadamo, Luca Mainardi, Shangkai Gao, Pieter-Jan Kindermans, John D Simeral, Reza Fazel-Rezai, Matteo Matteucci, Tiago H Falk, Luigi Bianchi, et al. Performance measurement for brain–computer or brain–machine interfaces: A Tutorial. *Journal of neural engineering*, 11(3):035001, 2014.

David B Ryan, GE Frye, George Townsend, DR Berry, S Mesa-G, Nathan A Gates, and Eric W Sellers. Predictive spelling with a P300-based brain–computer interface: increasing the rate of communication. *Intl. Journal of Human-Computer Interaction*, 27(1):69–84, 2010.

DHR Blackwood and WJ Muir. Cognitive brain potentials and their application. *The British Journal of Psychiatry*, 157(S9):96–101, 1990.

Monica Fabiani, Gabriele Gratton, Demetrios Karis, and Emanuel Donchin. Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. *Advances in psychophysiology*, 2(S 1):78, 1987.

Costanza D'Avanzo, Sami Schiff, Piero Amodio, and Giovanni Sparacino. A bayesian method to estimate single-trial event-related potentials with application to the study of the P300 variability. *Journal of neuroscience methods*, 198(1):114–124, 2011.

SMM Martens, NJ Hill, J Farquhar, and B Schölkopf. Overlap and refractory effects in a brain–computer interface speller based on the visual P300 event-related potential. *Journal of neural engineering*, 6(2):026003, 2009.

Ana Carla Leite Romero, Simone Fiuza Regacone, Daiane Damaris Baptista de Lima, Pedro de Lemos Menezes, and Ana Cláudia Figueiredo Frizzo. Event-related potentials in clinical

research: guidelines for eliciting, recording, and quantifying Mismatch Negativity, P300, and N400. *Audiology-Communication Research*, 20(2):VII–VIII, 2015.

Charles A Nelson and Joseph P McCleery. Use of event-related potentials in the study of typical and atypical development. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(11):1252–1261, 2008.

Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Dennis J McFarland and Jonathan R Wolpaw. Brain-computer interfaces for communication and control. *Communications of the ACM*, 54(5):60–66, 2011.

David E Thompson, Md Rakibul Mowla, and Jane E Huggins. Evidence of latency variation in the P3 speller brain computer interface. In *Neuroscience Meeting Planner*, page 760.19/L29. Society for Neuroscience, 2019.

Eric W Sellers and Emanuel Donchin. A P300-based brain-computer interface: initial tests by ALS patients. *Clinical neurophysiology*, 117(3):538–548, 2006.

Maarten AS Boksem, Theo F Meijman, and Monicque M Lorist. Effects of mental fatigue on attention: an erp study. *Cognitive brain research*, 25(1):107–116, 2005.

Md Rakibul Mowla, Jane E Huggins, Balasubramaniam Natarajan, and David E Thompson. P300 latency estimation using least mean squares filter. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1976–1979. IEEE, 2018a.

John Polich and Kathryn L Herbst. P300 as a clinical assay: rationale, evaluation, and findings. *International Journal of Psychophysiology*, 38(1):3–19, 2000.

Norbert Schwarz. Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4): 433–440, 2000.

Joseph P Forgas. Mood and judgment: the affect infusion model (aim). *Psychological bulletin*, 117(1):39, 1995.

Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.

John T Cacioppo and Louis G Tassinary. Inferring psychological significance from physiological signals. *American psychologist*, 45(1):16, 1990.

Manuel Oliva and Andrey Anikin. Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific reports*, 8(1):4871, 2018.

Daniel S Quintana, Adam J Guastella, Tim Outhred, Ian B Hickie, and Andrew H Kemp. Heart rate variability is associated with emotion recognition: direct evidence for a relationship between the autonomic nervous system and social cognition. *International Journal of Psychophysiology*, 86(2):168–172, 2012.

Arturo Nakasone, Helmut Prendinger, and Mitsuru Ishizuka. Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th international workshop on biosignal interpretation*, pages 219–222. Citeseer, 2005.

Robert W Levenson, Paul Ekman, Karl Heider, and Wallace V Friesen. Emotion and autonomic nervous system activity in the minangkabau of west sumatra. *Journal of personality and social psychology*, 62(6):972, 1992.

Valery Petrushin. Emotion in speech: Recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering*, volume 710, 1999.

Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12): 1424–1445, 2000.

Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

Tim Dalgleish. The emotional brain. *Nature Reviews Neuroscience*, 5(7):583, 2004.

Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1(2):66–84, 2014.

Christian Mühl, Dirk Heylen, and Anton Nijholt. Affective brain-computer interfaces: neuroscientific approaches to affect detection. In *Oxford Handbook of Affective Computing*, pages 217–232. Oxford University Press Oxford, 2015.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report a-8, The Center for Research in Psychophysiology, University of Florida, FL, USA, 2008.

Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25 (1):49–59, 1994.

Md Rakibul Mowla, Jane E Huggins, and David E Thompson. Enhancing P300-BCI performance using latency estimation. *Brain-Computer Interfaces*, pages 1–9, 2017.

G Townsend, BK LaPallo, CB Boulay, DJ Krusienski, GE Frye, CKea Hauser, NE Schwartz, TM Vaughan, JR Wolpaw, and EW Sellers. A novel P300-based brain–computer interface

stimulus presentation paradigm: moving beyond rows and columns. *Clinical Neurophysiology*, 121(7):1109–1120, 2010.

Pieter-Jan Kindermans, Martijn Schreuder, Benjamin Schrauwen, Klaus-Robert Müller, and Michael Tangermann. True zero-training brain-computer interfacing-an online study. *PloS one*, 9(7):e102504, 2014a.

Pieter-Jan Kindermans, Michael Tangermann, Klaus-Robert Müller, and Benjamin Schrauwen. Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller. *Journal of neural engineering*, 11(3):035005, 2014b.

Dean J Krusienski, Eric W Sellers, François Cabestaing, Sabri Bayoudh, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. A comparison of classification techniques for the P300 speller. *Journal of neural engineering*, 3(4):299, 2006.

Emanuel Donchin, Kevin M Spencer, and Ranjith Wijesinghe. The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE transactions on rehabilitation engineering*, 8(2):174–179, 2000.

Dean J Krusienski, Eric W Sellers, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. Toward enhanced P300 speller performance. *Journal of neuroscience methods*, 167(1):15–21, 2008.

Matthias Kaper, Peter Meinicke, Ulf Grossekathoefer, Thomas Lingner, and Helge Ritter. BCI competition 2003-data set IIb: support vector machines for the P300 speller paradigm. *IEEE Transactions on Biomedical Engineering*, 51(6):1073–1076, 2004.

Alain Rakotomamonjy and Vincent Guigue. BCI competition III: dataset II-ensemble of svms for BCI P300 speller. *IEEE transactions on biomedical engineering*, 55(3):1147–1154, 2008.

Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An efficient P300-based brain-computer interface for disabled subjects. *Journal of Neuroscience methods*, 167(1):115–125, 2008.

Hubert Cecotti and Axel Graser. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):433–445, 2011.

Gregory McCarthy and Emanuel Donchin. A metric for thought: a comparison of P300 latency and reaction time. *Science*, 211(4477):77–80, 1981.

Marta Kutas, Gregory McCarthy, and Emanuel Donchin. Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time. *Science*, 197(4305):792–795, 1977.

Anthony Magliero, Theodore R Bashore, Michael GH Coles, and Emanuel Donchin. On the dependence of P300 latency on stimulus evaluation processes. *Psychophysiology*, 21(2): 171–186, 1984.

Terence W Picton. The P300 wave of the human event-related potential. *Journal of clinical neurophysiology*, 9(4):456–479, 1992.

John Polich. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

Anders M Fjell, Hannah Rosquist, and Kristine B Walhovd. Instability in the latency of P3a/P3b brain potentials and cognitive function in aging. *Neurobiology of aging*, 30(12): 2065–2079, 2009.

David E Thompson, Seth Warschausky, and Jane E Huggins. Classifier-based latency estimation: a novel way to estimate and predict BCI accuracy. *Journal of neural engineering*, 10(1):016006, 2012.

Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp components—A Tutorial. *NeuroImage*, 56(2):814–825, 2011.

Ruijiang Li, Andreas Keil, and Jose C Principe. Single-trial P300 estimation with a spatiotemporal filtering method. *Journal of neuroscience methods*, 177(2):488–496, 2009.

Kristine B Walhovd, Hannah Rosquist, and Anders M Fjell. P300 amplitude age reductions are not caused by latency jitter. *Psychophysiology*, 45(4):545–553, 2008.

Iñaki Iturrate, Ricardo Chavarriaga, Luis Montesano, Javier Minguez, and JdR Millán. Latency correction of event-related potentials between different experimental protocols. *Journal of neural engineering*, 11(3):036005, 2014.

David E Thompson, John J Baker, William A Sarnacki, and Jane E Huggins. Plug-and-play brain-computer interface keyboard performance. In *Neural Engineering, 2009. NER'09. 4th International IEEE/EMBS Conference on*, pages 433–435. IEEE, 2009.

Adrien Combaz, Nikolay Chumerin, Nikolay V Manyakov, Arne Robben, Johan AK Suykens, and Marc M Van Hulle. Towards the detection of error-related potentials and its integration in the context of a p300 speller brain–computer interface. *Neurocomputing*, 80:73–82, 2012.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. ISBN 9780262018029.

Liu Chun-Lin. A Tutorial of the wavelet transform. *NTUEE, Taiwan*, 2010.

David E Thompson, Stefanie Blain-Moraes, and Jane E Huggins. Performance assessment in brain-computer interface-based augmentative and alternative communication. *Biomedical engineering online*, 12(1):1, 2013.

Md Rakibul Mowla, Jesus D. Gonzalez-Morales, Jacob Rico-Martinez, Daniel A. Ulichnie, and David E Thompson. A comparison of classification techniques to predict brain-computer interfaces accuracy using classifier-based latency estimation. *Manuscript submitted for publication*, 2020a.

Jerry J Shih, Dean J Krusienski, and Jonathan R Wolpaw. Brain-computer interfaces in medicine. In *Mayo Clinic Proceedings*, volume 87, pages 268–279. Elsevier, 2012.

Luigi Bianchi, Chiara Liti, and Veronica Piccialli. A new early stopping method for P300 spellers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(8): 1635–1643, 2019.

Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Krausz, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How many people are able to control a P300-based brain–computer interface (BCI)? *Neuroscience letters*, 462(1):94–98, 2009.

John Polich and Albert Kok. Cognitive and biological determinants of P300: an integrative review. *Biological psychology*, 41(2):103–146, 1995.

Yasuo Yagi, Kerry L Coburn, Kristi M Estes, and James E Arruda. Effects of aerobic exercise and gender on visual and auditory P300, reaction time, and accuracy. *European journal of applied physiology and occupational physiology*, 80(5):402–408, 1999.

P Aricò, F Aloise, F Schettini, S Salinari, D Mattia, and F Cincotti. Influence of P300 latency jitter on event related potential-based Brain-Computer Interface performance. *Journal of neural engineering*, 11(3):035008, 2014.

Jieping Ye. Least Squares Linear Discriminant Analysis. In *Proceedings of the 24th international conference on Machine learning*, pages 1087–1093. ACM, 2007.

Kibok Lee and Junmo Kim. On the equivalence of linear discriminant analysis and least squares. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Lukáš Vařeka and Pavel Mautner. Stacked autoencoders for the P300 component detection. *Frontiers in neuroscience*, 11:302, 2017.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Jonathan R Wolpaw, Herbert Ramoser, Dennis J McFarland, and Gert Pfurtscheller. EEG-based communication: improved accuracy by response verification. *IEEE transactions on Rehabilitation Engineering*, 6(3):326–333, 1998.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Yosef Hochberg and Ajit C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, 1987.

Leonard A Marascuilo and Maryellen McSweeney. Nonparametric post hoc comparisons for trend. *Psychological Bulletin*, 67(6):401, 1967.

Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference*. Springer, 2011.

Paul H Kvam and Brani Vidakovic. *Nonparametric statistics with applications to science and engineering*, volume 653. John Wiley & Sons, 2007.

Alessio Benavoli, Giorgio Corani, and Francesca Mangili. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, 2016.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80–83, 1945.

Md Rakibul Mowla, Rachael I. Cano, Katie J. Dhuyvetter, and David E. Thompson. Affective brain-computer interfaces: A tutorial to choose performance measuring metric, 2020b.

Rosalind Wright Picard et al. Affective computing. 1995.

Daniel Goleman. Emotional Intelligence. Why It Can Matter More than IQ. *Learning*, 24 (6):49–50, 1996.

Fritz Strack, Norbert Schwarz, Brigitte Chassein, Dieter Kern, and Dirk Wagner. Salience of comparison standards and the activation of social norms: Consequences for judgements of happiness and their communication. *British Journal of Social Psychology*, 29(4):303–314, 1990.

Christopher P Niemic and K Warren. Studies of emotion: A theoretical and empirical review of psychophysiological studies of emotion. *Journal of Undergraduate Research Rochester*, 1(1):15–19, 2002.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

Peter J Lang and Margaret M Bradley. International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings. Technical report b-2, The Center for Research in Psychophysiology, University of Florida, FL, USA, 1999.

Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

Bahar Güntekin and Erol Başar. A review of brain oscillations in perception of faces and emotional pictures. *Neuropsychologia*, 58:33–51, 2014.

Kalyani P Wagh and K Vasanth. Electroencephalograph (EEG) based emotion recognition system: A review. In *Innovations in Electronics and Communication Engineering*, pages 37–59. Springer, 2019.

Beatriz García-Martínez, Arturo Martinez-Rodrigo, Raul Alcaraz, and Antonio Fernández-Caballero. A review on nonlinear methods using electroencephalographic recordings for emotion recognition. *IEEE Transactions on Affective Computing*, 2019.

Shangfei Wang, Yachen Zhu, Guobing Wu, and Qiang Ji. Hybrid video emotional tagging using users' EEG and video content. *Multimedia tools and applications*, 72(2):1257–1283, 2014.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011.

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

Xiang Li, Dawei Song, Peng Zhang, Yazhou Zhang, Yuexian Hou, and Bin Hu. Exploring EEG features in cross-subject emotion recognition. *Frontiers in neuroscience*, 12:162, 2018.

Mohammad Soleymani, Frank Villaro-Dixon, Thierry Pun, and Guillaume Chanel. Toolbox for emotional feature extraction from physiological signals (TEAP). *Frontiers in ICT*, 4: 1, 2017.

Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 2017.

Shangfei Wang, Shiyu Chen, and Qiang Ji. Content-based video emotion tagging augmented by users' multiple physiological responses. *IEEE Transactions on Affective Computing*, 10 (2):155–166, 2017.

Mehmet Siraç Özerdem and Hasan Polat. Emotion recognition based on EEG features in movie clips with channel selection. *Brain informatics*, 4(4):241, 2017.

Gyanendra K Verma and Uma Shanker Tiwary. Affect representation and recognition in 3D continuous valence-arousal-dominance space. *Multimedia Tools and Applications*, 76(2): 2159–2183, 2017.

Youjun Li, Jiajin Huang, Haiyan Zhou, and Ning Zhong. Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. *Applied Sciences*, 7(10):1060, 2017.

Beatriz García-Martínez, Arturo Martínez-Rodrigo, Roberto Zangróniz Cantabrana, Jose Pastor García, and Raúl Alcaraz. Application of entropy-based metrics to identify emotional distress from electroencephalographic recordings. *Entropy*, 18(6):221, 2016.

Obada Al Zoubi, Mariette Awad, and Nikola K Kasabov. Anytime multipurpose emotion recognition from EEG data using a liquid state machine based framework. *Artificial intelligence in medicine*, 86:1–8, 2018.

Prima Purnamasari, Anak Ratna, and Benyamin Kusumoputro. Development of filtered bispectrum for EEG signal feature extraction in automatic emotion recognition using artificial neural networks. *Algorithms*, 10(2):63, 2017.

Jingxin Liu, Hongying Meng, Maozhen Li, Fan Zhang, Rui Qin, and Asoke K Nandi. Emotion detection from EEG recordings based on supervised and unsupervised dimension reduction. *Concurrency and Computation: Practice and Experience*, 30(23):e4446, 2018.

Cristian Torres-Valencia, Mauricio Álvarez-López, and Álvaro Orozco-Gutiérrez. Svm-based feature selection methods for emotion recognition from multimodal data. *Journal on Multimodal User Interfaces*, 11(1):9–23, 2017.

Maria Luiza Recena Menezes, A Samara, Leo Galway, A Sant'Anna, Antanas Verikas, Fernando Alonso-Fernandez, H Wang, and R Bond. Towards emotion recognition for virtual

environments: an evaluation of EEG features on benchmark dataset. *Personal and Ubiquitous Computing*, 21(6):1003–1013, 2017.

Ahmet Mert and Aydin Akan. Emotion recognition from EEG signals by using multivariate empirical mode decomposition. *Pattern Analysis and Applications*, 21(1):81–89, 2018.

Bahareh Nakisa, Mohammad Naim Rastgoo, Dian Tjondronegoro, and Vinod Chandran. Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. *Expert Systems with Applications*, 93:143–155, 2018.

D Jude Hemanth, J Anitha, et al. Brain signal based human emotion analysis by circular back propagation and Deep Kohonen Neural Networks. *Computers & Electrical Engineering*, 68:170–180, 2018.

Zhong Yin, Yongxiong Wang, Li Liu, Wei Zhang, and Jianhua Zhang. Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination. *Frontiers in neurorobotics*, 11:19, 2017a.

Hyun Joong Yoon and Seong Youb Chung. EEG-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm. *Computers in biology and medicine*, 43(12):2230–2237, 2013.

Shangfei Wang, Yachen Zhu, Lihua Yue, and Qiang Ji. Emotion recognition with the help of privileged information. *IEEE Transactions on Autonomous Mental Development*, 7(3): 189–200, 2015.

Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 2014.

Syed Naser Daimi and Goutam Saha. Classification of emotions induced by music videos and correlation with participants' rating. *Expert Systems with Applications*, 41(13):6057–6065, 2014.

Jorge I Padilla-Buritica, Juan D Martinez-Vargas, and German Castellanos-Dominguez. Emotion discrimination using spatially compact regions of interest extracted from imaging EEG activity. *Frontiers in computational neuroscience*, 10:55, 2016.

Rishabh Gupta, Tiago H Falk, et al. Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization. *Neurocomputing*, 174:875–884, 2016.

Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, 140:93–110, 2017b.

Andrea Clerico, Abhishek Tiwari, Rishabh Gupta, Srinivasan Jayaraman, and Tiago H Falk. Electroencephalography Amplitude Modulation Analysis for Automated Affective Tagging of Music Video Clips. *Frontiers in computational neuroscience*, 11:115, 2018.

Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17 (2):145–151, 2008.

Vincent Van Asch. Macro-and micro-averaged evaluation measures [[basic draft]]. *Belgium: CLiPS*, pages 1–27, 2013.

Jamie R Velo, Jennifer L Stewart, Brant P Hasler, David N Towers, and John JB Allen. Should it matter when we record? Time of year and time of day as factors influencing frontal EEG asymmetry. *Biological psychology*, 91(2):283–291, 2012.

John JB Allen and Samantha J Reznik. Frontal EEG asymmetry as a promising marker of depression vulnerability: Summary and methodological considerations. *Current opinion in psychology*, 4:93–97, 2015.

Jürgen Kayser and Craig E Tenke. Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: I. evaluation with auditory oddball tasks. *Clinical neurophysiology*, 117(2):348–368, 2006.

François Perrin, J Pernier, O Bertrand, and JF Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2): 184–187, 1989.

Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

James A Coan and John JB Allen. Frontal EEG asymmetry as a moderator and mediator of emotion. *Biological psychology*, 67(1-2):7–50, 2004.

John JB Allen, James A Coan, and Maria Nazarian. Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biological psychology*, 67(1-2):183–218, 2004.

Nikita van der Vinne, Madelon A Vollebregt, Michel JAM van Putten, and Martijn Arns. Frontal alpha asymmetry as a diagnostic marker in depression: Fact or fiction? A meta-analysis. *Neuroimage: clinical*, 16:79–87, 2017.

Peter Putman, Jacobien van Peer, Ioulia Maimari, and Steven van der Werff. EEG theta/beta ratio in relation to fear-modulated response-inhibition, attentional control, and affective traits. *Biological psychology*, 83(2):73–78, 2010.

Bo Hjorth. EEG analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970.

Anne Vakkuri, A Yli-Hankala, P Talja, S Mustola, H Tolvanen-Laakso, Tl Sampson, and H Viertiö-Oja. Time-frequency balanced spectral entropy as a measure of anesthetic drug effect in central nervous system during sevoflurane, propofol, and thiopental anesthesia. *Acta Anaesthesiologica Scandinavica*, 48(2):145–153, 2004.

Zeynab Mohammadi, Javad Frounchi, and Mahmood Amiri. Wavelet-based emotion recognition system using EEG signal. *Neural Computing and Applications*, 28(8):1985–1990, 2017.

Nicolas Lachiche and Peter A Flach. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 416–423, 2003.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

Digna R Velez, Bill C White, Alison A Motsinger, William S Bush, Marylyn D Ritchie, Scott M Williams, and Jason H Moore. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic epidemiology*, 31(4):306–315, 2007.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *Pattern recognition (ICPR), 2010 20th international conference on*, pages 3121–3124. IEEE, 2010.

Henry Carrillo, Kay H Brodersen, and José A Castellanos. Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy. In *ROBOT2013: First Iberian Robotics Conference*, pages 347–361. Springer, 2014.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 9780408709293.

Nancy Chinchor. MUC-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding*, pages 22–29. Association for Computational Linguistics, 1992.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

David MW Powers. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. *arXiv preprint arXiv:1503.06410*, 2015.

John Atkinson and Daniel Campos. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Systems with Applications*, 47:35–41, 2016.

Anatole Lécuyer, Fabien Lotte, Richard B Reilly, Robert Leeb, Michitaka Hirose, and Mel Slater. Brain-computer interfaces, virtual reality, and videogames. *Computer*, 41(10): 66–72, 2008.

Jan Van Erp, Fabien Lotte, and Michael Tangermann. Brain-computer interfaces: beyond medical applications. *Computer*, 45(4):26–34, 2012.

Md Rakibul Mowla, Jane E Huggins, and David E Thompson. Boosting BCI accuracy using wavelet enhanced cble scores as a classifier feature. In *Proceedings of the 6th International Brain-Computer Interface Meeting*. BCI Society, 2016. doi: 10.3217/978-3-85125-467-9-5.

Md Rakibul Mowla, Jane E Huggins, and David E Thompson. *Evaluation and Performance Assessment of the Brain–Computer Interface System*, chapter 33, pages 634–649. CRC Press, 2018b. doi: 10.1201/9781351231954-33.

James A Coan and John JB Allen. *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.

Karl Friston. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7(2):e1000033, 2009.

Appendix A

Reuse Permissions from Publishers

Academic UK Non Rightslink <permissionrequest@tandf.co.uk>
Fri 2/14/2020 03:50 AM
To: Md Rakibul Mowla <rakibulmowla@ksu.edu>

Dear Md Rakibul Mowla,

Material Requested: 'Md Rakibul Mowla, Jane E. Huggins & David E. Thompson (2017) Enhancing P300-BCI performance using latency estimation, Brain-Computer Interfaces, 4:3, 137-145, DOI: 10.1080/2326263X.2017.1338010'.

Thank you for your correspondence requesting permission to reproduce your **Authors Accepted Manuscript** from our Journal in your printed thesis and to be posted in the university's repository – **Kansas State University.**

We will be pleased to grant permission on the sole condition that you acknowledge the original source of publication and insert a reference to the article on the Journals website: http://www.tandfonline.com

This is the authors accepted manuscript of an article published as the version of record in **Brain-Computer Interfaces** © Taylor & Francis 2017, Informa UK Limited, trading as Taylor & Francis Group, https://doi.org/10.1080/2326263X.2017.1338010.

This permission does not cover any third party copyrighted work which may appear in the material requested.

Please note that this license does not allow you to post our content on any third party websites or repositories.

This licence does not allow the use of the Publishers version/PDF (this is the version of record that is published on the publisher's website) to be posted online.

Thank you for your interest in our Journal.

Yours sincerely,

Annabel
Annabel Flude – Permissions Administrator, Journals
Taylor & Francis Group
3 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN, UK.
Tel: +44 (0)20 7017 7617
Fax: +44 (0)20 7017 6336
Web: www.tandfonline.com
e-mail: annabel.flude@tandf.co.uk

Material Requested: Mowla, Md Rakibul, Jane E. Huggins, and David E. Thompson. "Evaluation and Performance Assessment of the Brain–Computer Interface System." In Brain–Computer Interfaces Handbook, pp. 635-650. CRC Press, 2018.

Order Details

1. Brain–Computer Interfaces Handbook : Technological and Theoretical Advances

Billing Status: Open

Order license ID	1018953-1
Order detail status	Completed
ISBN-13	9780367375454
Type of use	Republish in a thesis/dissertation
Publisher	Taylor and Francis Group LLC (Books) US
Portion	Chapter/article

0.00 USD

LICENSED CONTENT

Publication Title	Brain–Computer Interfaces Handbook : Technological and Theoretical Advances	**Rightsholder**	Taylor & Francis Group LLC - Books
		Publication Type	Book
Date	08/18/2019		
Language	English		

REQUEST DETAILS

Portion Type	Chapter/article	**Rights Requested**	Main product
Page range(s)	635-649	**Distribution**	Worldwide
Total number of pages	16	**Translation**	Original language of publication
Format (select all that apply)	Electronic	**Copies for the disabled?**	No
Who will republish the content?	Academic institution		
Duration of Use	Life of current edition	**Minor editing privileges?**	No
Lifetime Unit Quantity	Up to 499	**Incidental promotional use?**	No
		Currency	USD

NEW WORK DETAILS

Title	Applications of non-invasive brain-computer interfaces for communication and affects recognition	**Institution name**	Kansas State University
		Expected presentation date	2020-03-23
Instructor name	Md Rakibul Mowla		

ADDITIONAL DETAILS

The requesting person / organization to appear on the license	Md Rakibul Mowla

REUSE CONTENT DETAILS

Title, description or numeric reference of the portion(s)	Brain–Computer Interfaces Handbook : Technological and Theoretical Advances	**Title of the article/chapter the portion is from**	Evaluation and Performance Assessment of the Brain- Computer Interface System
Editor of portion(s)	Chang S. Nam, Anton Nijholt, Fabien Lotte	**Author of portion(s)**	Md Rakibul Mowla, Jane E. Huggins, and David E. Thompson
Volume of serial or monograph	1		
Page or page range of portion	635-649	**Publication date of portion**	2019-08-18