# A QUALITY METRIC TO IMPROVE WRAPPER FEATURE SELECTION IN MULTICLASS SUBJECT INVARIANT BRAIN COMPUTER INTERFACES

A DISSERTATION IN

Electrical and Computer Engineering
and
Telecommunications and Computer Networking

Presented to the Faculty of the University
Of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
JESSE SHERWOOD

BSEE, University of Missouri-Rolla, 1978
MSECE, University of Missouri-Columbia, 1985
MBA, Rockhurst College, 1988

Kansas City, Missouri
2011

# A QUALITY METRIC TO IMPROVE WRAPPER FEATURE SELECTION IN MULTICLASS SUBJECT INVARIANT BRAIN COMPUTER INTERFACES

Jesse Sherwood, Candidate for the Doctor of Philosophy Degree

University of Missouri-Kansas City, 2011

## ABSTRACT

Brain computer interface systems based on electroencephalograph (EEG) signals have limitations which challenge their application as a practical device for general use. The signal features generated by the brain states we wish to detect possess a high degree of inter-subject and intra-subject variation. Additionally, these features usually exhibit a low variation across each of the target states. Collection of EEG signals using low resolution, non-invasive scalp electrodes further degrades the spatial resolution of these signals. The majority of brain computer interface systems to date require extensive training prior to use by each individual user. The discovery of subject invariant features could reduce or even eliminate individual training requirements. To obtain suitable subject invariant features requires search through a high dimension feature space consisting of combinations of spatial, spectral and temporal features. Poorly separable features can prevent the search from converging to a usable solution as a result of degenerate classifiers. In such instances the system must detect and compensate for degenerate classifier behavior. This dissertation presents a method to accomplish this search using a wrapper architecture comprised of a sequential forward floating search algorithm coupled with a support vector machine classifier. This is successfully achieved by the introduction of a scalar Quality (Q)-factor metric, calculated from the ratio of sensitivity to specificity of the confusion matrix. This method is successfully applied to a multiclass subject independent BCI using 10 untrained subjects performing 4 motor tasks.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Graduate Studies have examined a dissertation titled “A Quality Metric to Improve Wrapper Feature Selection in Multiclass Subject Invariant Brain Computer Interfaces,” presented by Jesse Sherwood, candidate for the Doctor of Philosophy degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Reza Derakhshani, Ph.D., Committee Chair
Department of Computer Science and Electrical Engineering

Cory Beard, Ph.D.
Department of Computer Science and Electrical Engineering

Ghulam Chaudhry, Ph.D.
Department of Computer Science and Electrical Engineering

W. Daniel Leon-Salas, Ph.D.
Department of Computer Science and Electrical Engineering

Kenneth Mitchell, Ph.D.
Department of Computer Science and Electrical Engineering

CONTENTS

Chapter

Appendix

# LIST OF ILLUSTRATIONS

Figure Page

# LIST OF TABLES

Table Page

# ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude and love toward my family without whose endless support, patience and sacrifice, I would have never been able to accomplish this.

I would like to express my gratitude and appreciation toward my advisor and committee chairman Dr. Reza Derakhshani for his encouragement, leadership and guidance. His tireless efforts and advice were essential toward my obtaining this personal and professional achievement.

Additionally, I would like to thank my doctoral committee members Dr. Ghulam M. Chaudhry, Dr. Cory Beard, Dr. Kenneth Mitchell, Dr. W.D. Leon-Salas and Dr. Khosrow Sohraby for their advice and support through out this process.

In general, I also acknowledge the inspiration and contributions of the many teachers, educational administrators and scientific leaders, now living and many who have long since departed, each who has participated in my educational and professional journey.

# CHAPTER 1

# INTRODUCTION TO BRAIN COMPUTER INTERFACE SYSTEMS

## 1.1 Motivation

The brain computer interface (BCI) establishes a communications pathway for translating imagined human movements into computer commands for the purpose of controlling external devices, such as a computer or robotic device. BCI devices permit the human control of external devices through the use of thought patterns. Possible applications include communications and restoration of motor functions for disabled persons, remote operation of equipment in hazardous environments and gaming or virtual reality interaction. An additional aspect of BCI research is providing scientific insight into how the brain processes inputs and uses the information to control human actions.

While the implementations of BCI are numerous and varied, including invasive methods which require implants to non-invasive methods which employ externally worn devices, much remains unsolved in the BCI realm. The preponderance of BCI implementations are subject specific and require extensive training to learn the signatures for each of the brain states for each individual. In view of this, the possibility of a subject invariant brain computer interface (SIBCI) offers the advantage of reducing the overall training requirements for each individual and of providing additional insights into the human brain functionality for controlling movements. The SIBCI problem is complex as features that have been derived from electroencephalographic (EEG) signatures for brain states resulting from the planning and execution of movements have been shown to have both intersubject variance and intrasubject variance [107]. When multiple tasks or classes are included and the intraclass variation is small the probability of successful detection also decreases. The complexity of the problem is magnified by the requirement to produce a portable, non-invasive and general purpose BCI system. This proposition relies upon obtaining the best subsets of EEG features from a large population of candidate features. The techniques employed here include wrapper feature search, with a sequential forward floating search (SFFS) algorithm and support vector machine (SVM) classifiers. The performance of the final feature subset is dependent upon the quality of the feedback provided from the SVM, which is used to guide the SFFS search algorithm. SVM and discriminant classifiers may produce trivial or degenerate outputs under conditions, to be described later in this work, that are frequently encountered in multiclass SIBCI.

In this dissertation, the focus is on the derivation and implementation of a method to search a large group of features and improve the quality of feedback to guide the search. To explain this concept, a set theory interpretation of subject-invariant brain state feature space is presented. The impact of subject invariance and number of classes on the performance of SVM classifiers is analyzed and the resulting information is used to create an adaptive and dynamic feedback metric or Quality (Q-factor) factor, which detects and compensates for classifier degeneracies. The application of the Q-factor guides the search algorithm to produce an increased number of searches that converge to valid feature subset solutions. The goal is not necessarily to produce a lower error result but rather a less biased, higher validity result. It will be shown that as the number of brain states increases, the invalid or trivial solutions tend to dominate the solutions. While this dissertation addresses the SIBCI problem, this method can be applied to a much broader range of multiclass machine learning applications.

## 1.2 Overview

EEG signals corresponding to quasi-movement brain states are collected from a 10 untrained subjects through non-invasive scalp based electrodes. Each subject participates in two recording sessions scheduled 4-6 weeks apart, to exploit the intra-subject variability of EEG signals. Temporal, spectral and spatial features are extracted. From these extracted features, subsets of features are chosen, which demonstrate the best performance in training the system. The feature subset selection is based on a wrapper or data-driven approach. In the wrapper approach, the feature state space search is guided by the evaluation of the performance of a classifier. This produces a feature subset that accounts for the biases of a particular type of classifier, in our case the SVM.

There is an unfortunate drawback to the wrapper approach when the search algorithm encounters feature subsets that are highly uncorrelated with target brain states. When these sets are encountered, the classifier has a tendency to settle into one of two trivial or degenerate states where samples are classified as all belonging to either one class or another. The detrimental effects of these degeneracies are exacerbated by several other conditions that are characteristic of multiclass SIBCI applications. In multiclass arrangements where we try to detect more than two brain states using a one versus all the rest (OVR) arrangement of the input samples, the misleading result or bias worsens as the number of brain states being detected increases. Other degrading influences include the complex nature of the multiclass subject invariant feature space, the low signal to noise ratio of EEG environment, non-linear low pass filtering of the cranium, the low resolution of the 19 electrode scalp based EEG electrodes and the inadequacy of typical performance evaluation metrics. We can mitigate the effects of these factors by the proper choice of an evaluation function or performance metric which evaluates and compensates for the amount of degeneracy in the classifier.

In this dissertation we provide background on the state of the art of subject specific and subject invariant BCI including the evolution and development of EEG as a BCI tool. This is followed by a characterization of the feature space and signal environment encountered in BCI with emphasis on those attributes that compound the multiclass SIBCI problem. A description of the methods employed in feature selection and classification is provided with an analysis of their behavior pertinent to the multiclass SIBCI. The trivial or degenerate classifiers modes will be analyzed with emphasis placed on the derivation of a metric that can be used to detect such degenerate states. The acceptance criteria for a usable metric will be presented. The metric will be implemented and the results compared with uncorrected systems.

## 1.3 Novel Research Contributions

The application of the Quality factor in the detection of a subject invariant set of features corresponding to imagined quasi-motor tasks is the major novel contribution of this research. Sub-components of this contribution include:

- Derivation of a scalar metric reflecting the classification favorable qualities of the confusion (Quality factor) matrix. Feature selection performed by a wrapper method requires a scalar value that accounts for the accuracy of the classifier yet also considers the degeneracy of the classifier. Deriving this factor from the confusion matrix provides a relationship to sensitivity and specificity, which are both indicators of the quality or degeneracy of the classifier output. Building a metric based on sensitivity and specificity is a logical choice for constructing a quality metric. Our method detects and compensates for two distinct degenerate (or trivial) classifier modes.

- The implementation of the Q-factor, a specifically derived cost or fitness function, within a wrapper method to eliminate the impact of trivial classifier states on the search or induction algorithm. Preceding BCI performance metrics have been devised as a basis for comparison between systems to support model selection decisions. The metrics are generally applied to the output stage of the classifier. While the technique described in this dissertation may also be used to measure overall output performance, it is primarily intended to shape the cost/fitness function and guide the feature selection search. As such, this metric is applied within the feedback path of the wrapper rather than at the output of the overall classifier model.
- Application of set theory to devise a framework for the interpretation of brain state feature space. Characterization of the multiclass SIBCI feature space in terms of set theory provides a logical and structured manner for gaining insight into the complexities of the classifier problem as the number of subjects and/or tasks increase. Set representation is suitable for reflecting the inter-subject, intra-subject and inter-task variations seen in brain state features. As a conceptual tool, the set theory model supports conclusions otherwise presented in information theory based or probabilistic approaches for determining the performance of multiclass SIBCI. Admittedly, this approach is abstract and while it does not yield

quantitative solutions we hold it to be a useful tool for providing insight into the underlying feature behavior. Set representation provides a visual explanation of the difficulty in obtaining separable features as the number of subjects and/or classes increases. It also supports the assertion that as the number of subjects and/or classes increases there may exist a growing space of unclassifiable features.

- Characterization of support vector machine classifier output conditions in an accuracy-imbalance domain. While the existence of trivial solutions in discrete output classifiers has been well established, little has been done to fully identify the impact of these trivial or degenerate modes or measure the degree of degeneracy. The development of a performance metric that compensates for degeneracy of a classifier requires the development of a framework that captures the degree of degeneracy, preferably one that is an extension of existing and readily available metrics. By characterizing these conditions we establish a basis from which we can construct a detecting or compensating function or metric.

# CHAPTER 2

# HISTORICAL PERSPECTIVE AND STATE OF THE ART

## 2.1 Brain Physiology

The primary source of electrical activity measured by the EEG is a thin layer on the outer surface of the brain known as the cerebral cortex. The cortex consists of nervous system cells or neurons. These neurons convey and process information to and from the brain. The neurons contain a cell body or soma and are connected by a large web of membranes or apical dendrites. Each dendrite has thousands of branches each of which conduct electrical impulses to and from other neurons. Each axon acts as an output to a neuron and transports the electric signals back to the nervous system.

The signal is transmitted in the form of an electrical impulse from the neuron along the axon where it is transformed into an electrochemical signal or neurotransmitter at the junction or synapse between the terminal end of the axon and the next neuron. The information is transported across the synapse in the form of an electro-chemical reaction and is reconverted to an electrical impulse at the next successive neuron.

Each of the billions of neurons is receiving, summing and thresholding electrical signals received from a number of synaptic junctions to commence the firing of an action potential at the next neuron. This firing depends on the amplitude, number and timing of received pulses and the chemical makeup of the neuronal cells. Signals can either excite or inhibit the firing of an action potential.

The waveforms of the signals transported on the axons are characterized by strings of waveforms with uniform amplitudes and different firing rates. A high pulse rate indicates a high degree of neural activity. Over a billion neurons can be firing simultaneously in the human brain [86].

Different regions of the cerebral cortex are responsible for processing information and controlling the various bodily functions. Regions are devoted to planning, emotion, language interpretation, motion, and sensory detection. The brain consists of two lateral, symmetric left and right hemispheres. Within each hemisphere are four distinct lobes (frontal, occipital, temporal and parietal). Specialized functions are performed within each of these lobes. Most of the tasks associated with planning, emotion and control of muscular activity are processed in the frontal lobe area and to some extent in the forward region of the parietal lobe [1].

We cannot determine the current state of an intended motion or thought from the activity of a single neuron. This is precluded both by the number of neurons and our inability to know in advance which neuron will be associated with a particular activity. The use of a non-invasive measurement technique imposes the attenuation and spatial dispersion transfer characteristics of the intervening layers of bone, skin and fluids on the electrical signals of interest. However we can observe synaptic activity over a large scale by measuring the aggregation of signals from larger numbers of adjacent neurons, assuming that a large number of neurons fire in synchronization. When the pulses fire in unison we observe well-defined signals, when the pulses fire at random the signals resembles noise.

## 2.2 Electroencephalography

The modern EEG has its earliest roots in the 1790's when Italian physician Luigi Galvani demonstrated the electrical properties of the nervous system in his experimentation with frog tissues. Following the lead of Galvani's experiments other researchers including Carlo Matteucci and Emil Du Bois-Reymond explored the property of human tissues to generate electrical signals. Du Bois-Reymond, a German physician, built upon the work of Metteucci and discovered the existence of animal electricity establishing the foundation for modern neurophysiology. Du Bois-Reymond recorded the actual passage of an electrical impulse along a nerve tissue in 1848 [84].

Encephalographic research and the study of neurophysiology flourished in Western Europe during the 19th century and early 20th century. English scientist Richard Caton published the first recordings of animal brain activity in 1875 [23]. Caton's work was based on the placement of the electrodes of a reflecting galvanometer directly on exposed cortical tissue. Neurophysiologists were investigating the localization of the brain functions, postulating on the ideas of hemispheric specialization and the existence of specialized cortical regions. In 1861, French physician Pierre Paul Broca and later Karl Werninke of Germany examined language and motor specialization in human patients [84]. Actual analysis of EEG with human patients started in 1924 with galvanometer measurements on persons with skull bone defects in post WWI Germany. The electrical theories of human brain activities progressed when Hans Berger published the earliest human brain activity results in 1929 using non-invasive scalp placed electrodes [80]. Berger is credited as the discoverer of the modern EEG. Berger received his doctorate in medicine in 1897 from the University of Jena, where he became a professor in 1906. In 1919, he was named director of the psychiatry and neurology clinic at Jena [123]. Beginning in 1920, Berger devised a method for scalp-based measurements, recognizing that it was desirable to collect measurements from patients that did not exhibit skull bone defects. Berger's one-channel recordings were captured on photographic paper and used frontal-occipital leads. The leads were silver foils attached to the scalp with bandages and measured using a Siemens double coil galvanometer. Berger was able to resolve measurements as small as 100 microvolts [84]. Berger was investigating the phenomenon of alpha blockage where the amplitude of signals at certain frequencies attenuate when a subject opens their eyes. This band of frequencies or "Berger's wave", are now identified as alpha waves. This was the first documentation of oscillatory electrical behavior in the human brain due to the presence of external sensory stimuli which we today call event-related potentials (ERP). Burger observed that alpha waves decrease while accompanied by an increase in beta activity, a predecessor to the concept of desynchronization used in present day EEG analysis. In 1932, signal processing first occurred in EEG when Berger and G. Dietsch used Fourier analysis to detect the change in brain waves due to anesthesia [123]. This event constitutes the beginning of quantitative EEG (QEEG) as opposed to the previous qualitative EEG analysis methods. Qualitative EEG describes a general method of visually analyzing the EEG waveform patterns and quantitative utilizes mathematical and statistical techniques to present the data. Fourier analysis is still used today in QEEG and brain-computer interfaces. QEEG opened the door to the detection of the small event-related potentials amidst the larger number and magnitudes of random signals encountered in the normal EEG.

Additional advances in techniques to detect the ERPs were initiated with the invention of George Dawson's electro-mechanical signal averaging device a few years later [84]. Dawson, a researcher at the University of London, hypothesized that the ERPs occur at a fixed time after a stimulus occurs and have a specifically identifiable waveform. By averaging the wave signatures captured by multiple stimuli, it is possible to average out the random background EEG activity and clearly observe a waveform corresponding to the original stimuli.

A number of parallel research activities brought about technical improvements in the accuracy of EEG measurements in the period shortly after Berger's initial findings. A group at the Institute of Brain Research at Berlin-Bush was active at that same time as Berger. J.F. Toennes and A.E. Kornmüller, both researchers within this group developed the ink writing EEG, and electrical differential amplifiers for EEG, not unlike what we use today in BCI. Kornmüller and Oskar Vogt collected EEG samples from a greater number of locations on the brain and focused on the specific behaviors of EEG signatures collected from the different hemispheric and cortical regions [84].

During this period, there were also parallel developments and research activities in England, France, Belgium and Italy. There was little if any research in the United States at this time. During the period from the 1940s and 1950s clinical EEG focused around epilepsy and sleep disorder [84]. After 1960 the direction was toward the use of computers to provide faster, more accurate and automated diagnosis. These developments include the introduction of the Fast Fourier transform by Cooley and Tukey in 1965, and dipole source modeling and topographical mapping in the 1980s [26].

Going back to QEEG, which started with Berger and Deitsch's discovery of alpha blocking in 1929. Through using mathematical analysis alpha blocking was determined to be a quantifiable phenomenon. This behavior was renamed desynchronization to reflect this continuous change in the rhythm magnitude. Pfurtscheller has reported that alpha rhythms can become desynchronized, either entirely or partially, due to anticipation, reception or other processing of sensory information [95]. Rhythmic patterns in the human EEG are purported to originate from synchronization of neurons caused by inhibitory processes within the thalamacortical system [5,121] as mentioned earlier and from feedback generated by inhibitory and excitatory neurons [42]. These are periodic signals with regular features that lend themselves to quantitative analysis. In light of this, EEG signals may be described as a mixture of random, periodic, linear, non-linear, sinusoidal and non-sinusoidal signals characterized which also exhibit properties of intersubject variation and non-stationarity. We can extract spatial, temporal and spectral features from EEG signals to gain knowledge of the determining sources of the signals. Investigations into regional or spatial patterns of desynchronization may indicate the performance of different affective, cognitive or motor states.

Until the 1990's, EEG was mainly used in clinical settings as a diagnostic tool for neurological disorders. However, as the exploration of ways to use brain signals for controlling computers progressed the EEG became a logical candidate for a BCI engine. Other methods for monitoring brain function have appeared recently, such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG). EEG is attractive because of its relatively low-cost, portability and high temporal resolution.

## 2.3 Brain Computer Interface

The first use of brain signals as a method to command a computer without any physical contact with the computer was reported by Jacques Vidal in 1973, creator of the term “Brain Computer Interface” under the auspices of the University of California, Los Angeles [84,107,131]. Over the following two decades BCI activity grew along with improvements in the capture and processing of digital EEG signals and improvements in computer technology.

We can categorize EEG based BCI into the dependent and independent systems. Dependent systems rely on some degree of muscular control such as focusing on a particular highlighted square in a matrix. An image is formed on the visual cortex from the highlighted square but the system also detects movement of ocular muscles to permit focusing on the desired square. Independent systems rely strictly on non-muscular activity such as imagined motor or cognitive tasks.

Several organizations have been at the forefront in BCI research in the past two decades and their contributions are acknowledged. An overview is provided of the work of each of the major BCI labs.

The Berlin BCI group (http://www.bbci.de/) focuses on new sensor technology, improved understanding of the brain and the analysis of brain waves using modern machine learning methods. Berlin BCI consists of a cooperation of research agencies in Germany. The membership includes the Berlin Institute of Technology, Machine Learning Laboratory, led by Prof. Dr. Klaus-Robert Müller and Dr. Benjamin Blankertz, Fraunhofer FIRST institute, the IDA (Intelligent Data Analysis) research group, led by Dr. Benjamin Blankertz, and the Department of Neurology at Campus Ben Franklin, Charité - University Medicine, Berlin, led by Prof. Dr. Gabriel Curio. One of the earlier significant contributions from the Berlin group was the reduction of machine training times from weeks and months to less than 30 minutes [11,12,13,14,15].

Wadsworth BCI research, (http://www.wadsworth.org/) a part of the New York State department of health, is based out of Albany, NY. Dr. Jonathan Wolpaw leads Wadsworth BCI along with his colleague Dr. Gerwin Schalk. The primary focus of Wadsworth or Albany BCI activities is the development of BCI technology to restore communication and control to people who are severely paralyzed by amyotrophic lateral sclerosis (ALS), strokes, or other devastating neuromuscular disorders [113,129,136]. Wadsworth has developed a BCI system that lets a user direct the movement of a cursor on a computer screen. The BCI system detects sensorimotor rhythm (SMR) amplitudes in the $\mu$ (8-12 Hz) and β (18-16 Hz) frequency bands. The system relies on user practice to achieve the correct modulation of the SMR and used a regression algorithm to achieve the cursor movement. The Wadsworth group also published results for P300 based cursor control BCI. The P300 event related potential is a 300-600 msec positive signal generated in by a subject's response to the sensation of an expected stimulus. The P300 is used in conjunction with the oddball paradigm where stimulus events of low probability and events of high probability are presented to the subject [129].

The Graz BCI was developed by the Laboratory of Brain-Computer Interfaces, Institute for Knowledge Discovery at Graz University of Technology, in Austria (http://bci.tugraz.at/). Graz BCI developments were led by now Prof. Emeritus Gert Pfurtscheller. Prof. Christa Neuper presently leads the Graz activity [93]. Graz BCI implementations center on the use of the mu and beta rhythms to control devices using machine adaptation to control a neuroprosthetic device. Graz research activities include exploring sensors, feedback strategies, cognitive aspects and novel signal processing methods. Additionally, Graz is exploring general applications for BCI beyond assistive technologies for patients, such as virtual reality [91,92,93,94,95,96]. Recent work at Graz has addressed subject independent BCI, exploring both subject invariance and session to session transfer [2]. Specifically, it is this activity that has strong parallels to the research focus of this dissertation.

The Martigny or IDIAP BCI research project (http//:www.idiap.ch/) was similar to Berlin in the sense of the emphasis on relocating the burden of training the device to the machine learning algorithm. The Martigny project was conducted under the auspices of IDIAP Research Institute, Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne, Switzerland, by a team led by José del R. Millán, Ph.D. Professor Millán is now an associate professor with the Swiss Federal Institute of Technology in Lausanne (EPFL) where he continues research into the development of non-invasive brain-controlled robots and neuroprosthetic devices. The Martigny project employed a self-paced asynchronous BCI that responds every 0.5 seconds to one of three different mental tasks, plus an unknown state. The Martigny BCI used a periodogram algorithm to extract features and applied the features to a Gaussian classifier [79].

The development of BCI at the University of Tübingen, Germany was led by Drs. Thilo Hinterberger and Niels Birbaumer, of the Institute of Medical Psychology and Behavioral Neurobiology, within that university (http://www.ti.uni-tuebingen.de/BCI.856.0.html). Several approaches to BCI were undertaken in this organization focusing on mu rhythm, P300 and slow cortical potential signals. Notably, the development of the Thought Translation Device (TTD) was undertaken as a means to allow people to spell out letters with responses given by an EEG signal. While most BCI has been operated with visual feedback, research toward the use of auditory feedback was implemented to accommodate patients that have difficulties in focusing their gaze. BCI implementation at Tübingen employed support vector machine classifiers with features derived from auto-regressive coefficients [55,56].

Neil Squire Brain Interface lab at the University of British Columbia in Vancouver, BCI is home to research activities led by Jaimie Borisoff, Steve G. Mason and Gary E. Birch [17]. The initial focus has been on the development of assistive technologies in self-paced or asynchronous environments. This involves the issuance of commands only when control is intended and results in periods of 'command' or 'no command', fully allowing the user to determine when things happen. The Vancouver group and the parallel activity, the IDIAP group, were the two earliest labs to delve into self-paced BCI. This group was responsible for developing a BCI performance metric based on a ratio of true positive rate to false positive rate that could be viewed as a predecessor to the subject of this dissertation [8, 9,17,37,38,39].

While not an exhaustive list, these six labs represent a substantial sample of contributions to non-invasive EEG based BCI research over the last two decades. An extensive list of BCI research organizations and taxonomy of the research focus can be found in Mason and other sources which detail many successful approaches to BCI that have occurred over the past two decades [6,8,29,69,75].

While a majority of BCI approaches focus on the subject specific model using features with high subject-to-subject variances, subject invariant models have been appeared on the research landscape more recently [2,24,71]. One of the more recent goals has been toward obtaining reduced BCI training requirements through the identification and use of common features across either a single or a multi-subject population [2,33,37,70,110]. This approach is analogous to the scientific progress made in voice recognition technology in wireless phones, where early implementations required extensive calibration or training by the individual prior to using the device. Currently available voice recognition devices require little if any training and can be used right “out of the box”. The ideal BCI of the future can be expected to perform in the same manner. We investigate subject independent brain computer interfaces (SIBCI) with multiclass detection using a one versus the rest (OVR) classifier scheme. Comparisons between OVR and one versus one have been presented in the literature with few conclusions as to which provides better performance [34,102,123]. We chose the OVR scheme with a data driven feature selection or wrapper as described in section 2.2. OVR produces only one subset of features per class thus simplifying the architecture. We have devised a metric for evaluation of OVR classifier performance to guide the feature search algorithm.

The SFFS method was selected with the OVR multiclass SVM classifier. Several multiclass methods have been proposed in the literature [61,115,135], the OVR scales better than one versus one as the number of classes increases. For a simple 4-class BCI, which only uses one feature vector, a 4-class OVR will produce a maximum of 4 feature vectors. Using the same number of classes, the 4-class OVO method can require a maximum of 6 feature vectors, a 50% increase. The savings for OVO increases as the number of classes increases. Therefore with OVO, a smaller number of feature vectors must be evaluated to produce the final feature subsets.

We will also show that OVR discrete output classifiers such as the SVM, when faced with features that are not well separated can become trapped in degenerate states. When OVR is used, the performance metrics must account for these degeneracies. Metrics for measurement of BCI is an under-researched topic [111]. A simple calculation of correctly detected samples based solely on error rate produces a misleading result when degenerate cases are considered. Degeneracies are detrimental to the feature selection as search paths driven by misleading feedback can become trapped in local minima or infinite loops. A successful wrapper must detect and compensate for this condition. We propose a scalar metric that accounts for misleading classifier results based on the degree of degeneracy of the classifier.

## 2.4 SIBCI Challenges

The research into facets of BCI over the last 20 years has been extensive. This section details the most recent and relevant works in SIBCI. The goal of SIBCI is to provide a generalized system which can easily accommodate multiple users with a reduced training interval. Ideally, finding a universal set of features that reflect the brain state of all users would eliminate the need for any training for a BCI machine. Zero training BCI which avoids subject training altogether is at the forefront of current BCI research and recognizing that features representing brain states for multiple subjects exhibit a high degree of inter-subject variation, approaches are based upon using libraries of previously collected training samples. Using this paradigm, we explore methods to achieve subject invariance, while reducing the co-adaptation or training requirements for SIBCI. This problem is approached through the identification and analysis of the SIBCI obstacles and addressing them to improve the performance of the model. The connection between reduced training times and subject independence can linked by the idea that a system with zero training time is subject invariant. The ideal of zero training guarantees that a first time user can operate the system immediately, regardless of the previous history of the system. Such a system could be passed from one user to another without the need for training, thus achieving a pure form of subject invariance. Subject invariance is realized through 'pre-training' of a standard or common set of features reflecting each possible brain state to be implemented in the system. The features obtained from a common sample of the population would enable the self-paced SIBCI to be used with a reduced amount of training or even possibly with no training at all. The development of this approach calls upon two related areas of BCI research each of which contributes to our assertion that the incorporation of subject invariant features enables reduced training intervals for BCI algorithms. One research direction is the exploration of reduced training intervals for BCI and the second being the achievement of subject independence. Research was also drawn from the area of feature extraction methods, classifier architectures and BCI performance metrics.  This is a very active current research topic in BCI.

The reduction of training requirements for machine learning algorithms within BCI has been the underlying theme in research in the Berlin and Martigny BCI projects as early as 20 years ago. Achieving BCI with very sparse training data is still a very active BCI research area.

Achievement of subject independent feature patterns requires the collection and analysis of a variety of features within a high dimensionality feature space. Machine learning algorithms, which function in these high dimensionality spaces, must address a number of design complexities. A classifier must be chosen that provides sufficient Vapnik-Chervonenkis dimensionality to separate or shatter the feature space. While a classifier can usually be designed that achieves this goal, for high dimensionality feature space we run the risk of obtaining a classifier with an unacceptably large number of free parameters. This candidate classifier also faces the possibility of overtraining resulting from bias variance dilemma. An over-trained classifier will fit the training data precisely but cannot generalize to unseen data. A combination of techniques is employed to address these complexities. Our approach is to first reduce the dimensionality of the original features by searching for a subset of the features that correspond to our brain states. Other methods employed in this model include the use of a validation feature set to terminate the training phase, and by using a K-fold validation arrangement of the training and validation data sets.

### 2.5 Functional Model of a Brain Computer Interface System

While several proposals for BCI frameworks have been presented in the literature [63,72,76], one possible implementation of a BCI system, which consists of a human controlling a robotic arm, is shown in Figure 1.

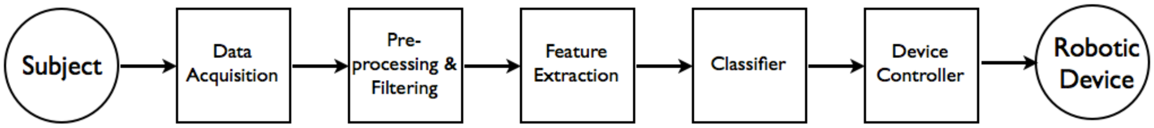


Figure 1. Conceptual Diagram of Brain Computer Interface Controller

The user is prompted by a visual instruction on a computer screen to execute a quasi-movement. The initiation of the visual prompt is accompanied by a 500 Hz audio tone to elevate the alertness of the subject at the beginning of each task. EEG signals are collected using non-invasive wet electrodes. The EEG amplifier digitizes and amplifies the EEG signals. The pre-processing stage selects the desired EEG channels and stores each epoch of the multichannel signals in a matrix format suitable for simultaneous batch processing in the next stage. This stage also consolidates the training data from each of the subjects into a single matrix. Where required, common spatial pattern filters are derived at this stage and applied to the signals. All feature extraction operations are applied in parallel to the output of the pre-processing stage. All of the feature vectors corresponding to an individual subject and trial are concatenated into a single feature vector. The feature vectors are reassembled into a matrix where each column is organized with respect to an individual, trial and task. The following stage comprises the wrapper feature selection. It is within this stage where this research is directed and as such, it will be the focus of most of the discussion in this dissertation. The output of this stage consists of a subset of the feature vectors applied to the wrapper input. The subsequent stage is a classifier that is trained by the subset of features selected by the wrapper. The output of the classifier is applied to a device controller algorithm that drives the robotic arm. The BCI may be used to drive another type of appliance or device, such as a cursor on a computer screen, or multimedia device by interfacing the output classifier to an appropriate type of device driver.

This functional model serves two purposes; first it clearly identifies the interfaces between the stages. This defines each stage allowing each function to be performed by a separate program, algorithm or subroutine. This provides interface locations and definitions suitable for troubleshooting and performance monitoring of intermediate stages. Second, it provides a degree of functional independence between the stages, permitting modularity where a function can be substituted by another different but albeit comparable function. An example of this would be to apply a different pre-processing algorithm such as blink artifact extraction, or the addition of a new and different feature modality, without requiring changes to the other stages.

Each stage consists of a single function or several closely related functions reflecting the current state of the art of BCI, signal processing and pattern recognition technologies. An advantage of this approach permits the comparison of this BCI implementation with that of others in terms of features used, types of classifiers and feature selection methods. This is an important consideration as there is no standardization for evaluation of BCI performance therefore a useful approach is through direct comparison of systems. Such an approach however, is imprecise at best due to the high degree of variation in the design of systems. The lack of a universally accepted BCI performance metric will be addressed in this dissertation.

## 2.6 BCI Applications

The preponderance of BCI research has focused on assistive technology that enables individuals without motor function control to communicate or control devices in their environment such as a wheelchair or neuroprosthetic device. The risk versus benefit of this type of application may be amenable to the application of wet scalp-based electrodes with complex preparation procedures or even invasive electrodes using implant technologies for severely disabled users. Applications for casual users or applications that require continuous sessions lasting more than a few hours would require dry electrodes.

Other potential applications have been identified for the general public and for industrial and or military applications. Gaming applications are clearly a candidate for BCI, as current implementations generally rely on the collection of EMG (electromyograph) potentials conducted by muscle and skin tissues. These signals are greater in magnitude than EEG and are easily detectable by simple dry electrodes. Research in the use of BCI to provide navigation through virtual reality environments is a strongly researched area at Graz. Some research has been focused on the control of a vehicle or in the performance of auxiliary functions required by pilots during the flight of an aircraft such as navigation or adjusting flight controls or speeding the response to the operation of a weapons control system. Such applications can depend on signals generated within the sensorimotor cortex region of the brain or extend to cognitive and/or affective domains.

While the methods and algorithms presented in this dissertation were applied to non-invasive scalp based EEG signals they are not uniquely dependent on this application method and can easily be applied to signals that are based on other technologies such as NIRS.

## 2.7 BCI Performance Metrics

Multiple reasons for an acceptable, and ideally, universal evaluation criteria for BCI systems can be identified. Comparative evaluation of different systems to allow model selection for a BCI is one of the foremost drivers for the use of BCI performance information. In light of the varied implementations of BCI, setting up a single method as the choice of a performance metric is a very challenging task. As a result of the variation in BCI implementations, several performance metrics have found use, with one group of metrics being more applicable to continuous output classifiers and another more apropos to discrete output classifiers. The classical choices include overall accuracy rates (or error rate), Cohen's Kappa coefficient, mutual information indices, response time and receiver-operator characteristic (ROC) based indices. Additional methods that have been recently introduced include the Utility metric [30] and the HF-difference method [58].

The confusion matrix has been demonstrated as the best descriptor for the results of an M-class classifier problem [4,34]. While our problem is actually a 4-class system, the use of an OVR arrangement permits the results to be shown as a confusion matrix for two classes with an unequal number of members in each class.

The most widely used evaluation method in BCI research is the overall agreement or classification accuracy (ACC) [111]. The primary weaknesses of this method are that it does not include any influence of the off-diagonal values of the confusion matrix and is also weighed in the direction of the more frequent classes in the OVR scheme, where the rest of majority class has a larger number of members.

Several multiclass BCI implementations use the Cohen's kappa coefficient as a performance metric. [112,126] The Cohen's factor uses overall agreement $p_o = ACC$ or classification accuracy and chance agreement $p_e$

$$p_e = \frac{\sum_{i=1}^{M} n_{:i} n_{i:}}{N^2} \tag{1}$$

the product of the sum of the ith column and ith row is represented by $n_{:i} n_{i:}$ where $\frac{n_{:i}}{N}$ is the a posteriori probability and $\frac{n_{i:}}{N}$ is the a priori probability [111].

The kappa coefficient $\kappa$ is calculated from [25]

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

and for M-classes with the number of samples equally distributed across classes the overall agreement or $p_o$ , a quantity which is equivalent to accuracy (ACC), is thus calculated from

$$p_o = \frac{M\kappa - \kappa + 1}{M} \tag{3}$$

Using the view of the BCI system as a communication channel between the subject and the external device, we can evaluate the channel in terms of an information transfer rate by applying the communication theory devised by Shannon [116].

As a starting point to the derivation of Wolpaw's information transfer rate, Farwell and Donchin [36] proposed this formula for the information transfer of I in bits for an M-class error free BCI system

$$I \ = log_2(M) \tag{4}$$

Given a confusion matrix where X represents the input of a BCI in the form of the users imagined brain state and Y represents the output presented by the classifier. Defining H(X) and H(Y) as the entropies of the respective discrete random variables

$$H(X) = -\sum_{j=1}^{M} p(x_j) \cdot log_2\left(p(x_j)\right) \tag{5}$$

and

$$H(Y) = -\sum_{j=1}^{M} p(y_j) \cdot log_2\left(p(y_j)\right) \tag{6}$$

where

$$p(y_j) = \sum_{i=1}^{M} p(x_i) \cdot p(y_j|x_i) \tag{7}$$

the information transfer for a general confusion matrix can be shown as [87]

$$I(X;Y) = H(Y) - H(Y|X) \tag{8}$$

where

$$H(Y|X) = -\sum_{i=1}^{M}\sum_{j=1}^{M} p(x_i)\cdot p\left(y_j|x_i\right)\cdot log_2\left(p\left(y_j|x_i\right)\right) \tag{9}$$

Therefore we can obtain the mutual information from

$$I(X;Y) = \sum_{i=1}^{M}\sum_{j=1}^{M} p(x_i)\cdot p\left(y_j|x_i\right)\cdot log_2\left(p\left(y_j|x_i\right)\right)$$

$$-\sum_{j=1}^{M} p\left(y_j\right)\cdot log_2\left(p\left(y_j\right)\right) \tag{10}$$

where the apriori probability for each class $x_i$ is represented by $p(x_i)$ and the probability to classify each $x_i$ as $y_j$ is described by $p(y_j|x_i)$. Wolpaw's expression for information transfer rate can be derived by assuming a BCI where each of the M classes has the same a priori probability $p(x_i) = \frac{1}{M}$. An additional assumption that the accuracy ACC is equal for each class, $p(y_i|x_i) = p_o = ACC$. For equally distributed incorrect choices where $j \neq k$, then $p\left(y_j\middle|x_i\right) = \frac{(1-p_o)}{M-1}$.

The entropy of Y, H(Y) and H(Y|X) are represented by

$$H(Y) = log_2(M) \tag{11}$$

$$H(Y|X) = -\sum_{i=1}^{M}\frac{1}{M}\left[p_o\cdot log_2(p_o) + \sum_{j\neq i}\frac{(1-p_o)}{(M-1)}\cdot log_2\left(\frac{(1-p_o)}{(M-1)}\right)\right] \tag{12}$$

which yields

$$I(X;Y) = \ log_2(M) + p_o\cdot log_2(p_o) + (1-p_o)\cdot log_2\left(\frac{(1-p_o)}{(M-1)}\right) \tag{13}$$

and this is equivalent to Wolpaw's formula [136] for information transfer rate. We use this metric to extrapolate the performance of a BCI with a given error rate and number of tasks to a different number of tasks. This is useful in showing the behavior trends as the number of classes M increases. Information transfer rate curves for 2,3 and 4 classes are provided in Appendix A.

ROC based performance metrics can only be derived from continuous output classifiers and are not applicable to our study. Likewise, HF-difference methods are applicable to self-paced BCI systems. Response time metrics do not directly address the separability of the data. The Utility metric described by Dal Sarno [30] addresses a complete system and could provide a useful alternative to the ITR, but not as a method for determining feature subsets within the context of a wrapper algorithm.

The use of true positive rate / false positive rate (TPR/FPR) as an evaluation metric for 2-class self-paced subject specific BCI systems was described by Fatourechi [38]. This method is closely related to the approach described in this dissertation with the following distinctions. The TPR/FPR ratio is used as a method for guiding a hybrid genetic algorithm for feature selection in self-paced BCI. However this metric does not perform successfully when zero false positive rates are encountered. While Fatourechi has identified that the occurrence of false positive outputs is a substantial detriment to self-paced BCI, the focus in this dissertation is toward a different issue. One of the major goals of this research was to address the behavior of classifiers exhibiting very low or zero false positive rates while also accompanied by very high false negative rates. The approach described in this dissertation offers a robust solution to this issue by performing successfully regardless of the occurrence and location of positives or negative outputs. Fatourechi also projected TPR and FPR into a 2-dimensional domain, identifying two analytical regions. The method described in this dissertation uses a similar 4-region characterization of the accuracy and imbalance domain.

While the methods described in this section can be useful for model selection, our application of performance metrics is distinctively different. Within the wrapper search approach each possible subset of features must be evaluated in sequence by a classifier, which then generates a cost function to guide the induction (or search) algorithm. This metric must be a scalar value suitable for a comparative evaluation of each of the feature subsets, while also compensating for biases and degeneracies in this classifier behavior.

Among these BCI performance metrics, there have been variations and hybrids, often the choice of metric in a given experiment is made without justification. Comparisons of BCI metrics are offered in in the literature [39,77,78,111].

## 2.8. BCI Feature Space

The challenge of subject independent BCI is magnified by the inclusion of multiple subjects with features that have high inter-subject variance. The inter-subject variance of electroencephalographic (EEG) signals [104] produces an increase in the task invariance of the features as more subjects are introduced as illustrated in Figures 2, 3 and 4. Four BCI tasks are projected into an arbitrary 2-dimensional feature space in Figure 2, where any possible feature vector can be represented as a point in the space.

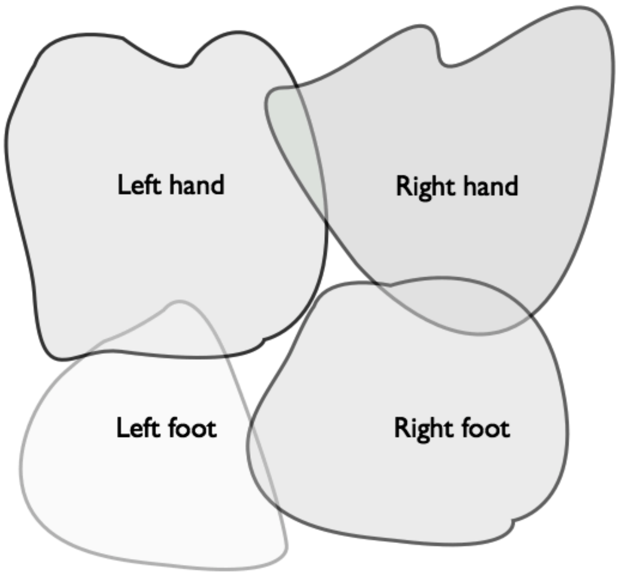


Figure 2. Four BCI Tasks represented in two-dimensional feature space

For a finite and fixed Vapnik-Chervonenkis dimension, the classification error increases as the number of subjects or tasks increases [88,127,128]. The Vapnik-Chervonenkis dimension identifies the capacity of a learning machine such as the support vector machine. Support vector machines can be designed with theoretically infinite V-C dimensionality however this does not produce a classifier that possesses favorable generalization qualities. Once a learning machine is designed it is constrained to $h$, a fixed VC dimension. The ability of a classifier to generalize indicated by the measured error or risk $R(\alpha)$ is bounded by [30,127]

$$R(\alpha) \leq R_{emp}(a) + \Phi\left(\frac{h}{N}\right) \tag{14}$$

with $\alpha$ representing the parameters of the classifier, $R_{emp}(a)$ or empirical risk term describing the error measured in the training set and $\Phi\left(\frac{h}{N}\right)$ describes the confidence interval as a function of *N,* the number of observation pairs and VC dimension. For a large VC-dimension, the confidence interval is large however the training error becomes very small. This relationship shows that an infinite VC classifier can "memorize" the training data and that the actual test error value can be much larger than that of the training data. A high VC dimension requires a very large number of points $N$ to assure a classifier with good generalization capability where the test error is less than or equal to the training error. The structural risk minimization principle (SRM) asserts that while the empirical risk term has an inverse relationship to the VC dimension, there is some minimum to the upper bound for the overall generalization error term. These two relationships describe the classifier generalization behavior known as the bias-variance dilemma. We can draw upon this knowledge of bias-variance related errors to introduce an idea regarding the behavior of classifiers. We cannot categorically state that if a given classifier incorrectly detects a specific test (or validation) sample then its performance will improve from increasing the VC dimension. Therefore, when faced with a large and diverse group of feature samples which exhibit varying degrees of inter-class and inter-subject variance it may be neither feasible or desirable to design a classifier that can separate or shatter all regions of the feature space. It may not be possible to design a perfect classifier and additionally, Bishop has shown that 2 class classifiers used in multiclass arrangements will always be unable to discriminate in some portion of the feature space [10]. Recognizing this, we propose a feature selection method to compensate for those regions in the feature space where inseparable features exist. To illustrate some of the basic challenges in achieving error-free SIBCI we draw from modern set theory. Modern set theory was introduced in 1874 by Georg Cantor and has been refined by many others since then [22]. To illustrate these regions we represent each set of features for M tasks and P subjects as $T_i^k$where $k \in \{1,2,3, \dots, \mathrm{M}\}$ corresponds to the number of tasks and $i \in \{1,2,3, \dots, \mathrm{P}\}$ the number of subjects. Figure 3 shows each of these entities as distinct regions in the arbitrary two-dimensional feature space, for *i*=2 and *k*=4.

In Figure 4 the subject regions for each task are conjoined to form distinct but overlapping task regions. The feature space for each task $T^k$ is depicted as

$$T^k = \bigcup_{i=1}^{P} T_i^k \qquad \forall i \in \{1,2, \dots, P\} \tag{14}$$

$$N(T^k) \geq N(T_i^k) \tag{15}$$

*N*(*T*) indicates the cardinality of each feature *T* and shows that the subject invariant task may occupy a larger area of the feature space than for any individual subject. Each *N(T)* increases as more subjects are included in $T^k$.

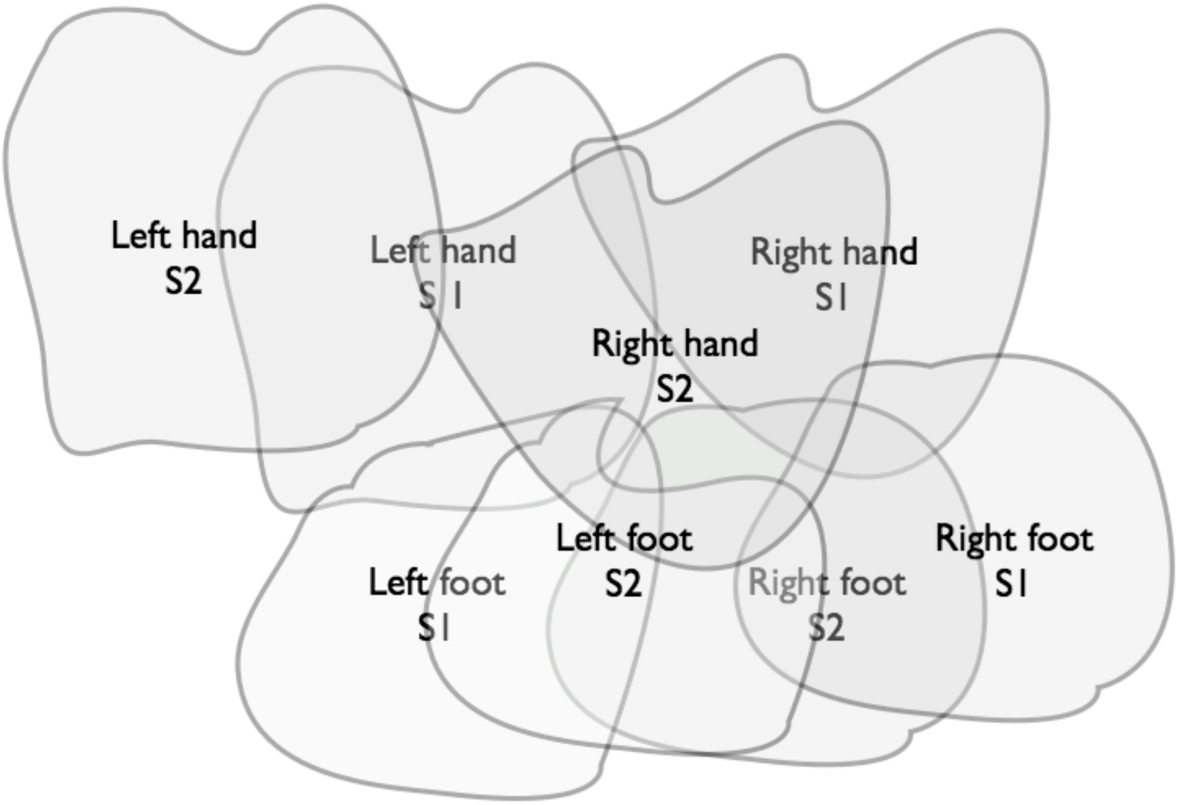


Figure 3. Four tasks represented in two-dimensional feature space for two subjects

For a subject specific BCI the ideal feature space for each subject $i \in \{1,2,3,\dots,\mathrm{P}\}$ would describe clearly separated features. For each possible combination of $j, k \in \{1,2,\cdots,M\}$ where $j \neq k$

$$T_i^j \cap_{j,k=1,j\neq k}^{M} T_i^k = \emptyset \tag{16}$$

describes a clear and distinct separation between the feature spaces of any pair of tasks, or no overlapping regions.

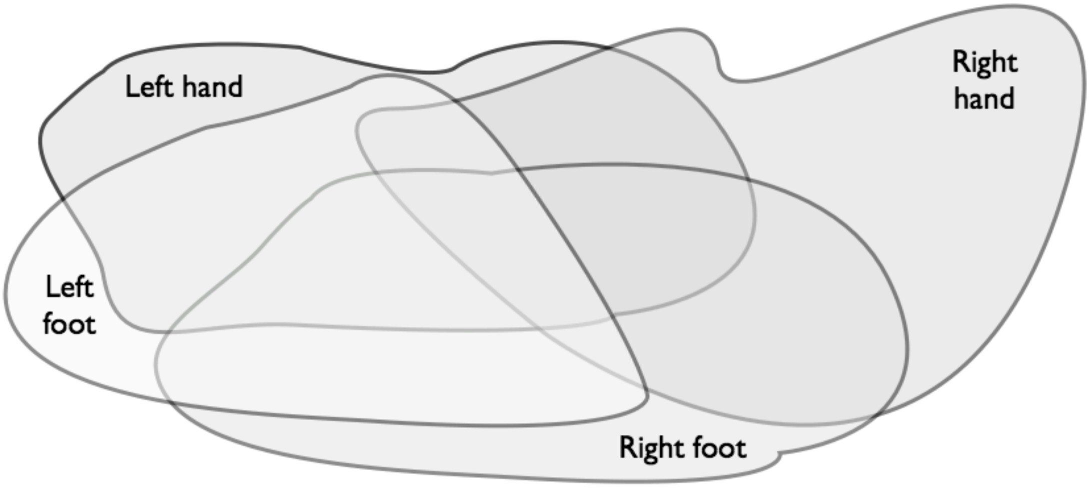


Figure 4. Four tasks represented in two-dimensional fature space for multiple subjects

For an ideal subject independent BCI we would observe the following relationship where subject indices are modified to indicate $i, \hat{\imath} \in \{1,2,3,\dots,\mathrm{P}\}$

$$\left(\bigcup_{i=1}^{P} T_i^j\right) \bigcap_{\substack{j,k=1 \\ j\neq k \\ i\neq \hat{\imath}}}^{M} \left(\bigcup_{\hat{\imath}=1}^{P} T_{\hat{\imath}}^j\right) = \emptyset \tag{17}$$

This also portrays a feature space where a clear separation exists between each of the $M$ task spaces. However, each task space is the union of only those features representing each specific task aggregated across all of the subjects.

It is well established in BCI that a degree of intersubject variance occurs in when subjects are asked to perform an imagined motor task [2,24] For a single BCI task, $j$, and two subjects, intersubject variance for task $j$, is represented by the area in region $V^j$

$$V^j = \left(T_1^j \cup T_2^j\right) - \left(T_1^j \cap T_2^j\right) \tag{18}$$

and for three subjects the variation term becomes

$$V^j = \left(T_1^j \cup T_2^j \cup T_3^j\right) - \left(T_1^j \cap T_2^j \cap T_3^j\right) -$$
$$\left(\left(T_1^j \cap T_2^i\right) - \left(T_1^j \cap T_2^j \cap T_3^j\right)\right) - \left(\left(T_1^j \cap T_3^j\right) - \left(T_1^j \cap T_2^j \cap T_3^j\right)\right) -$$
$$\left(\left(T_2^j \cap T_3^j\right) - \left(T_1^j \cap T_2^j \cap T_3^j\right)\right) \tag{19}$$

which can be rewritten as

$$V^j = \left(T_1^j \cup T_2^j \cup T_3^j\right) + 2\left(T_1^j \cap T_2^j \cap T_3^j\right) -$$
$$\left(\left(T_1^j \cap T_2^j\right) + \left(T_1^j \cap T_3^j\right) + \left(T_2^j \cap T_3^j\right)\right) \tag{20}$$

We want to minimize this expression, which grows in complexity as the number of subjects increase. We can see immediately that $V^j = \emptyset$ when $T_1^j = T_2^j = T_3^j$ using the relationship that for every $A = B \neq \emptyset, A \cup B = A \cap B = A = B$. In the strictest sense this requires that each $T_k^j$ has identical elements, for our case we can accept a weaker condition that the elements fall within the same common boundary and that there is no overlap between any of the pairwise combinations of $T_k^j$. In other words, the features representing all subjects performing the same task will be contained in a unique subregion of the feature space for each task. The variance is minimized when all three subjects occupy the identical feature space. In a real world BCI the value of $V^j$ is finite and grows as the number of subjects increases and as the dissimilarity between the individual subject feature spaces increases (indicating a smaller overlap region). For more subjects there will be a greater number of intersection terms that will contribute to the growth of $V^j$.

In contrast, for a BCI we wish to maximize the intertask variation in the feature space. We can perform a similar analysis for intertask variation. For two tasks, 1 and 2, assuming one subject $i$, the intertask variation between task 1 and 2 can be described by minimizing the overlap between the spaces, which is the equivalent to minimizing the expression

$$V_i = (T_i^1 \cup T_i^2) - (T_i^1 \cap T_i^2) \tag{21}$$

which expands for three tasks to a similar expression to that derived earlier

$$V_i = (T_i^1 \cup T_i^2 \cup T_i^3) + 2(T_i^1 \cap T_i^2 \cap T_i^3) - \left((T_i^1 \cap T_i^2) + (T_i^1 \cap T_i^3) + (T_i^2 \cap T_i^3)\right) \tag{22}$$

While this is a relatively complex expression for a concept that may be intuitively obvious it does show that the intertask variation is maximized by reducing all of the terms of the form $T_i^j \cap T_i^k$ to the null space $\emptyset$. This requires that there is no overlap between all possible pairwise combinations of task feature spaces. The relationship

$$\forall \mathrm{A}, \mathrm{B}, \mathrm{C} \quad A \cap B = \emptyset \Longrightarrow A \cap B \cap C = \emptyset \tag{23}$$

guarantees that any intermediate terms that consist exclusively of intersections of more than two spaces also vanish from the expression.

For a space with four tasks a general expression for $V_i$ can be written as

$$V_i = (T_i^1 \cup T_i^2 \cup T_i^3 \cup T_i^4) + \alpha(T_i^1 \cap T_i^2 \cap T_i^3 \cap T_i^4) - \left[\left(\beta_1(T_i^1 \cap T_i^2 \cap T_i^3) + \beta_2(T_i^1 \cap T_i^2 \cap T_i^4) + \beta_3(T_i^2 \cap T_i^3 \cap T_i^4) + \beta_4(T_i^1 \cap T_i^3 \cap T_i^4)\right) + \left(\gamma_1(T_i^1 \cap T_i^2) + \gamma_2(T_i^1 \cap T_i^3) + \gamma_3(T_i^1 \cap T_i^4) + \gamma_4(T_i^2 \cap T_i^3) + \gamma_5(T_i^2 \cap T_i^4) + \gamma_6(T_i^3 \cap T_i^4)\right)\right] \tag{24}$$

where $\alpha, \beta_i, \gamma_j \in \mathbb{Z}$, $\beta_i, \gamma_j \geq 0$ and $|\alpha| \leq \sum_i \beta_i + \sum_j \gamma_j + 1$.

Given the bounds on $\alpha$, $V_i$ is maximized when the all of the terms in square brackets are zero, implying no overlapping regions in the feature space. $V_i$ minima occur when all subspaces occupy exactly the same region.

While these expressions describe the ideal case, they illustrate, using set theory, basic concepts pertaining to separating features in BCI. The increased complexity of feature separation brought about by the inclusion of multiple subjects is shown. While these examples depict inter-subject variation in the feature space, the same presentation can be used to describe temporal intra-subject variations. Features collected at different time periods from one subject can be treated as being collected from a separate subject.

A unique feature space $T_{SI}^k$ for each subject invariant feature $k$ is required for correct operation of the classifier

$$T_{SI}^{k} = T^{k} - \bigcup_{j} T^{j} \qquad \forall j : j \in \{1,2,\dots,\mathrm{M}\}, j \neq k \tag{25}$$

As M increases, the faster growth of the second term on the right hand side reduces the probability of finding a location in the unique feature space for each individual task. This is presented graphically in Figures 3 and 4. As additional subjects are included, the region uniquely occupied by each $T_{SI}^k$ becomes smaller. The area of this region corresponds to the probability of correctly and uniquely mapping a feature to a task. This probability decreases as additional subjects or tasks are included [11,24,33,111].

Alternatively, Wolpaw [136] described this using the information capacity of a BCI channel, as derived from Shannon's theorem [51, 116]. A complete derivation of this formula was shown in Section 2.7. Information transfer rate (ITR) is expressed in bits per second ITR = B/T where B is bits per trial and T is the number of seconds per trial. B shows the effect of increasing the number of tasks or classes on the probability of correct detection

$$B = log_2 N + Plog_2 P + (1 - P)log_2\left(\frac{1-P}{N-1}\right) \tag{26}$$

N equals the number of tasks and P is the probability of correct detection. The difficulty of increasing the number of tasks has been evidenced in the results of existing BCI research as few designs using more than 4 or 5 tasks have been successfully implemented [8, 68, 89].

### 2.9. Wrapper Methods for Feature Selection

To obtain an efficient set of common features for subject independent BCI, we assembled a large number of high dimensionality feature vectors and conducted a non-exhaustive search for subsets of features to train the SVM classifiers. Figure 5 shows a wrapper method which choses features by using an induction algorithm guided by the performance of the actual classifier [64]. The search algorithm uses an optimization criteria derived from feedback from the SVM classifier.

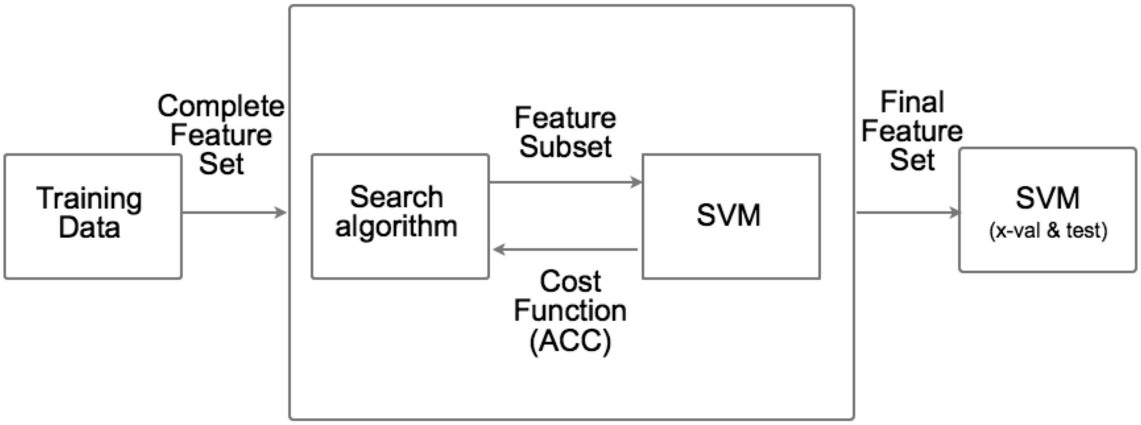


Figure 5. Wrapper Architecture Block Diagram

As such, it is imperative that this optimization criterion accurately reflects the outcome of the classifier. Any degenerate mode that skews or dominates the results so as to ‘swamp out’ meaningful results will deteriorate the performance of the wrapper. This can result in wrapper algorithms that converge to incorrect solutions or fail to converge at all. Given an appropriate correction for this behavior, a wrapper method for feature selection can be well suited for multi-class SIBCI.

## 2.10. Multiclass Classifier Design Issues

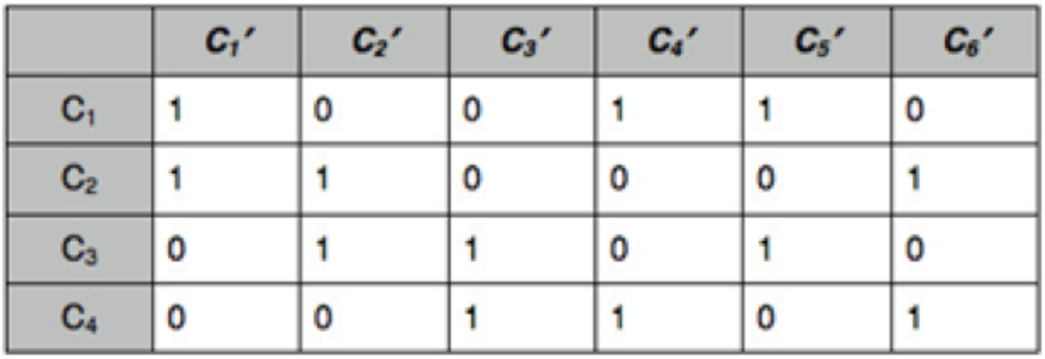
Our multiclass BCI classifier scheme requires us to create a model, based on a training data selected from each of M mutually exclusive classes that can accurately determine the class of a new test sample belonging to one of the M classes. While this is fundamentally a multiclass problem, the ease and flexibility of designing binary classifiers suggests recasting it as a two class or binary problem [3,37,102]. The support vector machine classifier was chosen as a classifier that has been successfully used for binary classification tasks [125]. The support vector machine has been previously described and produces discrete [-1, +1], as opposed to continuous, outputs. These can be scaled and offset to produce a one-zero result. The result indicates the classifier decision regarding membership of the data sample in one of two classes. While several independent approaches to adapting binary classifiers to multiclass problems have been discovered, they can be generally grouped into three families, of which we shall consider two that are significant (each recasts a multiclass problem into a binary one). The two groups of interest are the one versus the rest (OVR), sometimes called one versus all, or the one versus one (OVO), sometimes called all versus all (AVA) [102], in this dissertation we will use the former terminology.

A third option, the single-machine approach, where SVMs perform a single M-class optimization [19,67,125,134] is only of theoretical interest and included for completeness. In many ways, single machine methods are simultaneous reductions of either the OVA or OVO approaches and do not present any relief to the issues addressed in this dissertation. For example, the Directed Acyclic Graph support vector machine (DAGSVM) [27] still requires M(M-1)/2 internal binary SVM nodes and does not substantially reduce the number feature selection decisions over the OVO approach [114]. Evaluations found in the literature indicate that single machine implementations generally yield inferior performance to that of ether OVA or OVR methods when faced with finite-sized, high dimension data samples [102]. Single machine SVM's have been shown to only asymptotically approach the performance of OVA and OVR machines with infinite sets of data and at the cost of increased complexity of implementation and longer training periods [41,57,102]. We have not found any single machine SVM implementations in BCI. Beyond this immediate discussion, single machine implementations were given no additional consideration in this dissertation.

Both the OVA and OVO approaches to multiclass BCI can be found throughout the literature [8,27,47,68,88].

**2.10.1 One versus one classifier design.** The OVO approach uses binary classifiers with each class containing an equal number of samples, where a separate classifier is trained for each possible pairwise combination of classes [33,102]. OVO requires $\binom{M}{2}$ = M(M-1)/2 classifiers. Once these classifiers are created they must be aggregated in a manner that will produce M distinct outputs that reflect each of the possible input conditions. This requires a fusion method *D* that maps the L-metaclasses

$\mathbb{A}^L$into M-new metaclass groups according to $D: \mathbb{A}^L \rightarrow \mathbb{A}^M$ where $\mathbb{A} \in [0,1]$ and L=M(M-1)/2 and L>M. These classifier aggregation methods are known as ensemble, mixture, fusion or committee methods. Several ensemble strategies are available to accomplish this mapping or fusion. Fusion strategies have been shown to achieve higher classification accuracy and robustness for generalization, by cancelling out uncorrelated decision errors [10]. This particular application however may use the same techniques as an output level mixing method to perform the function *D* described above in our multiclass arrangement. The error correcting output code (ECOC) method is specifically designed for the multiclass problem, where the M class problem is split into a series of dichotomous problems [66]. If we split the M classes into L=M(M-1)/2 binary classes we can identify an M × L dimension code matrix, as shown Figure 6 for M=4.



| | $C_1'$ | $C_2'$ | $C_3'$ | $C_4'$ | $C_5'$ | $C_6'$ |
|---|---|---|---|---|---|---|
| $C_1$ | 1 | 0 | 0 | 1 | 1 | 0 |
| $C_2$ | 1 | 1 | 0 | 0 | 0 | 1 |
| $C_3$ | 0 | 1 | 1 | 0 | 1 | 0 |
| $C_4$ | 0 | 0 | 1 | 1 | 0 | 1 |

Figure 6. Error Correcting Output Code Matrix for M=4.

Each column represents the output for one of the L classifiers created by a pair of classes. For example, **$C_1$'** indicates a meta-classifier which is only discriminating between classes $\omega_1 \cup \omega_2$ as opposed to $\omega_3 \cup \omega_4$ as indicated by the first column word 1100, and **$C_2$'** is derived from classes $\omega_2 \cup \omega_3$ versus $\omega_1 \cup \omega_4$, (0110) and so forth. **$C_1$'** votes 'yes' in the presence of either $\omega_1 \cup \omega_2$ and 'no' otherwise. This continues for each of the other metaclassifiers **$C_2$'** through **$C_6$'** according to each respective column. Each row of this code matrix forms a codeword corresponding to a class, each column of the matrix corresponds to one of the classifiers. This produces a method for using balanced class label frequencies and binary classifiers to solve the multiclass problem. Each of the classifiers **$C_1$'** through **$C_6$'**, is a binary classifier where each class assignment is based on an equal frequency of positive and negative and class labels representing every possible pairwise combination of the M original classes. If we count the frequency of each metaclassifier's positive vs. negative targets over the entire dataset, each which has equal portions of $\omega_1$, $\omega_2$, $\omega_3$ and $\omega_4$, one can see that the ratio is now 50%-50% [118]. We can also see from this ECOC implementation the Hamming code distance is 4 between each pair of codewords or rows.

It is necessary to tally the results of the 6 classifiers to recover the original four-class information. Given a classifier which produces labels $(s_1, \cdots, s_L)$ for a given input **x**, the Hamming distance between the classifier outputs and the codewords is calculated. The decision for the correct label for **x** is assigned to the class with the shortest Hamming distance. Given a code matrix C where C($i$,$j$) indicates the *(i,j)th* element of the C where $i$ represents the number of metaclasses and $j$ represents the original class numbers. The presence of a 1 in the C(i,j)th element indicates membership or a 'yes' vote and 0 indicates a 'no' vote. The support for each class $\omega_j$ can be expressed according to [65]

$$\mu_j = \sum_{i=1}^{L} |s_i - C(i,j)| \tag{27}$$

where the decision for the correct class $\omega_j$ is made according to the minimum value of $\mu_j$.

Fusion was also employed to combine left and right hemispheric feature classifiers or feature modality classifiers through majority vote using plurality consensus patterns, the underlying justifications for these protocol and feature selection choices will be discussed in Chapter 6. This method is analogous to electoral theory with roots back to ancient Greek city-states and the Roman Senate [65]. Assuming that the label outputs of all L classifiers are expressed as M dimensional binary vectors $\left[d_{i,1}, \cdots d_{i,M}\right]^T \in \{0,1\}^M$, $i$=1, …, L, where $d_{i,j}$ = 1 if $D_i$ labels **x** in $\omega_j$ and 0 otherwise. The majority (plurality) vote will select class $\omega_k$ whenever

$$\sum_{i=1}^{L} d_{i,k} = \max_{j=1,M} \sum_{i=1}^{L} d_{i,j} \tag{28}$$

the total number of classifiers, L, is based on the number of feature modalities and hemispheres being combined in the classifier output. A variation of this which can be used in conjunction with ECOC by is implemented summing M(M-1)/2 element output vectors of the L classifiers and then thresholding and normalizing the scores. The resulting vector $\left[c_1, \cdots c_{M(M-1)/2}\right]^T \in \{0,1\}^{M(M-1)/2}$, is used to calculate the Hamming distance to the ECOC codewords where each $c_j$ for j = 1, …, M(M-1)/2 according to [65]

$$c_j = \left\lfloor a + \frac{1}{L} \sum_{i=1}^{L} d_{i,j} \right\rfloor \quad 0 \le a \le 1 \tag{29}$$

The constant term, $a$, in the sum represents a global bias that can be adjusted to provide a threshold value between a one (indicating class membership) and zero (indicating the complement of class membership) for each meta-classifier output. While $a$ is normally set to 1/2, the value may be adjusted to increase or decrease the sensitivity to class membership. A smaller value of $a$ will produce more zeros in the output vector and likewise a greater value of $a$ produces more ones.

**2.10.2. One versus the rest classifiers.** Alternatively, the OVR approach treats the M class problem as M two-class problems thus requiring only M classifiers. The advantage of using the OVR scheme in the induction algorithm within a wrapper feature selection is simplicity, a reduced number of operations, a reduced number of free parameters and a reduced number of decisions to make in the reassembly process. For each class, we need only conduct one set of searches for OVR. A distinct feature subset is produced for each of the M classes or tasks. In contrast, the one versus one arrangement increases the model complexity by requiring M(M-1)/2 iterations of searches. With OVO M(M-1)/2 individual feature subsets are generated. These feature subsets are attributed to pairs of tasks and none of which are uniquely assigned to one specific task. For example, with M=4, OVA produces 4 feature subsets, one for each of the classes. OVO produces 6 feature subsets, requiring additional steps to combine the features to produce four classes. This method introduces a number of additional free parameters and complexity into the classifier model.

Generally speaking, for each class, there is an optimal discriminant function $g_i(\boldsymbol{x})$, $i = 1,2,. . ., M$, so that $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \forall\, j \neq i,$ , if $x \in \omega_i$. Designing the discriminant function so that $g_i(\boldsymbol{x}) = 0$ separates class $\omega_i$ from all of the others where each classifier should produce $g_i(\boldsymbol{x}) > 0$ for $x \in \omega_i$ and $g_i(\boldsymbol{x}) < 0$ otherwise. Classification is achieved according to the rule: [123]

$$\text{Assign } \boldsymbol{x} \text{ in } \omega_i \text{ if } i = \arg\max_k \left[ g_k(x) \right] \tag{30}$$

Indeterminate regions where either more than one or no $g_i(\boldsymbol{x})$ is positive can result from the overlapping regions such as those shown in Figure 4.

An OVR classifier with an equal frequency of members in each of the M classes has M-1 greater number of training samples in the non-target class than in the target class. The unequal size of the two groups biases the overlapped region of Figure 4 toward the non-target class. The resulting classifier can fail to recognize members of the target class and impede the search algorithm.

Another shortcoming of the OVR classifier has been presented by Duda and Hart [34] and also by Bishop [10] Construction of a multiclass discriminant from two or more sets of two class discriminants invariably leads to regions of input space that are ambiguously classified. This can be easily shown for linear discriminant functions but can also apply to other two class classifiers such as support vector machines.

When the OVR classifier fails to recognize any members of the target class, the classifier accuracy calculated from the number of true positive samples divided by total samples is (M-1)/M %. In the reverse situation, where the classifier fails to recognize any members of the non-target class a 1/M % result is obtained. This is the same accuracy value calculation as for a random classifier. In the (M-1)/M% condition, the classifier produces misleading results of 75% for a 4 class machine and the result increases as M grows. This degenerate mode gives a meaningless accuracy result therefore it must be detected and adjusted. Detailed examples of this concept are presented in Section 4.1.

Some weaknesses in the kappa coefficient, introduced in Chapter 2.7, have been identified [50,111]. The kappa coefficient is a scalar value derived from the confusion matrix; it reflects the correlation between the actual and predicted classes. Cohen’s coefficient is a method of chance correction and has been shown to exhibit a bias based on trait prevalence (the existence or deficiency of true positives) as demonstrated by Gwet [50]. Chance correction does not mitigate the (M-1)/M % degeneracy problem since this degeneracy exhibits extreme trait prevalence behavior.

## 2.11 BCI as a Multiobjective Optimization Problem

BCI, like most phenomena modeled in the real world, represents a multiobjective optimization problem. A simple optimization problem with a single objective can be stated as [20,82,83]

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{subject to} \quad g_j(\boldsymbol{x}) \leq 0, j = 1,2,\cdots,M \tag{31}$$

*f(x)* is the objective function to be minimized, $\boldsymbol{x} \in \Omega$ is a decision vector in the parameter space $\Omega \subset \mathbb{R}^n$ and and $g_j(\boldsymbol{x}) \in \mathbb{R}$ is the set of constraints.

In contrast, a classical multiobjective optimization problem with j constraints and k objectives can be written as [82,83]:

$$\min_{\boldsymbol{x}} f(x) = \{f_1(\boldsymbol{x}), \cdots, f_k(\boldsymbol{x})\}$$

$$\text{subject to} \quad g_j(\boldsymbol{x}) \leq 0, j = 1,2,\cdots,M \tag{32}$$

$x \in \mathbb{R}^n$ is an n-dimension decision variable, and $g_j(\boldsymbol{x})$indicates the constraints on the decision vectors so that $\boldsymbol{x} \subset C$ where $C$ represents the region of feasibility $C \subset \mathbb{R}^n$ or; $f(x)$ consists of a set of k objective functions that are sought to be jointly minimized (or maximized, if the optimization problem is formulated in such manner). The weighted aggregation method is one way to solve this problem by creating a single linear composite objective function $z$ of the form [27,82]

$$\min_{x} \boldsymbol{z} = \sum_{i=1}^{k} a_i\, f_i(\boldsymbol{x}) \ , \tag{33}$$

$z$ is a weighted sum of objective functions and weighted by each $a_i > 0$. The objective vectors, $\boldsymbol{z},$ span the objective space. The feasible space is the subspace spanned by all p objective vectors that satisfy the constraints. Goal programming is a variation of this method where $|f_i(\boldsymbol{x}) - t_i|$ is used in lieu of $f_i(\boldsymbol{x})$. $|f_i(\boldsymbol{x}) - t_i|$ represents the magnitude of the deviation error from a predetermined target objective $t_i$.

A second approach, called the $\varepsilon$ constraint method, places constraints on all but one of the objective functions, which can be stated as [82]

$$\min_{x} f_j(x)$$

$$\text{subject to} \quad f_i(\boldsymbol{x}) \leq \varepsilon_i, \forall\ i = 1, N; i \neq j \tag{34}$$

where $\boldsymbol{\mathcal{E}} = \{\varepsilon_1, \cdots, \varepsilon_N\}$ describes the constraint limits.

The optimal solution that is optimal for all constraints is called the utopian solution, $\boldsymbol{x}_0$.

$$\boldsymbol{x}_0\ \in \Omega :\ \forall\ \boldsymbol{x}\ \in \Omega, f_j(\boldsymbol{x}_0) \leq f_j(\boldsymbol{x}) \tag{35}$$

$$\text{for } i = \{1,2, \cdots, N\}$$

This becomes a single objective problem for N=1, and equates to the global optimum which always exists although it may not be possible or difficult to determine. However, for multiobjective problems, N>1, the problem is ill-posed if the individual objective functions, $f_j$, are conflicting, as is typical in real world problems. In the scenarios where solutions represent compromises and trade offs, we identify an objective vector $z$ as optimal if none of its components, $f_j$, can be improved without degrading at least one of the others.

Italian economist Vilfredo Pareto devised a method to compare multiobjective optimization problems. While originally developed to describe allocation of economic resources, Pareto's concepts are also appropriate to describe engineering problems [82,83]. Given two sets of objective vectors, $z$ and $\boldsymbol{y}$, in a minimizer problem if every objective $f_j^z(\boldsymbol{x})$ contained in $z$ is less than or equal to every corresponding objective

$f_j^y(\boldsymbol{x})$ in $\boldsymbol{y}$, and at least one objective $f_j^z(\boldsymbol{x})$ is actually less than its corresponding $f_j^y(\boldsymbol{x})$ then $z$ dominates $\boldsymbol{y}$ denoted by $\boldsymbol{z} \succ \boldsymbol{y}$. If there is no other objective vector $\boldsymbol{y}$ that dominates $z$, then we describe the objective vector $z$ as Pareto optimal. The set of decision vectors $\boldsymbol{x}$ corresponding to this Pareto optimal objective vector as called the Pareto set, and the set all Pareto optimal objective vectors $z$ is the Pareto Front.

Multiobjective optimization clearly applies to the BCI classifier problem primarily in three areas.

- In the most general view, the real world SIBCI classifier problem is a very complex and highly ill-posed multiobjective optimization problem separating 4 tasks in feature space. The space for each task is composed of samples from multiple subjects each exhibiting different degrees of intersubject and intrasubject variance. The objectives consist of achieving the best average overall classification accuracy across the multiple subject population while disregarding the impact of artifacts at the expense of individual subject accuracy performance. The balance between conflicting objectives is also complicated by environmental uncertainties and inconsistencies. BCI is wrought with imprecisions from the lack of an error free and deterministic real world environment.
- The feature selection algorithm cost function is a balance two objectives, the best overall accuracy and degree of degeneracy in the classifier behavior.
- The support vector machine is a supervised learning algorithm whose implementation requires the solution of a multi-objective optimization. The problem can be stated as the determination of the location of hyperplanes that

minimize the classification error while maintaining maximum margins, subject to parameters that determine the softness of the margins.

While the concept of Pareto dominance is at first a seemingly attractive characterization of the BCI problem, the there are some pitfalls. Addressing the first point, the lack of deterministic objective functions and constraints keep us from applying the principles in all but the most abstract terms. Pareto dominance based classification methods are more often found in the evaluation of population-based in contrast to solution-based problems. Also, Pareto methods become less accurate as the number of constraints and problem dimensionality increases, thus making it unfeasible as an overall solution for this particular BCI implementation.

To the second point, our intention is to move the Pareto front to a more favorable section of the two dimensional accuracy-imbalance space. Without applying the Q-factor correction, in general, SIBCI solutions produce Pareto fronts which extend horizontally across the upper half of the space. In the worst cases, these uncorrected solutions produce a greater number of points in the upper right most quadrant of the space. The ideal location of the Pareto front in this two dimensional space is along the left vertical and lower horizontal axes. However we will consider any outcome that removes solutions from the upper right quadrant of the space to be successful, especially if the greatest concentration of solutions is placed in the upper left quadrant. The acceptable loci of Pareto fronts can be loosely described as the area under a diagonal connecting the upper left corner to the lower right corner, with the best solutions placed along the left and bottom axes of the space.

To the third point, the internal optimization algorithms of the SVMs used in this BCI implementation are well outside of the scope of this dissertation. Our research was confined only to the adjustment of parameters of the SVM and not to evaluate the performance of different implementation methods. The concept was presented solely for completeness of the discussion.

# CHAPTER 3

# EXPERIMENTAL DESIGN

A four-class SIBCI was implemented using scalp level EEG signals from 11 wet electrodes from the pre-frontal, pre-motor and motor cortex regions of the brain. Electrode locations Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4 and Cz from the pre-frontal and pre-motor cortex regions were included because of their role in planning of imagined motor tasks [32]. Our target classes consisted of left hand, right hand, left foot, and right foot imagined movements based on our own research [118,120] and that found in the literature [8,14,28,47,69,95]. The system stages include pre-processing, feature extraction, feature selection, and classification as shown in Figure 1.

## 3.1. Subject Protocol

EEG signals were collected from ten untrained volunteers under the authority of UMKC IRB# 090218. The students were chosen at random from the student body population of the university and consisted of 8 males and 2 females ranging in age from 20 – 32 years of age. The students were chosen from a variety of cultural and language backgrounds however all students spoke English and all instruction prompts were provided in English language. The signals were collected using a non-invasive 24 channel Electro-Cap shown in Figure 7 and configured according to the 10-20 electrode standard [64] depicted in Figure 8. These microvolt level signals were amplified by a NeuroPulse Systems MS-24R multichannel bioamplifier and 1.5 – 34 Hz bandpass filter with a 256 Hz sampling rate.

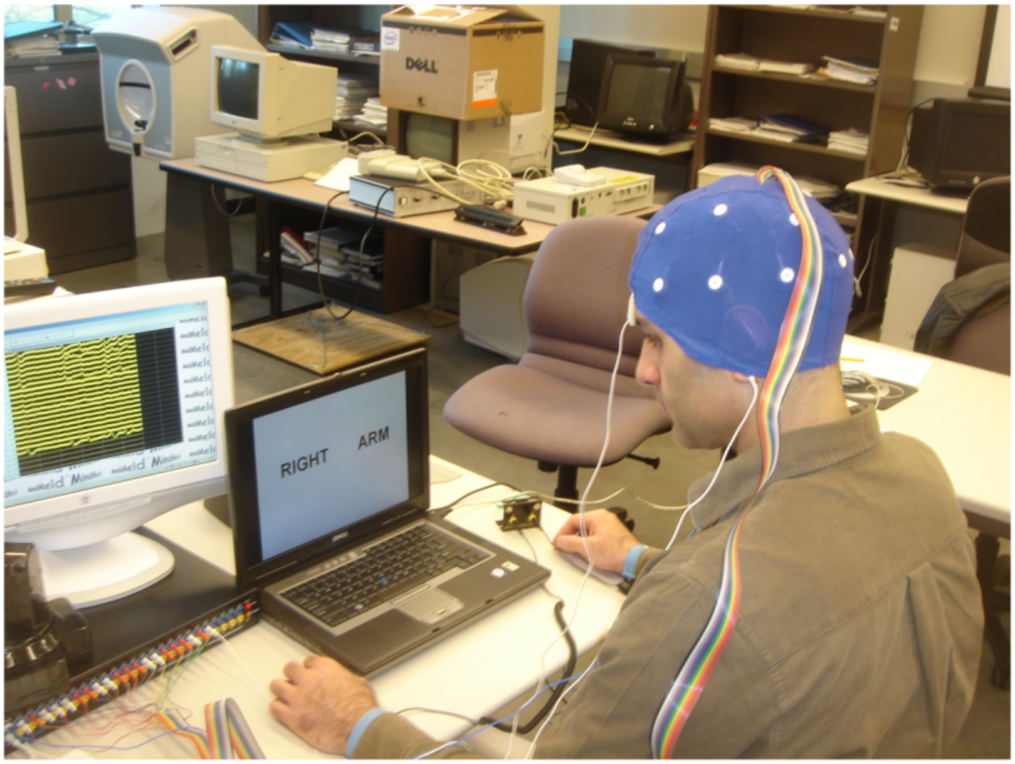


Figure 7. BCI operator wearing 24 channel EEG cap

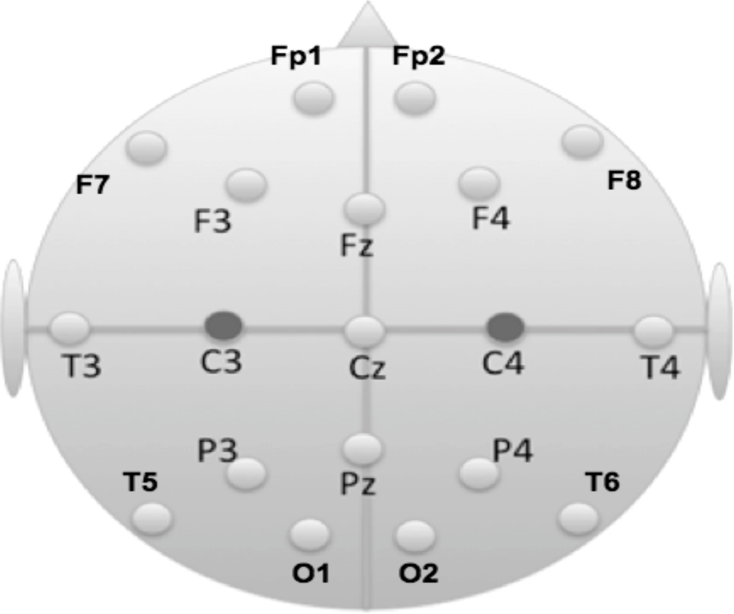


Figure 8. 10-20 International Electrode Placement System

Each volunteer was asked to perform four incremental or quasi-movements [85,118,120] without rehearsal for 8 seconds. Quasi-movements were chosen to simplify the imagination since none of our subjects had prior BCI experience. Two identical sessions were conducted 4-6 weeks apart, each consisting of 12 repetitions of the four intended movements. As a result, 960 8-second epochs were collected, each epoch consisting of the performance of one quasi-motor task. The first and last repetitions of each session were discarded to reduce the impact of fatigue, habituation and stress of performance, leaving 800 epochs for processing and feature extraction. Of these 800 epochs, 200 were left out for testing and 400 were used for training and 200 for cross validation in accordance with a 3-fold validation.

Each volunteer was visually prompted to produce a restrained hand or foot movement while gripping a rubber ball in each hand or pressing the appropriate foot into a rubber floor mat. The initiation of each visual prompt was accompanied by a short 500 Hz aural alert cue beep. The volunteers were instructed not to blink prior to and during the performance of the quasi-movement. A 10 second relaxation period was provided between each series of movements. Only the first 2 seconds of each movement event were analyzed. The protocol timing is shown in Figure 9.

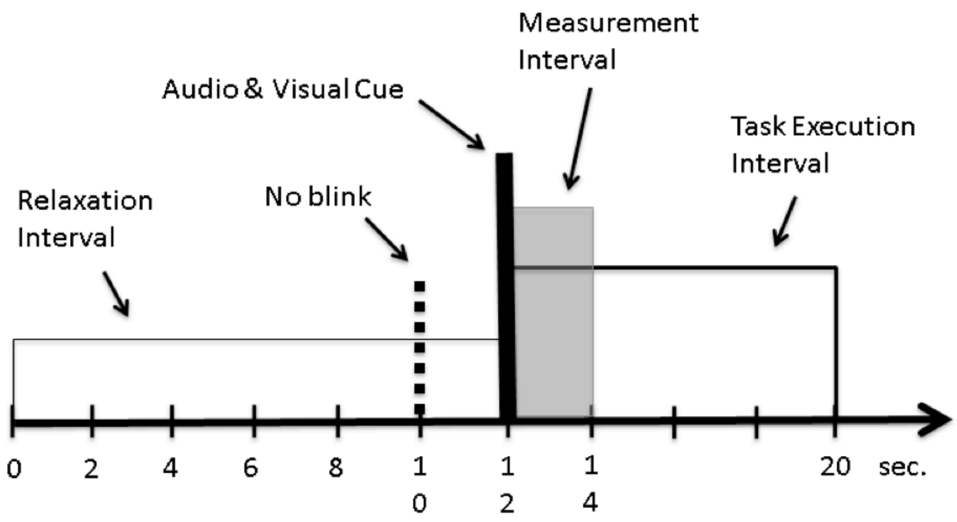


Figure 9. Recording timeline for SIBCI protocol

### 3.2. Feature Description

The pre-processing stage consists of collecting, filtering and segmenting the EEG signals into epochs. Common spatial filters were derived and applied to provide an alternate set of signals. The CSP signals were set aside and submitted to feature extraction in separate batches to produce an alternate set of features. Eight feature modalities were then extracted to the non-CSP (and later to the CSP) signals according to the parameters described below.

**3.2.1. Linear Predictive Coefficients.** Linear predictive coding (LPC) filter coefficients represent the time domain behavior of the EEG signals and have been successfully used in BCI applications [7]. The principle is to select a set of parameters that describes the electrical signals.

The LPC coefficients are derived from a forward linear predictor by minimizing the prediction error in the least squares sense. Specifically, the method uses the autocorrelation of autoregressive (AR) modeling to determine the filter coefficients. The Levinson-Durbin algorithm solves for an all pole infinite impulse response filter according to the autocorrelation sequence determined by the input signal. The transfer function H(z) of the filter is described by [105]

$$H(z) = \frac{1}{1+a_1(2)z^{-1}+\cdots+a_p(n+1)z^{-p}} \tag{36}$$

Our coefficients were chosen as the values $a_k$ where

$$\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k) \tag{37}$$

where $\hat{x}$ depicts the synthesized EEG signal, n represents the normalized discrete sample time, and k describes the linear parameters with maximum value p indicating the order of the model. LPC, AR or adaptive autoregressive (non-stationary) AAR models have been applied to EEG-based brain computer interfacing. As early as 1998, AAR models were used by Graz researchers to separate motor imagery from EEG signals [96].

After applying a downsampling of 3 to the signal, 18th order LPC coefficients were generated and used as features. The coefficient order was corrected for downsampling. Our investigation revealed a decrease in the classification error as the order increased to p = 58, then an increase in errors with a second minima at p=64 [118]. Beyond p=64 the errors increased. This increase in error for p > 64 is attributed to bias variance dilemma where the classifier is training according to noise in the training samples.

**3.2.2. Power Spectral Density.** Power spectral density, or some variation of it, is decidedly the most common source for EEG features found in BCI. As a result, many variations have been tested and evaluated since the earliest implementations of BCI. Dietsch first used Fourier analysis on EEG recordings prior to 1932 [123]. With the development of the Fast Fourier Transform in 1965 by Cooley and Tukey, [26] the computational complexity of discrete Fourier transform calculations was reduced to a level compatible with the current computing technology of that time. For a N samples of a discrete stationary EEG signal and real-valued $x(n) = \{x_0, x_2, \cdots x_{N-1}\}$, we assume that all values of x(n) are zero outside for n<0 and n > N-1. The autocorrelation function is given by $r_{xx}(k) = E[x_{n+k}x_n]$ where $r_{xx}(k) = r_{xx}(-k)$ which can be estimated from [105]

$$\hat{r}_{xx}(k) = \frac{1}{N}\sum_{n=0}^{N-1-k} x(n+k)x(n) = \frac{1}{N}x(k) * x(-k) \tag{38}$$

for lag values of $k$. This expression shows that the autocorrelation estimate can be viewed as a signal convolved with a time-reversed version of itself.

The Wiener-Khinchine theorem relates the power spectral density $S_{xx}(f)$ of a stationary random process to the autocorrelation function through the Fourier transform, stated in the discrete case by [99]

$$S_{xx}(f) = \sum_{k=-\infty}^{k=\infty} r_{xx}(k)e^{-j\omega k} \tag{39}$$

The Fourier transform of the autocorrelation function yields the power spectral density function. Using the estimated autocorrelation function in this Fourier transform, we obtain an estimate of the power spectral density or periodogram.

$$\hat{S}_{xx}(f) = \Im\{r_{xx}(k)\} = \Im\left\{\frac{1}{N}x(k) * x(-k)\right\} = F(s)F(-s)$$

$$= F(s)F^*(s) = |F(s)|^2 = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n)e^{-j\omega n}\right|^2 \tag{40}$$

There are several weaknesses in the periodogram method even when x(n) is stationary. The mean value of the periodogram can be shown to be [105]

$$E\left[\hat{S}_{xx}(f)\right] = \frac{1}{2\pi}S_{xx}(f) * \frac{1}{N}\frac{sin^2(\omega N/2)}{sin^2(\omega/2)} \tag{41}$$

The resulting expression results from the multiplication of the time domain signal by a window function. The effects of this term are smearing of the estimated spectrum making the periodogram unable to resolve fine details of the spectrum and leakage where energy from spectrum components is translated to other frequencies. Smearing and leakage can mask lower amplitude components and distort the location of narrowband peaks [105].

For a Gaussian process *x(n)*the variance of $\hat{S}_{xx}(f)$ [99] can is given as

$$Var\left[\hat{S}_{xx}(f)\right] \approx \hat{S}_{xx}^{2}(f)\left[1+\left(\frac{sin(\omega N)}{Nsin(\omega)}\right)^{2}\right] \tag{41}$$

where the variance does not converge to zero as the number of samples increases, resulting in undesirable estimation behavior. The leakage, smearing and estimation deficiencies are addressed by using estimation methods other than the periodogram. The Welch method applies windowing and averaging to address this.

Leakage effects produced by sidelobes in the $\frac{1}{N}\frac{sin^{2}(\omega N/2)}{sin^{2}(\omega/2)}$ term can be reduced by replacing the Bartlett window function of the basic periodogram with an alternative shape. In 1967 Peter D. Welch presented a method of spectral estimation which involves sectioning the data, taking modified periodograms of each section and averaging the modified periodograms. For $k$ segments of length L, possibly overlapping [134] and a data window W(j), j=0,…,L-1 we obtain k Fourier transforms $A_k$

$$A_{k}(n)=\frac{1}{L}\sum_{j=0}^{L-1}x_{k}(j)W(j)e^{-2kijn/L} \tag{42}$$

and $i=(-1)^{1/2}$ . The k periodograms are averaged to obtain the Welch spectral estimate $\hat{P}(f)$

$$\hat{P}(f)=\frac{1}{k}\sum_{k=1}^{k}\frac{L}{U}|A_{k}(n)|^{2} \text{ where } U=\frac{1}{L}\sum_{j=0}^{L-1}W^{2}(j) \tag{43}$$

and the mean of the estimates from the Welch method, for Gaussian processes are

$$E\left[\hat{P}(f)\right]=\frac{1}{2\pi LU}P(f)*|W(f)|^{2} \tag{44}$$

and the variance cannot be easily calculated in closed form for high amounts of overlap due to correlation effects, but for has been shown to be upper bounded by the expression [134]

$$\left[\hat{P}(f)\right]^2 \frac{L}{N} \leq Var\left[\hat{P}(f)\right] \leq \left[\hat{P}(f)\right]^2 \left[1 + \left(\frac{sin(\omega N)}{Nsin(\omega)}\right)^2\right] \tag{45}$$

The Hamming window, named for Richard Hamming, was selected for W(j). The Hamming window can be described as a Hann window (frequently referred to incorrectly as the Hanning window), sitting on top of a rectangular or Dirichlet window and is also known as a raised cosine window. The Hann window was named for Julius von Hann [45]. The endpoints of the Hann window are zero, the Hamming window has non-zero endpoints. The coefficients of the Hamming window are calculated by [90]

$$W(j) = 0.54 - 0.46\cos\left(2\pi \frac{j}{N}\right), 0 \leq j \leq N \tag{46}$$

This window has a discontinuity at the boundaries resulting in smearing and leakage effects however the raised cosine portion is designed to reduce the amplitude of the nearest sidelobes to the main lobe, improving selectivity performance of the filter in resolving spectral peaks. Peak resolution is the strongest advantage of the Hamming window.

Event-related desynchronization of neuronal signals as a result of imagined motor task EEG signals have been reported in the literature [92,105]. This desynchronization has been observed as changes in the amplitude of power spectral density (PSD) energy in the EEG signals.

Welch's periodogram method of power spectrum estimation was selected. Welch's method segments the time series into smaller sections and computes a periodogram estimate of each section. The estimates are averaged together to produce a result based on the estimate.

By using the Welch periodogram spectral method and Hamming window length 33 with 97% overlap, spectral amplitude coefficients were generated. The choice of window function, overlap and window length were selected by trial and error. The coefficients representing frequency components from 1 - 48 Hz were aggregated into twelve 4 Hz frequency bins. The spectral energy in each bin was calculated and the coefficients were used as PSD features.

**3.2.3. Cepstrum.** Coefficients of the non-linear Fourier Transform or Cepstrum were collected and used as features. Cepstrum was included for its possibility to detect non-linear features [90]. Cepstrum is commonly encountered in seismology, image processing and voice recognition applications, primarily for echo detection and removal. Bogert used cepstrum to detect time delayed echoes resulting from seismic waves propagating through geologic layers in the earth [16].

The real cepstrum transform does not preserve phase information. Cepstrum suffers from similar issues as power spectral density, such as windowing, aliasing and oversampling.

While cepstrum analysis is not a widely used method to derive BCI features, there have been a few recent implementations found in the literature. A recent cursor movement implementation in South Africa employed a favorable and encouraging comparison of cepstral and wavelet packet features using linear discriminant analysis and SVM classifiers for the subject specific paradigm [67]. Additional use of the cepstrum feature was done by Salvetti and Wilamowski [107].

The cepstrum coefficients are calculated from the inverse Fourier transform of the natural logarithm of the magnitude of the Fourier transform $X\left(e^{j\omega}\right)$ of signal x(t) [90]

$$c_x = \frac{1}{2\pi}\int_{-\pi}^{\pi} log\left|X\left(e^{j\omega}\right)\right| e^{j\omega n} d\omega \tag{47}$$

Cepstrum provides a method for representing convolution in the time domain as additive in the cepstrum frequency or quefrency domain.

For a signal input *x(t)* applied to a filter with a time domain impulse response of *h(t)* we obtain an output *y(t):*

$$y(t) = x(t) * h(t) \tag{48}$$

which can be described in the Fourier transform domain by

$$Y(\omega) = X(\omega)H(\omega) \tag{49}$$

After performing natural logarithm of the squared magnitude we obtain [90]

$$Y'(\omega) = log|Y(\omega)|^2 = log|X(\omega)H(\omega)|^2 = log|X(\omega)|^2 + log|H(\omega)|^2 \tag{50}$$

where $Y'(\omega)$represents the cepstrum feature vector. Otherwise, in signal reconstruction applications, each of the terms on the far right hand side can be inversely transformed to recover the terms

$$y'(t) = \ x'(t) + h'(t) \tag{51}$$

where either *x'(t)* or *h'(t)* can be estimated from the known source or filter in the problem and the unknown variable can be determined.

One of the disadvantages of the cepstral feature is the large number of coefficients that are produced. This may have discouraged its use as an EEG feature for BCI implementations. Since the objective of this research is to implement a wrapper feature selection, the length of the feature vector is not a detractor.

**3.2.4. Short Time Fourier Transform.** Unfortunately, none of the feature extraction methods presented thus far provide any information based on the time occurrence of spectral components. For non-stationary signals where the spectral composition is time dependent different techniques are required. Dennis Gabor introduced the short time Fourier transform (STFT) method in 1946 [43]. STFT is the simplest method of time-frequency analysis and is performed by dividing the time series data x(n) into segments in the time domain and Fourier transforming each segment independently. We can represent a continuous time short time Fourier transform STFT by the expression [105]

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau)W(\tau - t)e^{-jf\tau}d\tau \tag{52}$$

$W(\tau - t)$ represents a time shifted window which determines the resolution in time for the STFT algorithm. The discrete time STFT can be inferred from the continuous time model for frequency $f$ and discrete time m, and x(n) is a block of length 2N over the interval from –N to N-1 [99]

$$X(m,f) = \sum_{n=-N}^{N-1} w(n)x(n+m)e^{-j2\pi fn} \tag{53}$$

The spectrogram is analogous to power spectral density and can be obtained from the STFT according to $S_x(t,f) = |X(t,f)|^2$. STFT's are used to capture the non-stationary properties of signals. The STFT can be used to isolate short intervals of signal and such as is required in speech processing or as exhibited in the time-frequency behavior of EEG signals [105].

The disadvantage of the STFT is that the time and frequency product or resolution is constant across all frequencies and times of interest for the signal under analysis. We have to use the same frequency resolution for low frequency signals as for high frequencies. We also have to use the same time resolution for slowly varying signals as for rapidly varying signal, this is inefficient and does not always permit the best observation of all features. This is attributed to the Heisenberg uncertainty principle which establishes a lower bound to the product of time variation $\sigma_t$ and frequency variation $\sigma_f$ so that it cannot become infinitesimally small.

$$\sigma_t \sigma_f \geq \frac{1}{4\pi} \tag{54}$$

This inequality also guarantees that an increase in the time resolution (small $\sigma_t$) is accompanied by a decrease in the frequency resolution (large $\sigma_f$). STFT produces a fixed $\sigma_t \sigma_f$ resolution window regardless of the frequency or time of interest.

The Heisenberg relationship also has implications for the previously discussed feature extraction algorithms (PSD and Cepstrum). When signals are oversampled at a high rate (small $\sigma_t$), the transforms produce lower resolution output in the frequency (or quefrency) domain. Features that reside within the spaces between the output coefficients are missed; also, higher frequency noise components can be mistakenly identified as features [68].

STFT can be used to localize features which were not detectable using the LPC, PSD or Cepstrum methods. Short time Fourier transform signals were extracted on the downsampled EEG signals using a sliding window length of 1 second and an 84 sample overlap. The signal amplitude coefficients with the frequency span of 1 – 48 Hz were aggregated into four 12 Hz frequency bins and the energy coefficients of each bin were used as features. These parameters were selected by trial and error.

**3.2.5. Wavelets.** Wavelet functions have been well established as an EEG feature extraction method. The history of wavelet development is rooted in Fourier analysis.

Well established since the early 19th century as a tool for analysis of the properties of linear time-invariant signals, Fourier analysis was not suited for the examination of transient or non-stationary phenomena. However, Fourier analysis established the idea of decomposition and reconstruction of signals into basis functions, an essential foundation for wavelet theory. Much of the development of wavelet, originated during the 20th century from multiple independent sources. Mathematician Alfred Haar who studied orthogonal systems of functions realized the first wavelets in 1909. He was responsible for the concept of wavelet families, scaling, and translation. In the 1930s, Paul Levy, a French physicist found that the functions devised by Haar performed better than Fourier basis functions in his investigation of the local regularity elements of Brownian motion [54]. During the 1930s, physicist Hermann Wigner was exploring the use of wavelet transformation in quantum mechanics. Jean Ville extended Wigner's work to determine the energy distribution of signals in the time frequency domain in terms of the Wigner-Ville transform:[74]

$$W(t,\omega) = \int_{-\infty}^{+\infty} f\left(t+\frac{\tau}{2}\right) f^*\left(t-\frac{\tau}{2}\right) e^{-i\omega\tau} d\tau \tag{55}$$

The quantum physics heritage of wavelets is evidenced in terminology as wavelet functions are also referred to as time-frequency atoms. Dennis Gabor, a Hungarian electrical engineer, credited with the invention of holography in 1947, was the first to use continuous frequency wavelets to localize communication signals into time and frequency components. Gabor also defined the windowed Fourier transform. Gabor proposed wavelets or logons (named for the Greek word for information) [43] as the basis functions or building blocks for speech signals. Speech compression research in the late 1970's ushered in the development of non-invertible filter banks and conjugate mirror filters by Croiser, Esteban and Galand. Smith, Barnwell, Vaidyanthan and Vetterli [130] established the recovery of electrical signals by inverse filtering. During the same 1980's time frame, geophysicists Grossmann and Morlet, [49] while using wavelet theory to study seismological phenomena and the propagation of sound waves, extended the use of wavelets to study non-stationary signals. Morlet specifically was processing transient patterns in backscattered seismic signals propagating through geological layers. The transients yielded information about the thickness and composition of the layers comprising the earth. Prior to Morlet's work the analytical tools available consisted of windowed Fourier analysis and Gabor wavelets, either of which introduced artifacts. The collaboration of Morlet and Grossmann's work resulted in the formalization of the continuous wavelet transform in 1984 [49]. Morlet and Grossman formalized the distinction between analysis and synthesis wavelets and established the first general definition of a wavelet in the space $\boldsymbol{L}^2(\mathbb{R})$ of finite energy functions [59].

Stephane Mallat and Ingrid Deaubechies achieved the extension of the continuous version of wavelet transforms to the digital signal processing area. Deaubechies, a Bell Labs scientist, invented smooth orthonormal wavelet bases with pre-determined regularity and compact support [31], including biorthogonal wavelet families as well as the theory of frames.

Because of the transient nature of EEG signals, particularly event related potentials (ERPs), they generally have time varying frequency composition. This non-stationary behavior is not amenable to modeling with periodic basis functions such as the sine or cosine. The use of wavelet methods improves on the time-frequency localization of features over those described earlier by STFT methods. The constant $\sigma_t\sigma_f$ product or time frequency resolution referred to in the STFT discussion is mitigated in the wavelet paradigm.

While our signals were digitally sampled, which necessitates the use of discrete wavelet methods, the basic understanding of discrete wavelet methods can easily be derived from continuous wavelet theory. We can represent a signal by a set of $k$ coefficients $w_k$ representing a series expansion of a family of basis functions $\varphi_k(t)$ by calculating the correlation or inner product between the original signal *x(t)* and the basis function as shown by the relationship in continuous time [105]

$$w_k = \int_{-\infty}^{+\infty} x(t)\, \varphi_k(t) dt = \langle x(t), \varphi_k(t)\rangle \tag{56}$$

The discrete version can be inferred using the Riemann sum of the integral for step size of 1/N with N equal to the sampling frequency

$$w_k = \sum_{n=0}^{N-1} x(n)\, \varphi_k(n) \tag{57}$$

Wavelet basis functions can be generated through the scaling and translation of a single function or “mother” wavelet $\psi_{s,\tau}(t)$

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) \tag{58}$$

where $s > 0$, and represents the compactness of the function. Smaller values of $s$ shrink or dilate the waveform resulting in a higher frequency composition. Translation is determined by $\tau$. The continuous one-dimensional wavelet transform can be derived from these equations as [109]

$$w(s,\tau) = \int_{-\infty}^{+\infty} x(t) \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) dt = \langle x(t), \psi(t)\rangle \tag{59}$$

which is linear and invariant under translation and dilation. When the Fourier spectrum $\Psi(s)$ of the real valued function $\psi(t)$ satisfies the admissibility criterion [21]

$$C_{\psi} = \int_{-\infty}^{+\infty} \frac{|\Psi(s)|^2}{|s|} ds < \infty, \qquad (60)$$

this condition can be otherwise stated as constraining the wavelet expansion to strong convergence in the Hilbert space denoted by $\mathbb{L}^2$ or convergence in energy. This property is significant as it indicates filters that are well behaved in the iteration process [21,122, 130].

The admissibility criterion requires both that $\Psi(\infty) = 0$ and also that $\Psi(s)$ meets the zero mean condition

$$\Psi(0) = \int_{-\infty}^{+\infty} \psi(x)dx = 0, \qquad (61)$$

then the amplitude spectrum of the wavelet resembles the transfer function of a bandpass filter [21]. This condition reveals that any bandpass filter impulse response that rapidly decays to zero with increasing frequency can serve as a basic continuous wavelet. As a result, this permits us to implement wavelet transforms in terms of banks of linear convolution filters [122].

Calderon, Morlet and Grossman have shown that when the admissibility criteria is met the original function *x(t)* may be obtained through the reconstruction formula [49]

$$x(t) = \frac{1}{C_{\psi}} \int_{0}^{\infty} \int_{-\infty}^{+\infty} w(s,\tau)\, \psi_{s,\tau}(t) d\tau \frac{ds}{s^2} \qquad (62)$$

For a basic wavelet function of shift $\tau = 0$, we can define the reflected complex conjugate as

$$\breve{\psi}(t) = \psi^{*}(-t) = \frac{1}{\sqrt{s}} \psi^{*}\left(-\frac{t}{s}\right) \qquad (63)$$

where

$$\breve{\psi}(t) = \psi(t) \qquad (64)$$

for all real and even functions $\psi(t)$. Using this relationship we can restate the continuous wavelet transform as [109]

$$w(s,\tau) = \int_{-\infty}^{+\infty} x(t)\,\psi^*(\tau - t)dt = x(t) * \psi^*(\tau - t) \tag{65}$$

where for any given value of scale *s*, $w(s,\tau)$ represents the convolution of *x(t)* with the wavelet at the given scale value.

By substituting the reflected complex conjugate into the reconstruction formula we obtain Calderon's identity [49]

$$x(t) = \frac{1}{C_\psi}\int_0^{\infty}\int_{-\infty}^{+\infty} [x(t) * \psi^*(\tau - t)]\,\psi(\tau - t)d\tau\frac{ds}{s^2}$$

$$= \frac{1}{C_\psi}\int_0^{\infty}[x * \psi^* * \psi]\,(t)\,\frac{ds}{s^2} \tag{66}$$

This relationship consistently shows that we can apply the wavelet filters to reconstruct *x(t)*.

The continuous wavelet transform contains redundant information and transforms the original one dimensional function *x(t)* into a two dimensional function $w(s,\tau)$. Because of the computational inefficiency and redundancy it is desirable to convert the parameters to discrete form.

The derivation and implementation of the discrete wavelet transform follows along the same lines as the CWT, only the notable differences are presented here.

The wavelet function can be discretized through dyadic sampling of the translation and dilation parameters, where k and j are integers

$$s = 2^{-j} \tag{67}$$

$$\tau = k2^{-j} \tag{68}$$

obtaining the discretized wavelet function [21,108]

$$\psi_{j,k} = 2^{j/2}\psi\left(2^{j}n - k\right) \tag{69}$$

when substituted in the continuous wavelet transform, yields the discrete wavelet transform for the dyadic wavelet

$$w_{j,k} = \sum_{n=0}^{N-1} x(n)\, 2^{j/2}\psi\left(2^{j/2}n - k\right) \tag{70}$$

and the wavelet series expansion or inverse DWT is described by

$$x(n) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} w_{j,k}\psi_{j,k}(n) \tag{71}$$

The advantage of wavelets in multiresolution analysis can be seen from the scaling capability. The wavelets can be expanded or contracted to produce a complete set of basis functions which can simultaneously provide a coarse and a detailed view of the underlying signal, as well as multiple views for various degrees of resolution. As a finer resolution detail is obtained we can also achieve a better approximation or reconstruction of the signal. If we represent a low resolution (or scale) approximation of signal *x(t)* by $x_j(t)$, where *j* represents a scale level and *j* +1 indicates the next higher or more detailed resolution level. If the incremental detail is described by [105] as

$$y_j(t) = x_{j+1}(t) + x_j(t) \tag{72}$$

where the complete signal *x(t)* can be represented by the coarse part and the sum of the additional details

$$x(t) = x_j(t) + \sum_{k=j}^{\infty} y_k(t) \tag{73}$$

If *x(t)* belongs to the space $L^2(\mathbb{R})$ and $x_j(t)$ is contained in scaling or approximation subspace $\boldsymbol{V}_j$ then we have [122]

$$\boldsymbol{V}_0 \subset \boldsymbol{V}_1 \cdots \subset \boldsymbol{V}_{j-1} \subset \boldsymbol{V}_j \subset \boldsymbol{V}_{j+1} \cdots \subset \boldsymbol{L}^2(\mathbb{R}) \tag{74}$$

where *x(t)* exists in the entire space and the sequence of spaces meets the condition of completeness such that $x_j(t) \rightarrow x(t)$ as $j \rightarrow \infty$. This notation used for this relationship appears differently in several sources as Mallat and Debauchies define the space notation in descending order, ie. [31,74]

$$\cdots \subset \boldsymbol{V}_{j+1} \subset \boldsymbol{V}_j \subset \boldsymbol{V}_{j-1} \subset \cdots \tag{75}$$

in this dissertation we adopt the ascending notation order as used by Strang. A wavelet subspace $\boldsymbol{W}_j$ describes the difference between the consecutive approximation subspaces $\boldsymbol{V}_j, \boldsymbol{V}_{j+1}$ [122].

$$\boldsymbol{V}_{j+1} = \boldsymbol{V}_j \oplus \boldsymbol{W}_j \tag{76}$$

In order to take advantage of multiresolution capabilities the scaling function which establishes the relationship between the compression or resolution of the wavelet and the time shift parameter. Scaling functions consist of an orthonormal set of functions, which span the corresponding subspaces $\boldsymbol{V}_j$ in the space $L^2(\mathbb{R})$. Similarly, the set of wavelet functions span the corresponding subspaces $\boldsymbol{W}_j$ in the space $L^2(\mathbb{R})$. In multiresolution, wavelet decomposition we are essentially carving up the lower resolution $V_j$ subspace in order to construct lower order detail subspaces $\boldsymbol{W}_0, \cdots, \boldsymbol{W}_{j-1}$ and more importantly to derive new wavelet bases from them.

In summary, the relationship between these functions can be described by the top-down algorithm in that once we have a discrete low pass filter impulse response $h_0$ and the scaling function $\varphi$, we can then determine the discrete high-pass impulse response $h_1$or wavelet vector [21]

$$h_1 = (-1)^k h_0(-k+1) \tag{77}$$

where the basic wavelet is derived from

$$\psi(t) = \sum_k h_1(k)\, \varphi(2t - k) \tag{78}$$

and an orthonormal wavelet set is produced.

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right) \tag{79}$$

A filter bank interpretation for the wavelet time-frequency decomposition follows from predecessor frequency spectrum analysis. In the purely frequency domain the Fourier series coefficients can be portrayed as the output magnitudes from a bank of 1 Hz wide bandpass filters. This concept can be extended to wavelet basic functions. The discrete representation of the reflected complex conjugate wavelet is specified by a pair of conjugate mirror filters with an impulse responses related by [130]

$$h_1[n] = (-1)^{1-n} h_0[1 - n] \tag{80}$$

This relationship assures that the impulse response times of $h_0$ and $h_1$will be comparable. This is also represented by the transfer function relationship where $H_0$ represents the low pass transfer function and $H_1$ represents the high pass transfer function for a simple two level decomposition.

$$H_1^2(s) = 1 - H_0^2(s) \tag{81}$$

Iteratively we can perform successive decompositions for example, so that the low pass signal is broken down into many lower resolution components. Structurally this represents a wavelet decomposition tree and can be continued until the number of samples *i* produced by each individual filter outputs consists of a single sample, *i*=1. A signal *x(n)* of length N can be expanded in $j \leq 2^{N/2}$ ways, where *j* indicates the number of filter banks [74]. For large numbers N of sample *x(n)* the size of the decomposition tree and the number of wavelet coefficients can also be large so methods are introduced to produce a lower dimension metric which can also be useful as a wavelet feature. The energy $E2(w)$ or 2-norm of wavelet expansion coefficients $s_{i,j}$ produced at each of the filter bank outputs reduces to a single scalar value.

$$E2_j(s) = \sum_i |s_{i,j}|^2 = \|s_{i,j}\|_2^2 \tag{82}$$

This expression reflects the energies marginalized over shift or time.

The method of iterating the low-pass filter by decomposing each approximation space assumes that the lower frequencies within our range of interest contain more information than higher frequencies. The wavelet packet method is a generalization of the wavelet decomposition described above. Instead of dividing only the approximation spaces $\boldsymbol{V}_j$ to produce detail spaces and wavelet bases we can set $\boldsymbol{U}_j = \boldsymbol{W}_j$ divide these detail spaces to derive new wavelet bases. Using a binary tree where nodes indicate a space $\boldsymbol{W}_j^p$ where j represents the depth of the node and p represents the number of the number of nodes at the same level and to the left, the splitting of the spaces occurs according to

$$\boldsymbol{W}_{\boldsymbol{j}+1}^{2p} \oplus \boldsymbol{W}_{\boldsymbol{j}+1}^{2p+1} = \boldsymbol{W}_{\boldsymbol{j}}^{p} \tag{83}$$

resulting in a binary tree of packet wavelet spaces where each parent node is subdivided into a pair of orthogonal subspaces [74]. For a discrete signal of size N, the maximum depth of the wavelet packet tree is $\log_2$N.

Unfortunately smooth orthonormal functions that meet the conditions required for wavelet basis functions with compact support lack either symmetry or anti-symmetry properties. In general, the biorthogonality condition is defined for two basis functions $\boldsymbol{r_i}, \boldsymbol{v_j}$ by the inner product

$$\langle \boldsymbol{r_i}, \boldsymbol{v_j} \rangle = \delta(i-j). \tag{84}$$

For biorthogonal wavelets, we have two wavelet families $\psi$ and $\hat{\psi}$, one is used for analysis or decomposition and the other is used for synthesis or reconstruction. From Mallat [74]. Theorems 7.2 and 7.11, the two wavelets are related by the biorthogonality condition, that for any $(j, k, m, n) \in \mathbb{Z}^4$

$$\langle \psi_{j,m}, \hat{\psi}_{k,n} \rangle = \delta[m-n]\delta[j-k] \tag{85}$$

and for any $f \in \boldsymbol{L}^2(\mathbb{R})$ there are two possible decompositions

$$f = \sum_{n,j=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \hat{\psi}_{j,n} = \sum_{n,j=-\infty}^{+\infty} \langle f, \hat{\psi}_{j,n} \rangle \psi_{j,n}, \tag{86}$$

where either wavelet of the pair can be used for either analysis or synthesis. The filter bank implementation of biorthogonal wavelets requires twice as many discrete filter banks as for orthogonal filter banks. Two scaling vectors or lowpass filters with impulse responses $h_0(n)$ and $\hat{h}_0(n)$ meeting the biorthogonality condition

$$\langle h_0(n), \hat{h}_0(n) \rangle = \delta[0] \tag{87}$$

and additional wavelet vectors or bandpass filters are [21]

$$g[n] = (-1)^n h_0(1-n) \text{ and } \hat{g}[n] = (-1)^n \hat{h}_0(1-n). \tag{88}$$

The use of biorthogonal splines developed by Cohen, Daubechies and Feauveau have closed form solutions for the discrete filter impulse responses or scaling vectors described by [74]

$$h(\omega) = \sqrt{2} exp\left(\frac{-i\epsilon\omega}{2}\right)\left(cos\frac{\omega}{2}\right)^p \tag{89}$$

$p$ indicates the number of vanishing moments, and $\epsilon = 0$ for $p$ even and $\epsilon = 1$ for $p$ odd. The number of vanishing moments describes the local regularity of the wavelet. The number of vanishing moments comes from the Taylor series approximation order of the wavelet function indicating a relationship between the differentiability and the rate for the coefficients to decay to zero. The greater the number of vanishing moments the faster the delay of the wavelet transform $w(s,\tau)$ as described by Sheng [117], for $\tau = 0$, using a Taylor series expansion at t=0, where n represents the expansion order

$$w(s,0) = \frac{1}{\sqrt{s}}\left[\sum_{p=0}^{n} f^{(p)}(0)\int_{-\infty}^{+\infty}\frac{t^p}{p!}\psi\left(\frac{t}{s}\right)dt + O(n+1)\right] \tag{90}$$

the p^th moment $M_p$ calculated by

$$M_p = \int_{-\infty}^{+\infty} t^p \psi\left(\frac{t}{s}\right)dt \tag{91}$$

gives the expression

$$w(s,0) = \frac{1}{\sqrt{s}}\left[\sum_{k=1}^{n}\frac{f^{(k)}}{k!}M_k s^{(k+1)} + O(s^{n+2})\right] \tag{92}$$

From the admissibility condition $M_0 = 0$, hence the absence of the k=0 term in the summation term. From this expression if we define the first n moments as vanishing, then only the $O(s^{n+2})$ term remains, thus establishing the upper bound on the decay of the signal. By increasing the number of vanishing moments to k, all coefficients of order $s^{(k+1)}$ become zero or vanish. The wavelet function will only have terms of $s^{(k+2)}$ and greater, giving it the capability to represent all polynomials of degree less than or equal to k+1 in the scaling space $\boldsymbol{V_j}$. This is also referred to as the accuracy of the filter [123].

**3.2.7. Wavelet families.** The classification performance of biorthogonal (and reverse biorthogonal) spline, Morlet, Coiflet, symlet, Gaussian, Daubechies, Meyer wavelets and their variants were examined. Biorthogonal and reverse biorthogonal splines are compactly supported wavelets for which symmetry and exact reconstruction are possible with finite impulse response filters, this is not possible for orthogonal wavelets except for the Haar wavelet [81,123].

Symlets are compactly supported wavelets designed by Debauchies [31] with the least symmetry and the highest number of vanishing moments for a given support width. The support width is 2N-1 and the filter length is 2N. Given the transfer function for a real, causal filter h[n] : [74]

$$H(\omega) = \sum_{n=0}^{N-1} h[n]e^{-in\omega} \tag{93}$$

which has a zero of order p at $\omega = \pi$ producing a wavelet with p vanishing moments and a polynomial

$$H(\omega) = \sqrt{2}\left(\frac{1+e^{-i\omega}}{2}\right)^{p} R\left(e^{-i\omega}\right) \tag{94}$$

where the square root of the polynomial $R\left(e^{-i\omega}\right)$ represents the phase term. By optimizing this square root polynomial to the most linear solution produces a symmetric wavelet with a support of [-p+1,p] with p vanishing moments.

Continuous wavelet transform methods have been used for feature extraction in BCI competition [18]. Wavelets were evaluated as one of the candidate features in the Thought Translation Device [55]. Subject independent BCI studies using autoregressive coefficients of stationary wavelet packet transforms have been reported in self-paced mental task BCI [35,125,133]. A wavelet packet entropy method was employed in BCI research by Zhiwei, Minfen [138] Graimann [47] also reported on the development of BCI for detection of ERP and ERD/ERS from electrocortiogram signals using wavelet packet analysis.

**3.2.7. Common Spatial Pattern (CSP).** Common Spatial Pattern Filtering (CSP) of the signals prior to feature extraction was used to generate additional sets of features. CSP is a data-driven supervised decomposition of the EEG signals. With CSP we maximize the variance of EEG signals from one task while minimizing the variance of signals from another task. As CSP is essentially a binary (two-task) method, we again use the OVR technique, which will be described below. Bandpower of bandpass filters is equivalent to the variance of the signals we can use this method to optimize our EEG signals with respect to event related desynchronization effects [101]. After performing CSP filtering we proceed to extract features as was done for non-CSP signals. CSP filtering for EEG signals has been used extensively in the extraction of BCI signals [13,15,48,73,101].

For our multiclass problem we projected each EEG signal epoch matrix $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{C}\times\boldsymbol{T}}$, (where C indicates the number of channels and T the number of samples) into a CSP decomposition $\boldsymbol{Z} \in \mathbb{R}^{\boldsymbol{C}\times\boldsymbol{T}}$, according to a projection matrix, $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{C}\times\boldsymbol{C}}$, consisting of the common spatial filters, according to

$$\mathbf{Z}\mathbf{=}\ \mathbf{W}^{\mathrm{T}}\ \mathbf{X} \tag{95}$$

The elements of $\mathbf{W}$, specifically each column vector $\boldsymbol{w_j} \in \mathbb{R}^{\boldsymbol{C}}$ where $j = 1, \cdots, \mathrm{C}$ , represents a spatial filter. A spatial pattern or mixing matrix $\mathbf{A}$ can be retrieved from the inverse of the spatial filters or demixing matrix $\mathbf{W}$, using $\boldsymbol{A} = (\boldsymbol{W}^{-1})^{\mathbf{T}} \in \mathbb{R}^{\boldsymbol{C}\times\boldsymbol{C}}$ where each column vector $\boldsymbol{a_j} \in \mathbb{R}^{\boldsymbol{C}}$ where $j = 1, \cdots, \mathrm{C}$ represents one pattern [15,101]. Our features were extracted from the elements of the CSP decomposed signals $\boldsymbol{Z}$**.**

A covariance matrix $\Sigma_k^{i,j}$ was calculated for each subject *i,* trials *j* and task *k* by from each of the data matrices $\mathrm{X}_k^{i,j}$ according to $\Sigma_k^{i,j} = \left(\mathrm{X}_k^{i,j}\right)^T \mathrm{X}_k^{i,j}$ . For each task $k \in [1, \cdots, 4]$, an average covariance matrix $\Sigma_k$ is calculated [43]

$$\Sigma_k = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \Sigma_k^{i,j} \tag{96}$$

where m represents the total number of subjects and n represents the number for trails recorded for each subject. A trial consists of one complete set of 4 tasks performed by one subject.

For a binary or two class problem, the CSP filter is produced by simultaneously diagonalizing the two covariance matrices $\Sigma_+, \Sigma_-$ so that [43,101]

$$W^T \Sigma_+ W = \Lambda_+ \tag{97}$$

$$W^T \Sigma_- W = \Lambda_- \tag{98}$$

where $\Lambda_{(+,-)}$ is the diagonal eigenvalue matrices and $\Lambda_+ + \Lambda_- = I$, where I is the identity matrix. To apply this method to a four task model we generate 4 sets of filters by setting $\Sigma_+ = \Sigma_k$ to the desired class $k \in [1, \cdots, 4]$ and setting $\Sigma_- = (\Sigma_1 + \Sigma_2 + \Sigma_3 + \Sigma_4)/4$.

While CSP has been commonly used in BCI to derive features based on the actual patterns, our method used the CSP filters to project the time sampled signals from the original sensor space **X** into comparable time sampled signals in a surrogate sensor space **Z**. The feature extraction and selection steps were then performed in the same manner as for the non-filtered signals.

### 3.3. Feature Selection

The objective of feature selection is to find a subset of features that leads to the smallest classification error [60,61]. To this end, we chose a Sequential Forward Floating Selection (SFFS) within our wrapper method. SFFS avoids the nesting problems of the Selective Forward method, where once a feature is chosen it cannot be discarded and in the Selective Backward algorithm, where a feature cannot be included in the final subset once it has been discarded [100,124]. SFFS re-evaluates previously discarded features for inclusion while previously selected features are reevaluated for discard. A detailed treatment of SFFS is provided by Theodoridis [124]. In general, we consider a set of *m* features and build the best subset *k* for $k = 1, 2, \ldots, l \leq m$ where a cost criterion C is optimized where $X_k = \{x_1, x_2, \ldots, x_k\}$ is the set of the best combination of features. $Y_{m-k}$ represents the remaining unselected $m - k$ features. Retaining all of the remaining lower dimension subsets $X_2, X_3, \ldots, X_{k-1}$ we build on the sets by creating the next subset $X_{k+1}$ by choosing an additional element from $Y_{m-k}$. The newly created features are compared with the original subsets and if the cost function is improved the new feature is retained. As the classifier outputs producing less favorable cost functions are less relevant as they are discarded. For a desired final N-dimension feature vector SFFS has a time complexity of *O(N)* [100].

If the array of cost functions $C_2, C_3, \ldots, C_{k+1}$ associated with $X_2, X_3, \ldots, X_{k+1}$ contains a most favorable cost function resulting from a degenerate classifier, the SFFS search algorithm may converge to an incorrect feature subset or fail to converge. As the number of classes increases to the point where the value of (M-1)/M% is greater than the probability of correct detection from Eq. (4) the chance of this failure to converge increases with M.

SFFS has been used in BCI for channel selection [52,53] and has been shown to be a very successful feature selection algorithm [40,58,100].

### 3.4. Support Vector Machine Classifier

Once the features have been extracted our objective is to separate the feature space by establishing boundaries between the regions occupied by each of the classes based on measureable criteria. In two dimensional feature space this boundary would be curves or lines establishing regional boundaries, in three dimensional feature space this boundary becomes a surfaces and in higher dimensional feature spaces the boundaries are hyperplanes. Given $x_i, i = 1,2,\cdots,N,$ where each $x_i$ is a feature vector of a training set of data, where each vector belongs to either class $\omega_1$ or $\omega_2$ and $y_i \in [-1,+1], i = 1,2,\cdots,N,$ are class labels indicating membership in either $\omega_1$ or $\omega_2$ [20]. For a set of feature vectors that can be perfectly separated by a hyperplane, the projection weights **w** and bias $b$ can be scaled to produce the condition for all data points

$$y_i(\boldsymbol{w^T} x_{\boldsymbol{i}} + b) \geq 1 \tag{99}$$

known as the canonical form of the hyperplane [124]. Where the point closest to the hyperplane meets the condition

$$y_i(\boldsymbol{w^T} x_{\boldsymbol{i}} + b) = 1 \tag{100}$$

where for this point the constraint is active. The distance between this closest point and the hyperplane is described by $\frac{1}{||w||}$ which can be maximized by solving the optimization problem [20]

$$\min_{w,b} \frac{1}{2} ||\boldsymbol{w}||^2 \tag{101}$$

subject to $y_i(\boldsymbol{w^T} x_{\boldsymbol{i}} + b) \geq 1,\ \ i = 1,2,\cdots,N.$

This formulation assumes that the data points in the feature space are linearly separable, in instances were this is not the case, it may not be possible to obtain an exact separation between the classes. There may be some outliers that are not able to be correctly separated in the feature space. This can occur in situations where the number of features is high, there is a limited time to perform a calculation, the distribution of the data is not known, there is a nonlinear map between the inputs and outputs or the convergence is not convex so the solution may fall into a local minimum [107].

This is resolved by the introduction of a slack variables $\xi_n \geq 0$ where n= 1,..., N for each training point. Correctly classified points will have $\xi_n = 0$, points on the decision boundary will have $\xi_n = 1$, points on the wrong side of the boundary will have $\xi_n > 1$, and points on the correct side of the boundary but inside the margin will have $1 > \xi_n > 0$. This condition can be handled by introducing the constraint [20]

$$y_i(\boldsymbol{w^T} x_{\boldsymbol{i}} + b) \geq 1 - \xi_n, \ \ i = 1,2, \cdots, N. \tag{102}$$

Since we now wish to maximize the margin while keeping the number of points for which $\xi_n \geq 0$ to a minimum we must now minimize the condition

$$\min_{w,b} C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\boldsymbol{w}\|^2 \tag{103}$$

where the C-parameter C > 0 controls the relative weight or trade-off between the two conflicting terms.

While the support vector machine is well suited for BCI applications which exhibit many of the characteristics described above one of the major limitations is the high computational requirements posed by quadratic programming solver techniques used to implement the SVM. Quadratic programming solvers typically take *$O(N^3)$* with memory requirements on the order of *$O(N^2)$*. This provides additional incentive, beyond bias-variance, to reduce the dimensionality of feature sets.

CHAPTER 4.

# DEGENERACY IN SUPPORT VECTOR MACHINES

## 4.1. Degeneracy Detection

In much of the classifier literature the terms random classifier and trivial classifier are used interchangeably. Within the scope of M-class discrete output classifiers where M=2, the distinction is insignificant as all trivial classifiers converge to the same solutions as we will demonstrate. For the purpose of analyzing classifiers where M > 2 it becomes important to draw a distinction. Figure 10 depicts the results of an ideal discrete output classifier such as an SVM for 2 classes of equal frequency. Figure 11 shows the one versus the rest multiclass version of the ideal classifier for 4 classes of equal frequency. An non-ideal non-trivial classifier is shown in in Figure 12 for 2 classes.

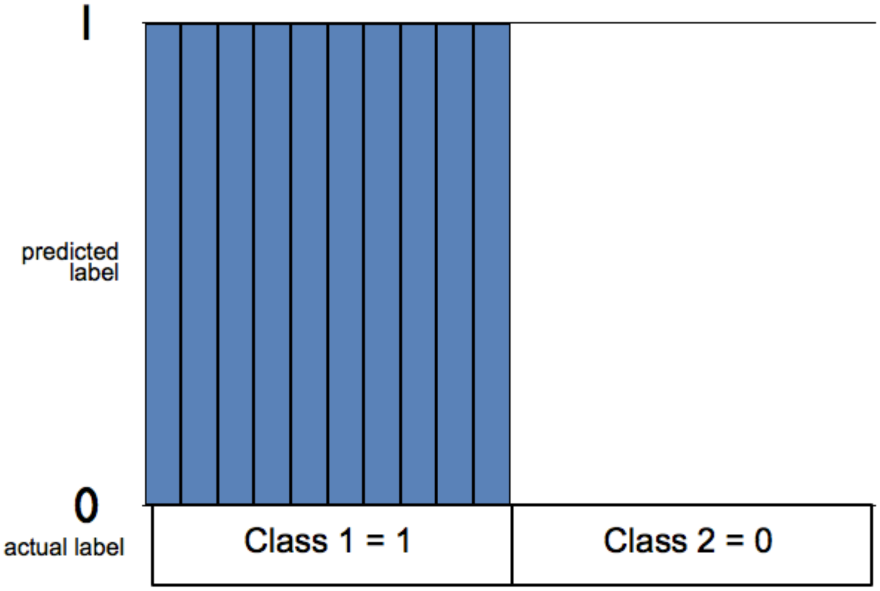


Figure 10. Ideal discrete Output Classifier (for 2 classes of equal frequency)

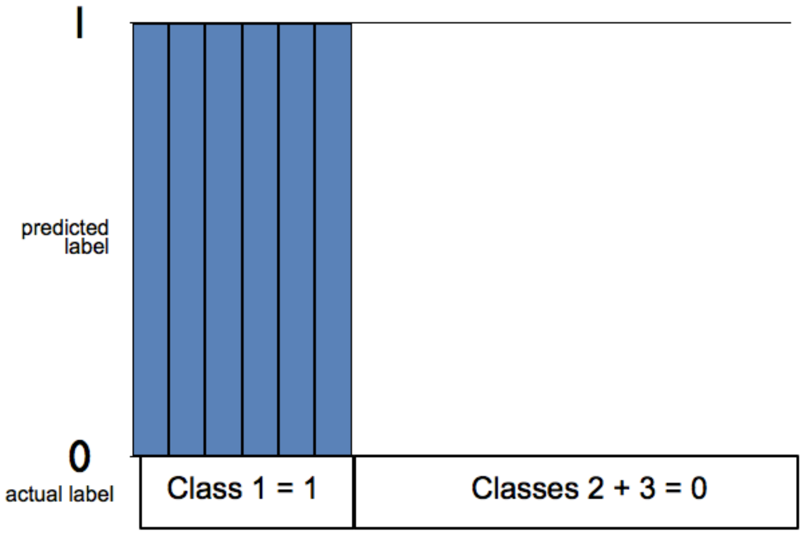


Figure 11. Ideal discrete Output Classifier (for 3 classes of equal frequency)

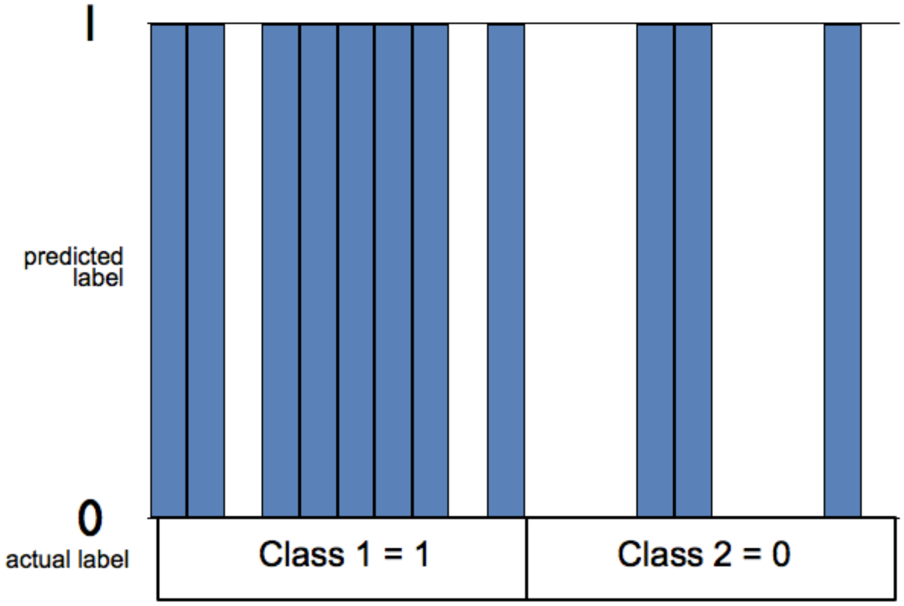


Figure 12. Non-ideal non-trivial discrete Output Classifier (for 2 classes of equal frequency)

It is useful and more accurate to categorize trivial classifiers into two distinct subsets. The random classifier is one subset of trivial classifiers. We identify a random classifier as one that actually makes distinctions between input signals but randomly distributes errors across the population to the point where the classifier is determined unreliable. The accuracy of the random M-class classifier output is indicated by ACC = 1/M. Figure 12 shows an example of a random classifier for M=2. The same concept can be extended to examples of M>2.

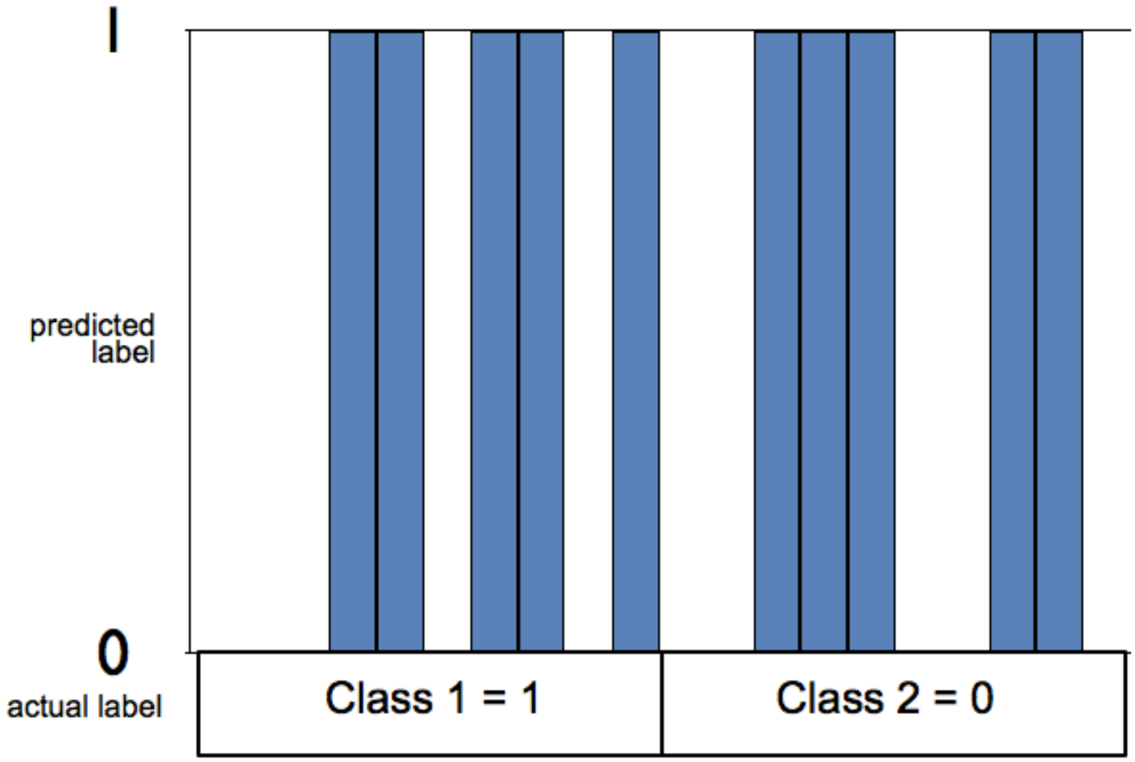


Figure 13. Random discrete Output Classifier

The second subset of trivial classifiers consists of degenerate classifiers and describes models that assign all input signals to either one class or the other regardless of the composition of the input data. This subset is the focus of our research topic as these classifiers become increasingly problematic as the number of classes M increases.

Within this subset of degenerate classifiers we will make a further distinction between those that indicate all positive (minority class) and those indicating all negative (majority class). Figure 14 shows the all negative degenerate classifier and Figure 15 shows the all positive degenerate version. These figures depict M=2 classifiers and as explained earlier, can be extended to cases where M>2. The majority or minority distinction only applies to cases where M>2 and there is an unequal number of samples between the two groups of classes. To overcome this we refer to the all negative as Mode I degeneracy and the all positive as Mode II, regardless of the value of M.

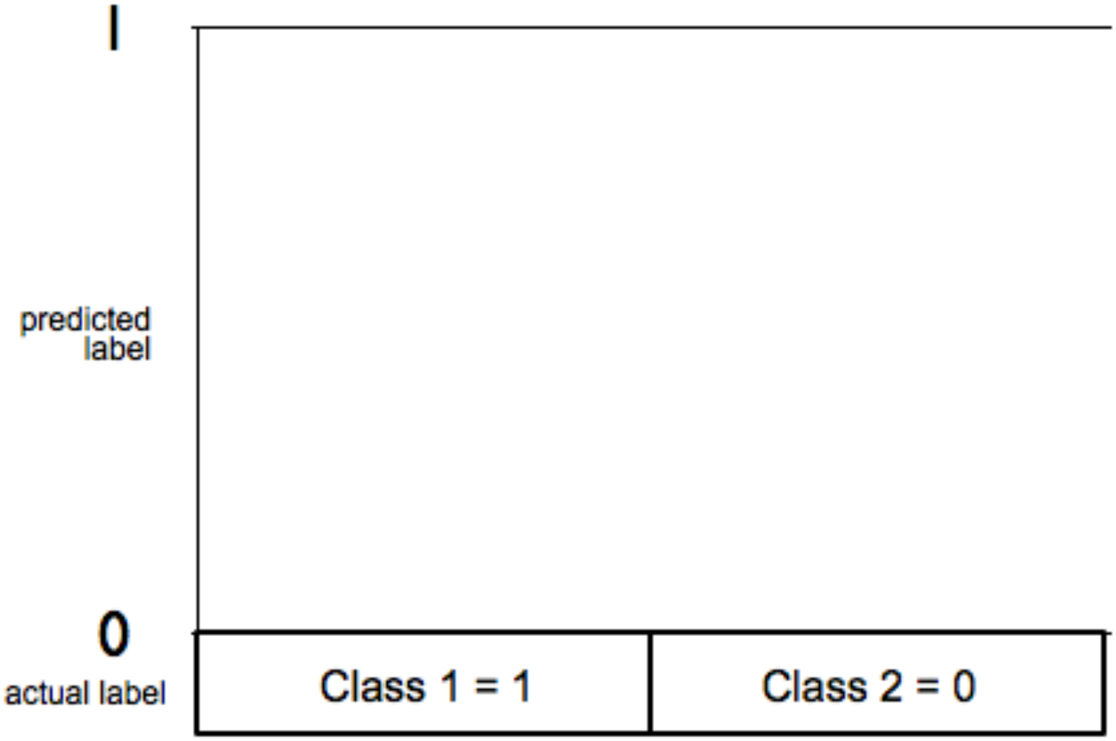


Figure 14. All Negative Degenerate Classifier (Mode I)

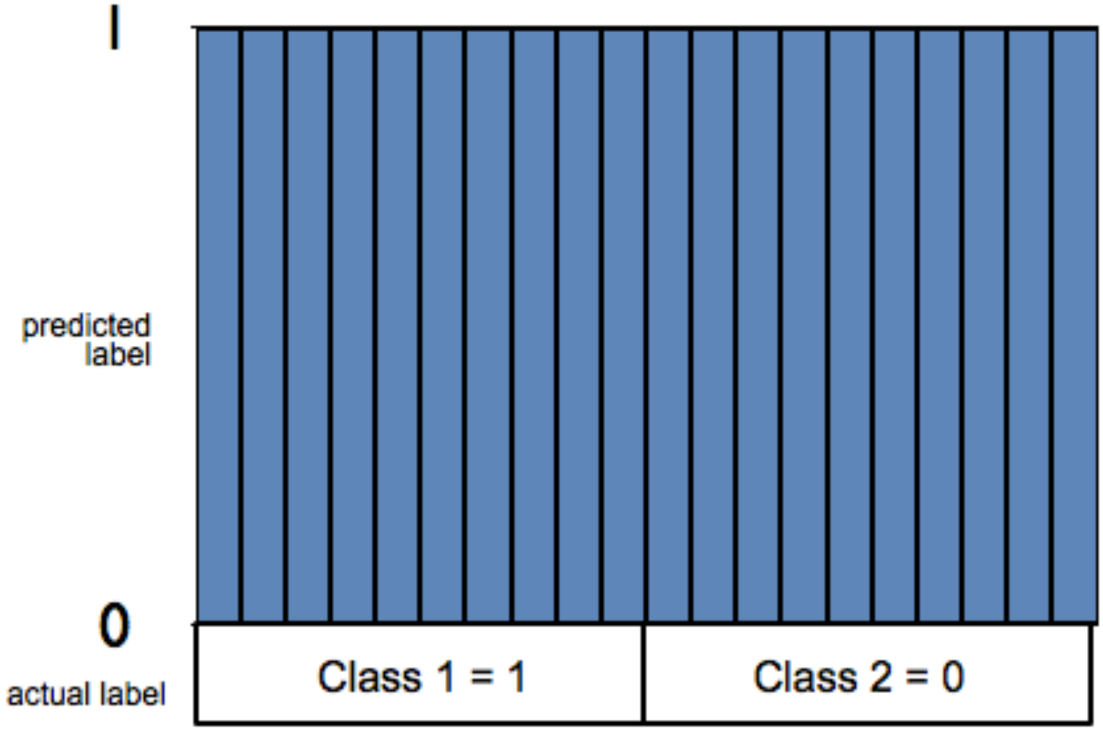


Figure 15. All positive degenerate classifier (Mode II)

Detection of degenerate and non-degenerate classifiers can be accomplished by using elements of the confusion matrices of the classifier output. Unfortunately the confusion matrix is not a scalar value and two confusion matrices cannot be easily compared. For this reason, a comparison of the Sensitivity and Specificity of the confusion matrix is valuable. The probability of error is not the only method for evaluating the performance of classifiers and in instances where the frequencies of the two classes are unequal it is unsatisfactory. For discrete output classifiers where ROC AUC methods cannot be applied, the confusion matrix is useful. The confusion matrix **A** = [*A( i , j )*] is defined as each element A( i , j ) is the number of data points whose true class label was *i* and was classified to class *j*. For a one versus the rest classifier where

$\mathbf{A} \in \mathbb{R}^{2\times 2}$ the approach requires a transformation of the confusion matrix into a scalar value. This transformation must preserve both the accuracy information and compensates for degeneracy in the solution. To achieve this we have devised the Q –factor which will be derived in the following two steps, first by converting the matrix to a vector quantity and then to a scalar as shown

$$\mathbb{Q}: \mathbb{R}^{2\times 2} \in \mathbb{R} \tag{104}$$

Describing the elements of $\mathbf{A}$, A(1,1) represents TP or the number of True Positive elements and A(2,2) represents TN, the number of True Negatives. A(1,2) indicates FP False Positives, and A(2,1) represents FN False Negative elements. From these elements we obtain accuracy (ACC), the precision of the first class or Sensitivity (SENS) and the precision of the second class or Specificity (SPEC),

$$ACC = \frac{1}{N}\sum_{i=1}^{K} A(i,i) = 1 - \text{probability of error} \tag{105}$$

$$SENS \ = \frac{A(1,1)}{\sum_{i=1}^{K} A(1,i)} = \frac{TP}{TP+FN} \tag{106}$$

$$SPEC \ = \frac{A(K,K)}{\sum_{i=1}^{K} A(1,K)} = \frac{TN}{TN+FP} \tag{107}$$

For every value of ACC, SENS and SPEC

$$0 \leq ACC, SENS, SPEC \leq 1.0 \tag{108}$$

for the OVR classifier $K$ =2 and N is the total number of data points. Summarizing the possible classifier states and performance metrics, we have

ACC = SPEC = SENS = 1.0 ideal classifier (Figs. 10, 11)

SENS/SPEC = 1

| | |
|---|---|
| $0 < \text{ACC} \approx \text{SPEC} \approx \text{SENS} < 1.0$<br>$\text{SENS/SPEC} \approx 1.0$ | non ideal classifier (Fig. 12) |
| $\text{SENS} = 0$; $\text{SPEC} = 1.0$; $\text{ACC} = (\text{M-1})/\text{M}$<br>$\text{SENS/SPEC} =0$ | Mode I degenerate classifier (Fig.14) |
| $\text{SENS} = 1.0$; $\text{SPEC} = 0$; $\text{ACC} = 1/\text{M}$<br>$\text{SPEC/SENS} = 0$ | Mode II degenerate classifier(Fig.15) |

M is equal to the total number of classes or tasks aggregated in the OVR classifier. We identify the trivial classifier that always selects instances as negative as Mode I degenerate. The trivial classifier that always selects instances as positive, we identify as Mode II degenerate.

Mode I is most detrimental to wrapper search convergence as we see that $\text{ACC} = \text{M-1/M} \geq 1/\text{M}$ for all possible values of M.

The SPEC/SENS ratio indicates the imbalance in the off-diagonal elements of confusion matrix **A** and indicates of the degeneracy in the classifier model (Fig. 16). Using the SPEC/SENS ratio provides a consistent measure of the degeneracy in the classifier independent of the number of classes and relative sizes of the majority and minority groups, unlike the similar statistic, the f-metric which is sensitive to the number of classes [132].

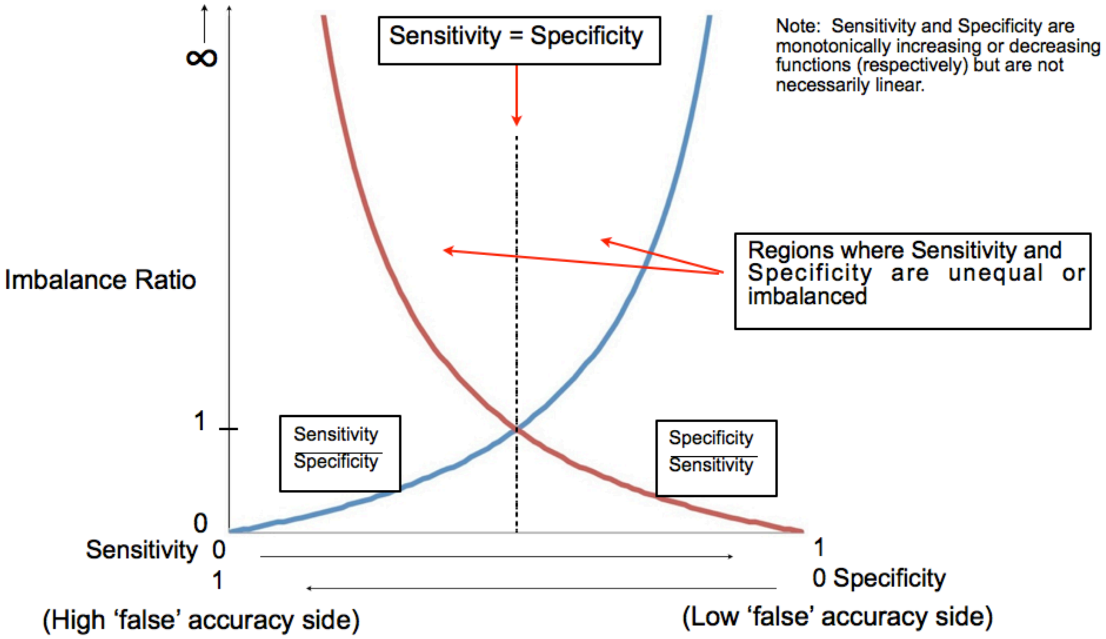


Figure 16. Sensitivity and Specificity as indicators of classifier degeneracy

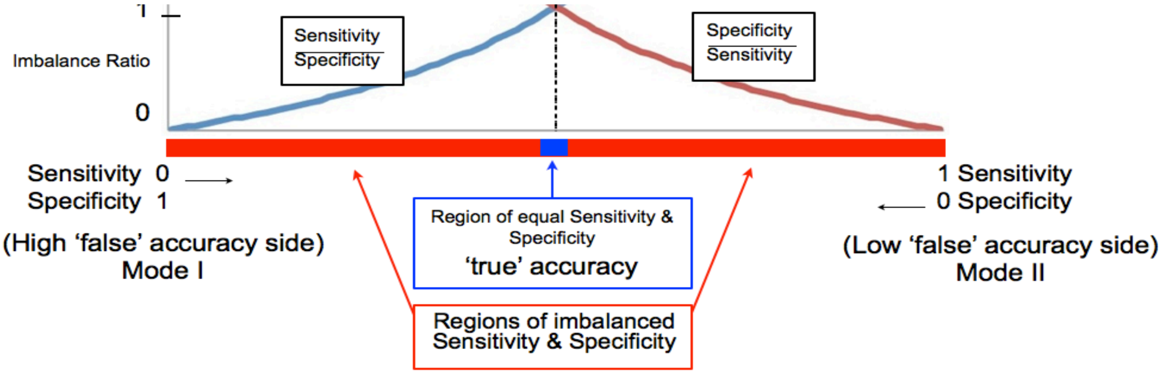


Figure 17. Imbalance ratio derived from Sensitivity and Specificity

By looking at the lower portion of the curve from Figure 16, we can derive an Imbalance ratio or fitness as shown in Figure 17. The Imbalance ratio is also an indicator of fitness or validity of the accuracy of the classifier and can be calculated from

$$IMB = \min\{\text{SPEC/SENS}, \text{SENS/SPEC}\} \tag{109}$$

which is the minimum of either the SPEC/SENS or its inverse. This detects both modes while simultaneously eliminating any infinite values. We now transform the confusion matrix to a 2 dimensional vector **F**(x,y) where x = *IMB* and y = ACC.

## 4.2. Degeneracy Correction

In wrapper methods, when a low ACC value occurs in the array of cost functions feedback the search will chose a feature set based on a higher value. Imbalance in the confusion matrix elements indicates an invalid result (Figures 13, 14). To address this we examine a two dimensional space consisting of simple accuracy (ACC) versus imbalance as shown in Figure 18. ACC can also be described as relevance when it is used in a wrapper search for feature selection in the sense that the features with lower accuracy or overall agreement will be considered less relevant to the solution and subsequently disregarded. Dividing the accuracy vs. imbalance space into 4 separate regions reveals region 1 of high relevance and high validity, where the most desirable results occur. Region 2 of high relevance and low validity are where the most detrimental results occur. Regions 3 and 4 are low relevance. The most desirable results on the plot will be located along the upper left boundary of Region 1 along the vertical axis. Using an arbitrary set of data points from a toy problem a plot of ACC against Imbalance is shown in Figure 19.

Our features are described by a two dimensional value (imbalance, accuracy). While a simple accuracy calculation (Fig. 18.) transforms the information to a one dimension scalar value as required by the wrapper, it disregards the high imbalance.

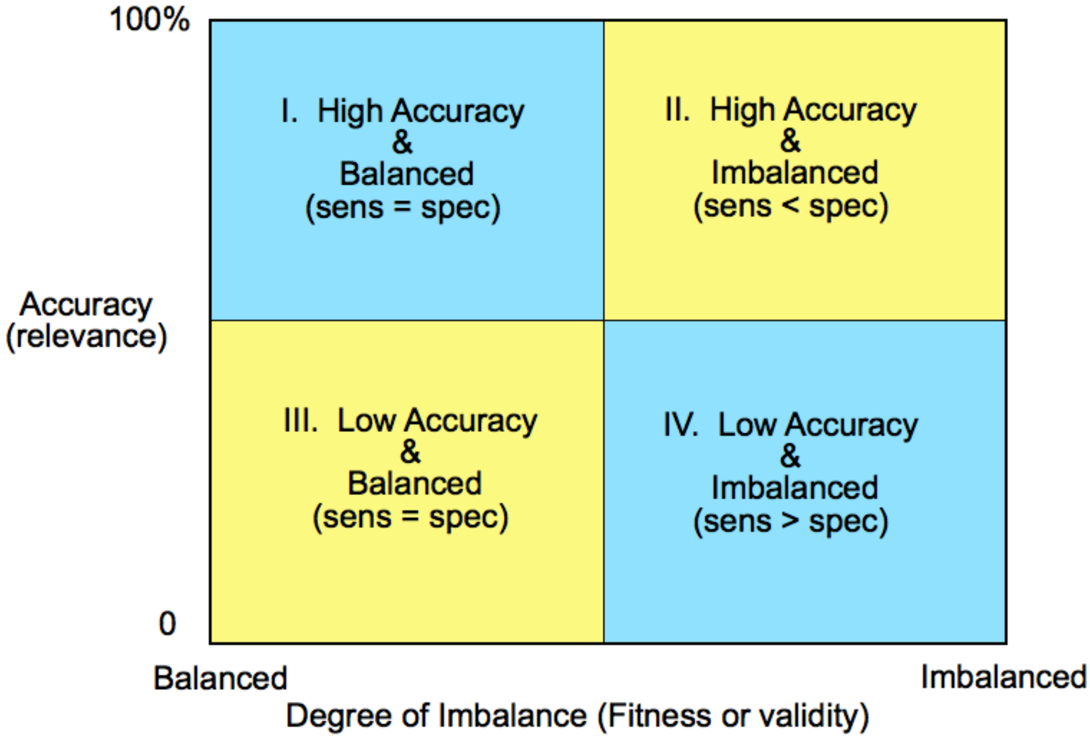

100%
I. High Accuracy & Balanced (sens = spec)
II. High Accuracy & Imbalanced (sens < spec)
Accuracy (relevance)
III. Low Accuracy & Balanced (sens = spec)
IV. Low Accuracy & Imbalanced (sens > spec)
0
Balanced
Imbalanced
Degree of Imbalance (Fitness or validity)

Figure 18. Imbalance versus Accuracy Regions

### 4.3 Q-factor Derivation

Conversion of the two-dimensional classifier metric into a scalar value requires the following transformation

$$F(x,y) \Rightarrow F(y') \quad \text{where } y' = f(G(x), y) \qquad (110)$$

$$G(x) = \min\{SENS/SPEC, SPEC/SENS\} = 1/\max\{SENS/SPEC, SPEC/SENS\} \qquad (111)$$

Either expression for G(x) can be used as needed for convenience of calculation. This is implemented in the form of a correction to the classifier cost function in the feedback path of the wrapper as shown in Figure 20. This accomplishes our objective of moving any F(x,y) with a large value of Imbalance into a region of lower relevance while insuring that y' ≈ y for smaller values of imbalance.

This transformation results in clearing out the elements located in Region 2 and moving them into Regions 1, 3 and 4 as shown in Figure 21. Transformation is accomplished by the Q-factor where Q = ACC x *IMB*

$$Q = \frac{ACC}{\left(\max\left(\frac{SENS}{SPEC}, \frac{SPEC}{SENS}\right)\right)^{p}} \tag{112}$$

*p* is used to adjust the imbalance threshold and is nominally set to *p* = 1. A larger value of *p* will move the boundary between Region 1 and Region 2 closer to the left vertical axis. The additional inclusion of a multiplicative constant in the SENS/SPEC ratio is suggested for further study as an adjustment factor for imbalance in the sizes of the majority and minority classes.

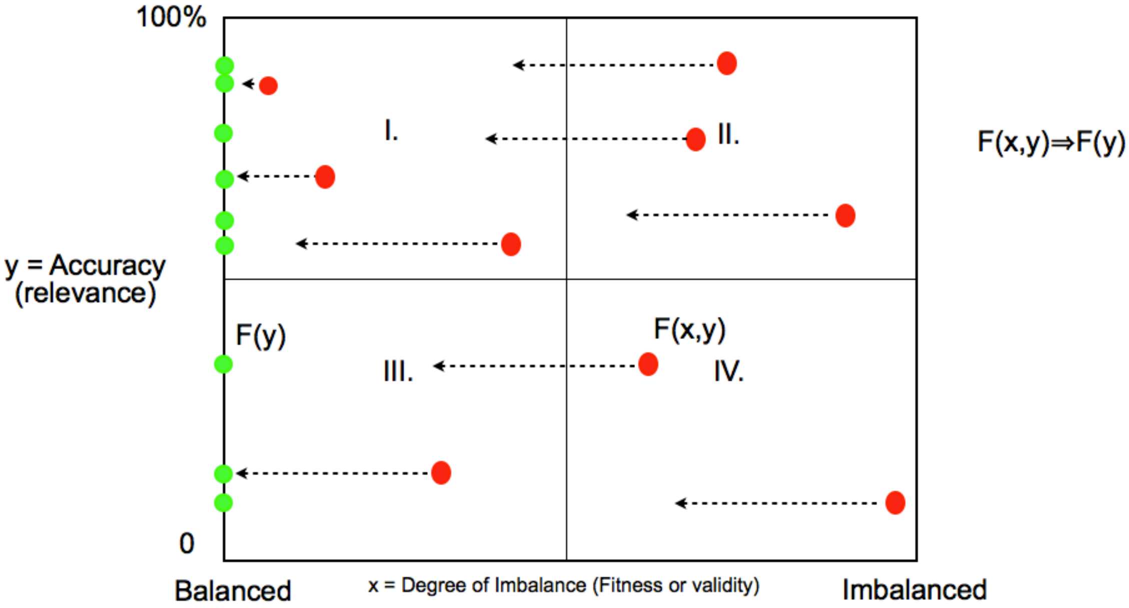


Figure 19. Example of Arbitrary Data Using Imbalance versus Accuracy

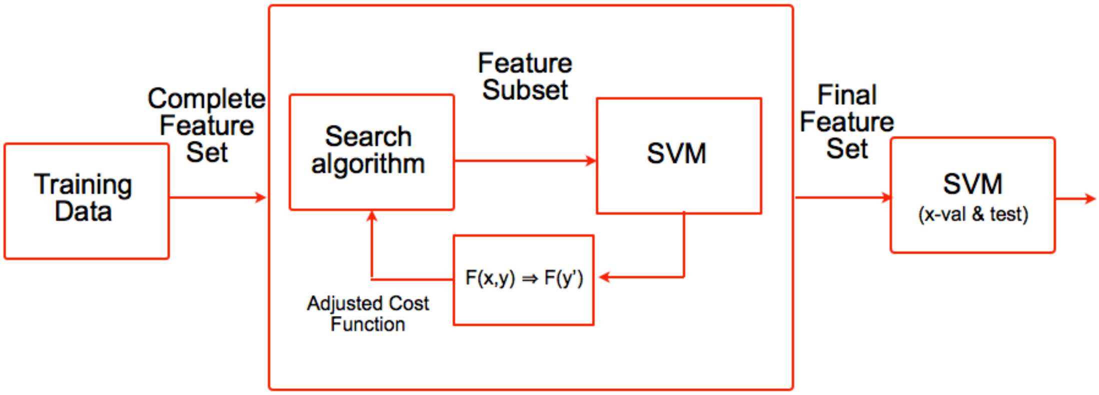


Figure 20. Wrapper Method with Q-factor Correction

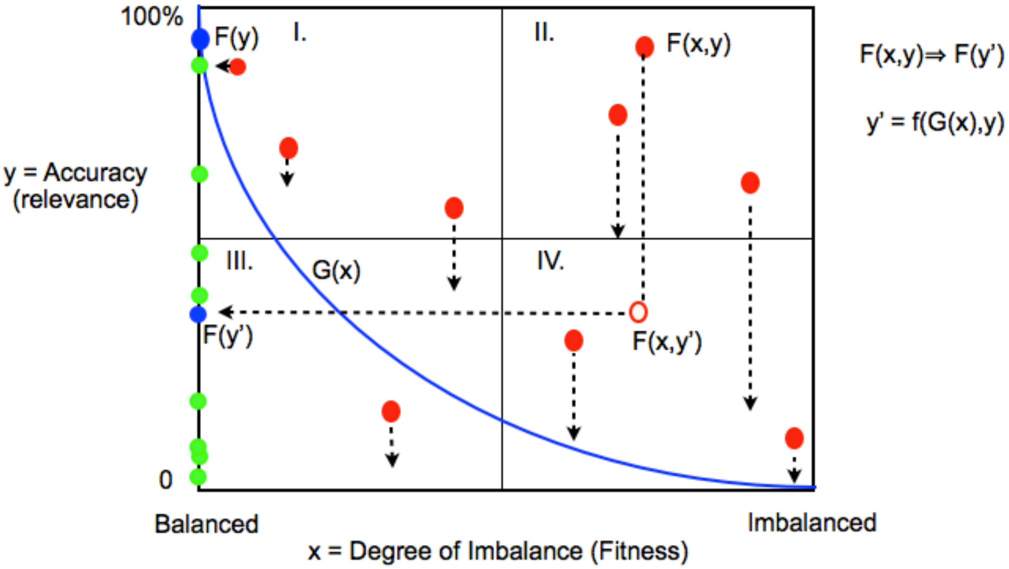


Figure 21. Arbitrary Data Corrected Using the Q-factor

### 4.4 Application to multiclass BCI

A plot of accuracy versus M, the number of classes is shown in Figure 22. The Mode I line describes the locus of the Mode I degeneracy state as described in Section 2.2. The Mode II line corresponds to the 1/M% Mode II degeneracy state. The dashed line represents the constant bit rate line for an estimated value of our SIBCI from M =2 to M = 10, based on (1). For a fixed bit rate channel, the probability of a correct decision decreases as the number of classes increase. For an OVR classifier, the impact of Mode I degenerate results becomes more severe as M increases. If the degenerate result is near or greater than the theoretical limit for a given class, the wrapper has a greater chance of basing its search on the degenerate results.

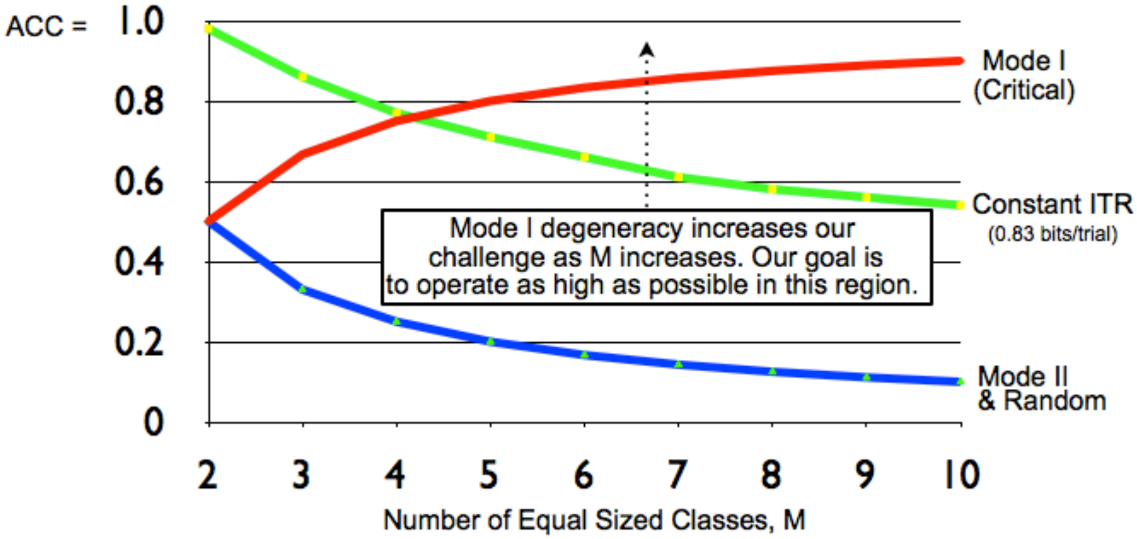


Figure 22. Plot of Degeneracy Modes versus Number of Classes

# CHAPTER 5.

# DISCUSSION OF RESEARCH

The subject independent BCI study described in this dissertation progressed in three phases each of which were conducted entirely by the author. The third and last phase also relied on contributions produced from a separate study, not yet published, led by another CIBIT Lab researcher, Vikas Gottemukkula, with collaboration by this author. Those contributions are acknowledged.

This research focuses on subject independent BCI using in-group subject independence. In-group subject independence methods includes training data from all subjects in the study in contrast with out-of-group studies consisting of machine learning systems that are trained from a diverse group of subjects while validating and testing on other subjects that are not part of the training set. While out of group SIBCI may also be considered the ultimate goal of subject independence, the state of the art has not yet consistently achieved high rates of accuracy with out of group methods. Out of group SIBCI was outside of the scope of this study.

Prior to conducting subject independent studies, the initial research activities focused on a less complex single subject BCI for the purpose of refining the subject protocol, observing the performance of different features and determining the sensitivity of the feature modalities and SVM to parametric variations. This work established a baseline for comparison for later SIBCI study phases. It also provided guidance for starting points with feature parameters when the more complex SIBCI models are implemented in later stages. Data was collected for motor tasks, cognitive tasks and affective (emotional) tasks in this single subject study.

As a result of the single subject work, the subject independent protocol was realized and revisions were made to the tasks were based upon quasi movements. This activity also provided the background for refining the timing and cues for the subject prompts and the number of movement tasks for the BCI. The decision to investigate the eight feature modalities was based on these single subject activities. Classifiers types were chosen from the results of single subject studies. During this research, the one versus the rest classifier method was used for SVMs. The influence of trivial classifiers was observed in the overall accuracy results. The results of this activity were presented at the 2009 IEEE INNS International Joint Conference on Neural Networks and published in the Conference proceedings [120].

The second phase of the research introduced the subject independent BCI approach. The quasi-movement data collection protocol was introduced. Individual feature modalities (periodgram, wavelets, cepstrum) were examined and adjusted to refine both feature and classifier parameter settings, the best performing wavelet families were selected and the methods for combining SVM classifiers to produce a subject independent system were investigated. During this phase the OVO SVM classifier was implemented, in an attempt to correct the influences of trivial classifiers on the final results encountered in the previous phase. Ensemble methods for combining classifier outputs were used for combining the multiple binary classifiers. Comparisons of majority vote, hamming and error correcting code methods were made. This research provided some additional benchmark values for SIBCI methods for use in future studies. The results of this particular study were published [118].

Additional SIBCI research was conducted in collaboration with another researcher in the CIBIT Lab to evaluate performance of the SIBCI among various classifiers and wrapper feature selection methods based on the modalities and parameters chosen above. This work included an initial but cursory implementation of the Q-factor method. The author was not the primary researcher in this activity but acted in a supporting role, providing input on parameters selection and methods. This activity was accepted for publication in Feb 2011 in the International Journal of Computational Intelligence Theory and Practice (IJCITP). This author was the second author for this publication.

The third research phase was further refined by inclusion of SFFS feature selection implemented within a wrapper method. This classifier was designed with SVM in an OVR configuration. The decision to use OVR was justified by recognizing that the complexity and number of free parameters encountered when combining the OVO classifiers each with a different feature subset produces an overly complicated model. The OVR method was used for the consolidation of the feature sets. The Quality-factor was implemented to compensate for the occurrence of trivial classifiers in the one versus the rest arrangement and eliminate biases that produced misleading results. The necessity for using the Q-factor was examined closely for its role in guiding the search algorithm within the wrapper approach. An additional refinement made to this study was the elimination of the first and last trials from each of the individual sessions, reducing the effects of subject experience and habituation. The results in this study are compared between Q-corrected and uncorrected versions of SIBCI, CSP filtered SIBCI, subject specific BCI. The results of this study have been submitted for publication in February 2011 to the Journal “Pattern Recognition Letters” [119].

# CHAPTER 6.

# RESULTS

## 6.1. Wavelets and Single Subject BCI

Our first study investigated the behavior of single subject BCI by looking at imagined motor, cognitive and affective tasks [120]. Two independent recording sessions were conducted with an 8 day separation between the sessions. Twelve repetitions of each of the motor, cognitive and affective tasks were executed during each session. The subject was cued by a visual prompt from a computer screen to initiate each specific task, a 10-second relaxation interval was inserted between each sequential task.

Signals for the motor cortex were captured from electrodes C3 and C4. Imagined motor tasks selected were Left Hand and Foot combined, right Hand and Foot combined, both Hands and both Feet. This choice of tasks was a test paradigm to investigate hemispheric specialization of the brain. We could not find a precedent for this in our literature search and included it as a novel alternative to explore contra-laterality [115]. Additionally we included tongue movements to the left and right, which have demonstrated strong separation in EEG studies [91]. The clockwise rotation of a 3-dimensional cube, a two-digit multiplication problem and the mental composition of a letter were selected for three cognitive tasks. Affective states were induced by images with positive and negative emotional valences, selected from the Internation Affective Picture System 1999, produced by the National Institute of Mental Health (NIMH) Center for the Study of Emotion and Attention, University of Florida, Gainesville, FL, provided to our lab by the Dr. Christopher Lovelace of the University of Missouri – Kansas City psychology department. Cognitive signals were collected from electrodes O1 and O2 [28] and affective signals from electrodes Fp1 and Fp2 [97]. Spatial Large Laplacian references were chosen for the motor cortex signals. Bipolar references were chosen for the cognitive and affective signals, as the location of these electrodes are on the outer circumference of the EEG cap and not amenable to spatial Large Laplacian reference.

TABLE I
SUBJECT TEST PROTOCOL

| Task | Duration (Seconds.) |
|---|---|
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Left Hand and Foot (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Right Hand and Foot (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Both Hands (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Both Feet (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Tongue Left Movement (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Tongue Left Movement (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Relax (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| 2-Digit Multiplication Problem (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Clockwise rotation 3-D Cube (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Letter Composition (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Relax (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Pleasant Visual Image (No Blinks allowed) | 8 |
| Relax (Blinks allowed) | 10 |
| Relax (No Blinks allowed) | 2 |
| Unpleasant Visual Image (No Blinks allowed) | 8 |

A diagram of the entire classifier scheme for this phase of the study is shown in Figure 1.

For the single subject, BCI, we used two methods to eliminate blink artifact signals. The subject protocol was designed to instruct the subject to refrain from blinking during the data collection and task performance interval. Additionally, any remaining artifacts were examined manually and removed. Efforts were made to align zero crossings and match waveform slopes when reassembling the time domain signals after artifact removal. Manual artifact removal was not performed during the subject independent data sets, as this was not practical based on the sheer volume and number of actual data sets.

During this project, the emphasis was placed upon determining which wavelet methods and mother wavelet families would yield the best performing classifiers. The 77 types of orthogonal and biorthogonal wavelets from the MATLAB wavelet tool box were selected. For comparison the direct feed of temporal coefficients into a Feed-Forward Neural Network (FFNN) were also used as features. The classifier results were examined using SVM classifiers according to a OVR arrangement. Separate classifiers were built separately for motor, cognitive and affective tasks. The results were presented in two formats, first classifiers were built (trained, validated) only from the first session ignoring the effects of any session to session variation in EEG patterns. These classifiers were then blind tested on signals exclusively from the second session. Secondly, a battery of classifiers were built with signals from both sessions (training and validation) and then blind tested on a smaller subset of the signals collected from the end of session 2. This permitted an analysis of session to session transfer behavior on EEG signals for BCI.

During this study, SVM C-parameters were adjusted in addition to wavelet decomposition levels. This allowed us to determine the relationships of hard and soft margins to classifier accuracy along with the performance of wavelet decomposition methods and wavelet families. We also were able to gather a sense for the accuracy levels that can be obtained for cognitive and affective tasks.

Finally, this research laid the groundwork for later subject independent studies. We were able to explore the concept of BCI, establish a subject protocol for data collection, identify critical parameter settings, identify performance benchmarks and determine an overall framework for our future investigations. As a result of this study, the 77 wavelet families were categorized in terms of accuracy performance and the top performing were employed in future stages of the research. The calculated accuracy results of this study also revealed some behaviors related to the occurrence of trivial classifiers and the number of tasks. These trivial classifiers became apparent at the extremes of parameter settings for classifiers and feature decomposition levels, they were duly noted. The results were manually inspected and adjusted to eliminate this bias from the resulting accuracy calculations and was determined to have a negligible influence in the results of the study. This was easily accomplished as the degeneracies only played a part in final accuracy calculations and were not used to guide any type of search algorithm within a wrapper. It became apparent that the bias would become more substantial as the number of classes increased in a one versus all the rest classifier arrangement. The effect of this bias was more noticeable for poorer performing features and classifiers. Scores from the inferior performing multitask classifiers were skewed toward a higher accuracy as the number of tasks increased which at first observation, appeared to be counter-intuitive and investigated further. The presence of trivial classifiers that identify all (or most) samples as being members of the majority class was determined to be the cause of this behavior. The only solution to the problem at this point was to manually eliminate the purely degenerate classifiers from the calculation of the results. A purely degenerate classifier, for this purpose is defined as one that identifies all samples as belonging to the majority group (class non-membership). However, this did not address any classifiers that tended toward this degenerate state but were not fully degenerate but exhibited some varied degree of degeneracy. The worst case scenario, consisting of classifiers that detected no true positives and a very small number of false positives, still skewed the results. Such classifiers still produced misleading accuracy results approaching M(M-1)/2 %. The influence of this can be seen in those results which show a tendency to cluster near the M(M-1)/2 % and was noticeable in the raw coefficients classifiers where the features possess higher feature dimensionality.

The best results were dominated by symmetric wavelets including biorthogonal wavelets a brief summary of results is presented in table 2. A full discussion of the methods and results are presented in [120].

Table 2. Summary of results for Wavelet and Wavelet Packet Features for Single Subject BCI

Motor Tasks - scheme B
Classifier accuracy -Wavelet Decomposition

| Wavelet | Validation | Testing | Decomp Order | C parameter |
|---|---|---|---|---|
| rbio3.7 | 83.8% | 82.8% | 2 | $10^9$ |
| rbio2.8 | 83.5% | 82.3% | 6 | $10^9$ |
| bior5.5 | 83.4% | 82.3% | 2 | $10^9$ |
| rbio2.2 | 83.3% | 82.9% | 1 | $10^9$ |
| rbio1.1 | 83.3% | 82.5% | 2 | $10^8$ |

Motor Tasks - scheme B
Classifier accuracy -Wavelet Packet Decomposition

| Wavelet | Validation | Testing | Decomp. Order | C parameter |
|---|---|---|---|---|
| rbio3.3 | 83.5% | 82.1% | 3 | $10^9$ |
| rbio3.1 | 83.3% | 82.6% | 1 | $10^9$ |
| rbio1.1 | 83.2% | 82.4% | 1 | $10^8$ |
| bior1.1 | 83.2% | 82.4% | 1 | $10^8$ |
| dB1(haar) | 83.2% | 82.4% | 1 | $10^8$ |

## 6.2 Ensembles of Classifiers and SIBCI

Equipped with a set of wavelet family candidates, the next stage of the research shifted from subject specific to the subject independent paradigm. The collection protocol was modified to accommodate four imagined motor tasks. The protocol was adjusted to left hand, right hand, right foot and left foot quasi-movements. This modification allowed for a more intuitive set of tasks requiring less co-ordination for the BCI volunteers who had no previous BCI experience. These four tasks are also a more appropriate mapping for a practical BCI using individual hand and foot movements rather than left side versus right side or both hands versus both feet. Prior research was found using quasi-movements or small but restrained movements for BCI [85]. Quasi-movements were selected as this was expected to be more intuitive for inexperienced volunteers that purely imagined tasks. The four tasks were expected to take better advantage of the contra-laterality characteristics of event-related desynchronization (ERD) of the mu and beta rhythms [95]. In order to explore this, 10 volunteer subjects were recruited (8 males and 2 females, aged from 21-28 years), this the same subject population described in the earlier analyses. The sequence of visual prompts used for the single subject data collection was shortened to the four motor tasks. An additional 1/2 second, 500 Hz audio alert tone was added to the beginning of each task execution interval. This alert cue was added to raise the volunteer's attention level at the beginning of the measurement interval. Declines in attention level are known to produce increases in alpha frequency waves [5,104] which could be detrimental to BCI accuracy. As in the previous stage, each subject participated in two recording sessions separated by an interval of 4 – 6 weeks between recordings.

Each session consisted of 12 repetitions of the 4 tasks. A 10 second rest interval was inserted between each task. The time between each of the 12 repetitions varied from one to several minutes. The recording timeline is presented in Figure 8. While data was collected from all electrodes, for this study the motor cortex signals were collected from C3 and C4. Spatial Laplacian filters were applied to both of these signals. Manual blink artifact removal was not performed as the protocol instructed volunteers not to blink. It was expected that a small number of blinks would occur in the data signals. For a practical BCI either blinks would have to be tolerated or processed using an automatic algorithm. Our expectation was that blink artifacts would be detrimental to the classifier performance as they would be uncorrelated with specific tasks. A determination of the exact impact of blinks on the results of this study is a possible subject for future investigation, as well as the impact of different artifact removal methods.

Features were selected which exploit the temporal and spectral characteristics of the EEG signals. Selected features include linear predictive coefficients (LPC), power spectral density (PSD) amplitude coefficients, spectrogram or short time Fourier transform (STFT) coefficients, cepstrum coefficients, and wavelet and wavelet packet. Wavelet families were chosen based on the findings described for the first phase of the study. The parameter settings for the other features were determined by trial and error. Each set of features was applied to an SVM classifier and the overall accuracy was determined. To mitigate the trivial classifier issue discovered from the prior BCI study we used the OVO approach with ECOC output level fusion. As a six-fold validation was employed and classifiers were built from features collected from both left and right hemispheres, 12 classifiers were combined for each feature modality. A block functional diagram of the feature extraction and classifier architecture is shown in Figure 23.

TABLE 3. SUBJECT INDEPENDENT BCI MOTOR TASKS - ENSEMBLE CLASSIFIER ACCURACY

| Feature | Accuracy |
|---|---|
| LPC | 40.6% |
| PSD | 25.6% |
| STFT | 50.6% |
| Cepstrum | 30.0% |
| WDC | 56.9% |
| WPC | 28.8% |
| WDE | 56.9% |
| WPE | 42.5% |
| Ensembles | 58.8%/59.4%/70.4% |

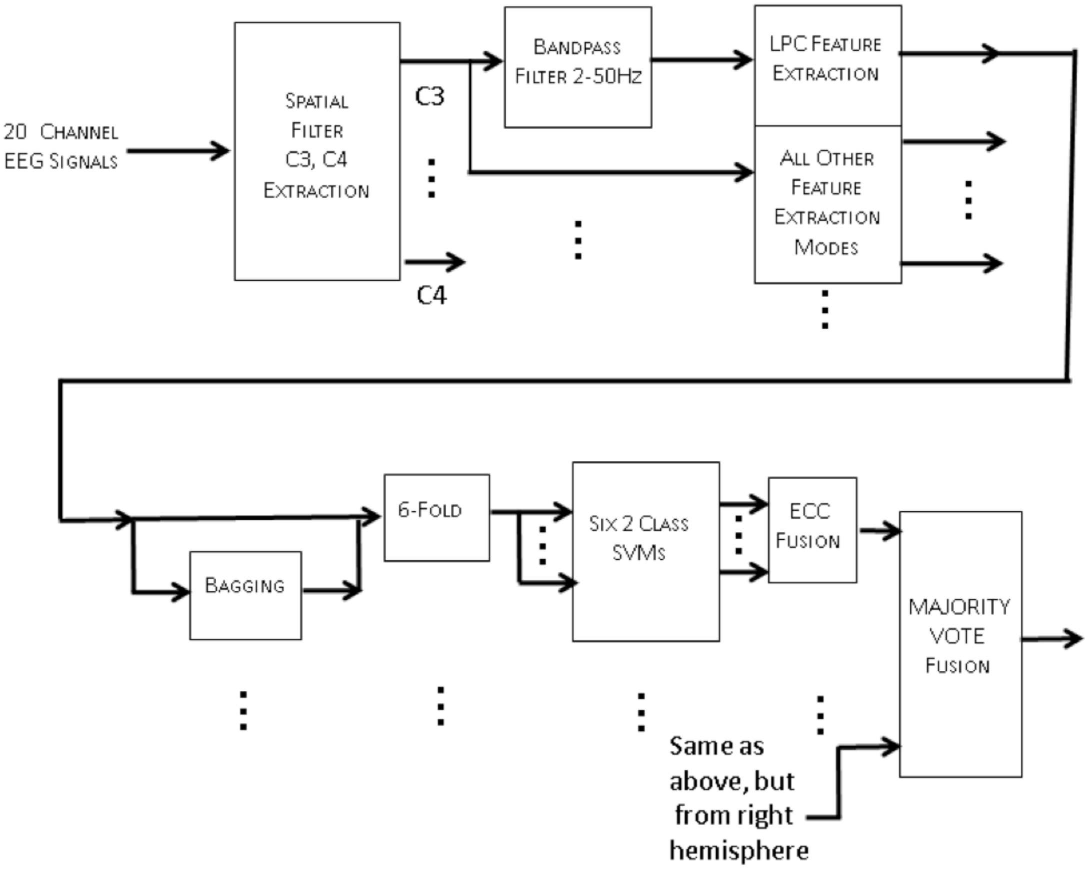


Figure 23. Functional Diagram of Ensemble SIBCI

By eliminating single modalities which classification accuracy rates of less than 30%, and combining the better performing modalities, accuracies of 59.4%, 58.8% and 70.4% were obtained for validation results. The sensitivities and specificities of these results were examined. It was determined that while these classifiers were much less than ideal none of these classifiers exhibited any detectable degeneracies, in all cases meeting the conditions of SENS/SPEC ≈ 1. In all three cases, sensitivity, specificity and accuracy were approximately equal. While these results were poorer than comparable single subject BCI studies they do reflect the difficulty in establishing a 4 task SIBCI. These results reflect the use of training data from session 1 and session 2 against a blind test set from session 2. In other words, the division of the training and testing data were made in a chronological order where testing data was always chosen from the last repetitions of the second session. Calculations of ITR for these accuracies map to detection rates as high as 80-92% for a corresponding 2 task BCI with a comparable information rate. This compares favorably to other comparable SIBCI investigations found in the literature [2]. The results are briefly summarized in Table 3 and a full description of the study is provided in [118].

### 6.3 Wrapper Feature Selection and SIBCI

The most recent phase of research continued the focus on SIBCI and utilized the same 10 volunteer dataset described above. The first and last set of task repetitions for each subject and session were removed from the study to allow time for some adaptation to the protocol and reduce the effect of habituation. The best results for the prior study required apriori knowledge of the particular feature modalities for creation of ensembles. For a working subject independent BCI it cannot be assumed that one particular modality will yield the best results. Furthermore the output level fusion of the balanced two class SVMs involves a greater number of additional free parameters. Several fusion methods are available to choose from. While these factors introduce a degree of flexibility and choices, Occam's razor suggests that a simpler approach may offer a better solution. For BCI implementations based on large numbers of tasks, the idea of simplicity bears consideration. To this end, we seek out a common set of features using one separate classifier for each task requiring the use of OVR classifier arrangements. Each task will be detected based on a subset of the total concatenated feature vector. The concatenated vector is a very high dimensional feature, too high to be useful by itself for classification. Selection of a subset of features from the full set is achieved through a wrapper method. The wrapper employs an SFFS induction algorithm and the SVM classifier. As was discussed previously, the OVR support vector machines can produce misleading results when degenerate classifier outputs are generated. The impact of degenerate classifiers in the wrapper feedback loop results in wrappers converging to false minima, or not converging at all and becoming trapped in infinite loops. This is avoided by use of a quality metric calculated from the sensitivity and specificity values of confusion matrix of the classifier output. Sensitivity and specificity provide a measure of the degenerate behavior in the classifier output. A metric calculated from sensitivity and specificity can be used to evaluate the fitness of the classifier solution during each search iteration. Hence, a Q-factor corrected accuracy factor was used as the cost factor within the wrapper method. The study was conducted for a subject independent paradigm with Q-factor correction, without Q-factor correction, and repeated for the same features collected from sampled time domain signals subjected to CSP filtering prior to feature extraction. Figure 24 shows a comparison of SIBCI results before and after applying the Q-factor correction. This figure shows the aggregation of all four tasks, plots of SIBCI for each individual task are provided in Appendix B.  Figure 25 shows the results for CSP.

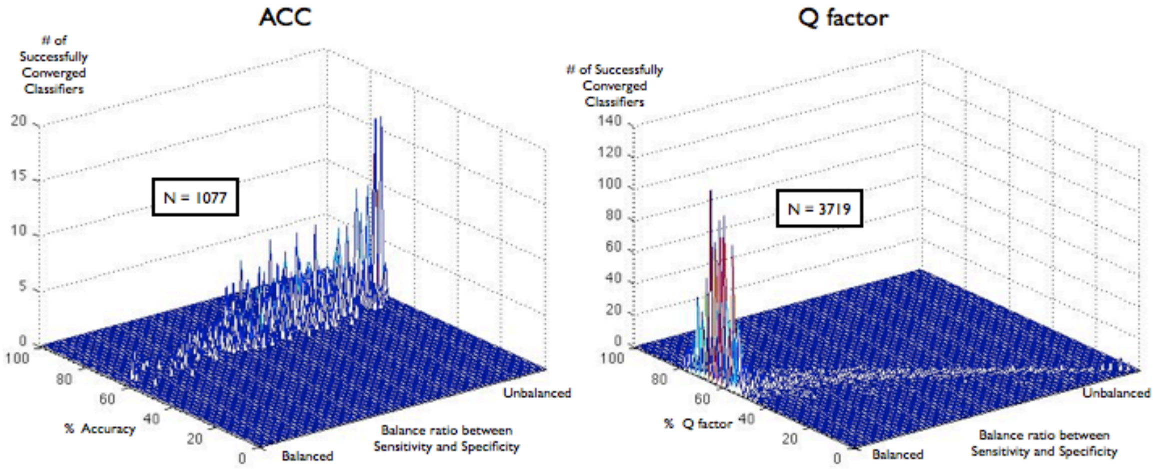


Figure 24. SIBCI results without and with Q factor.

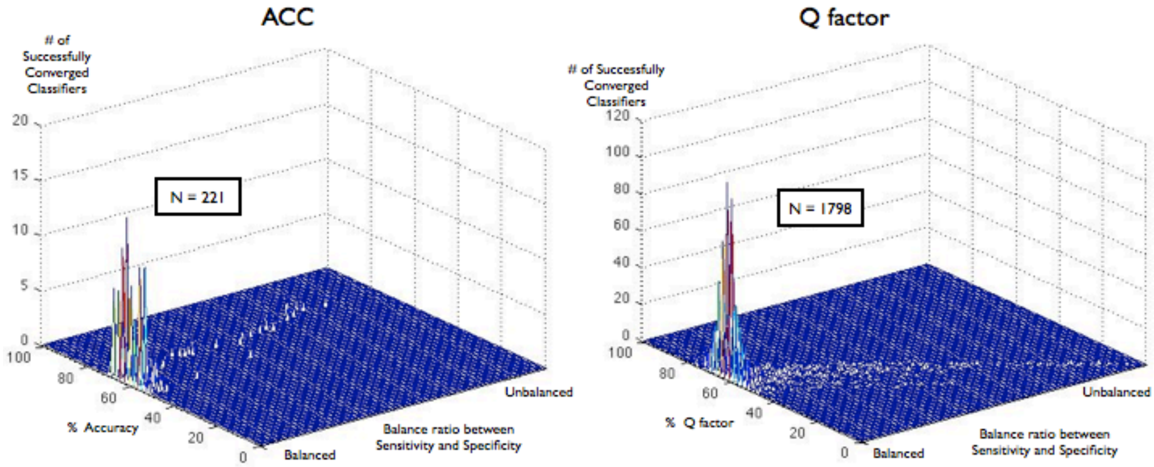


Figure 25. SIBCI results without and with Q factor (CSP).

Additional studies were conducted upon each of the 10 subjects using a subject specific BCI, where all the classifiers were trained only from data collected from the same subject. Classifiers were implemented with and without Q-factor correction and all performance results were evaluated.  The subject specific results combined for all subjects are shown in Figure 26.  Addition subject specific results for each individual subject are presented in Appendix B.

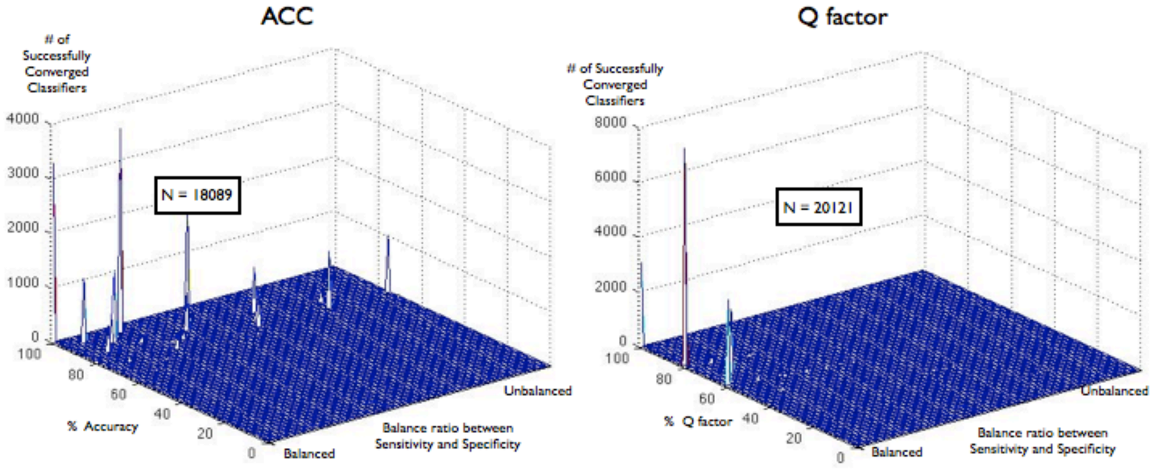


Figure 26. Subject Specific BCI results without and with Q factor.

Accuracy and imbalance values were documented for all cases with and without Q-factor correction. The number of classifiers produced from converging wrapper solutions were also noted for each scenario. The influence of Q-factor corrections on SIBCI was clearly observed within these studies. The influence of accuracy results biased toward (M-1)/M% accuracy from the degenerate classifiers was reduced by the Q-factor. The number of feature searches actually increased by a 4-fold amount after Q-factor correction. The impact of Q-factor correction on subject specific models was substantially less, subject specific data sets are easily more separable for reasons presented earlier and do not produce as many degenerate classifiers. By applying CSP filtering to input data signals prior to feature extraction, an increased amount of separation is induced in the projected EEG signals, thus reducing the occurrence of degenerate classifiers within the feature space, some slight improvements were realized. Information transfer rate calculations on the overall q-corrected results reveal greater performance with this approach when compared with other SIBCI studies using comparable scale and task protocols [2].

TABLE 4. SUMMARY OF RESULTS FOR SIBCI WITH Q FACTOR.

| | Imbalance | | Accuracy (%) | | Number of Convergent Classifiers |
|---|---|---|---|---|---|
| | Mean | Std. Dev. | Mean | St. Dev. | |
| | | | Uncorrected | | |
| SI | 0.7051 | 0.1930 | 72.75 | 40.48 | 177 |
| SS | 0.2991 | 0.2829 | 89.44 | 8.64 | 18089 |
| SI CSP | 0.0707 | 0.636 | 56.78 | 9.70 | 221 |
| | | | Q factor Corrected | | |
| SI | 0.1635 | 0.2695 | 64.41 | 7.30 | 3719 |
| SS | 0.0384 | 0.0958 | 80.34 | 12.82 | 20121 |
| SI CSP | 0.0920 | 0.1794 | 56.94 | 11.26 | 1798 |

Mean and standard deviation values for Subject Independent (SI) and Subject Specific (SS) BCI features with and without Q factor corrections. Imbalance is calculated by 1- min{sens/spec, spec/sens} where 1 is unfavorable and 0 is favorable. The accuracy calculation is based on ACC for uncorrected models and Q-factor for corrected models. CSP indicates common spatial patterns. The number of SFFS feature vectors that produced converged classifiers is in the rightmost column.

With this implementation we were able to overcome the shortcomings of the OVR classifier scheme, eliminate the effects of classifier degeneracies and produce a useful metric that can guide the induction algorithm in a wrapper based feature selection. This permits the use of very high dimensionality feature vectors created from multiple concatenated feature vectors.

As described earlier, one of the obstacles to creating OVR BCI classifiers from large numbers of tasks stems from the behavior of degenerate classifiers. The misleading accuracy results produced by these degenerate conditions produce biases that obscure the performance of the non-degenerate classifiers. This approach overcomes the aforementioned obstacle and can be useful in producing classifier arrangements that can separate a greater number of classes.

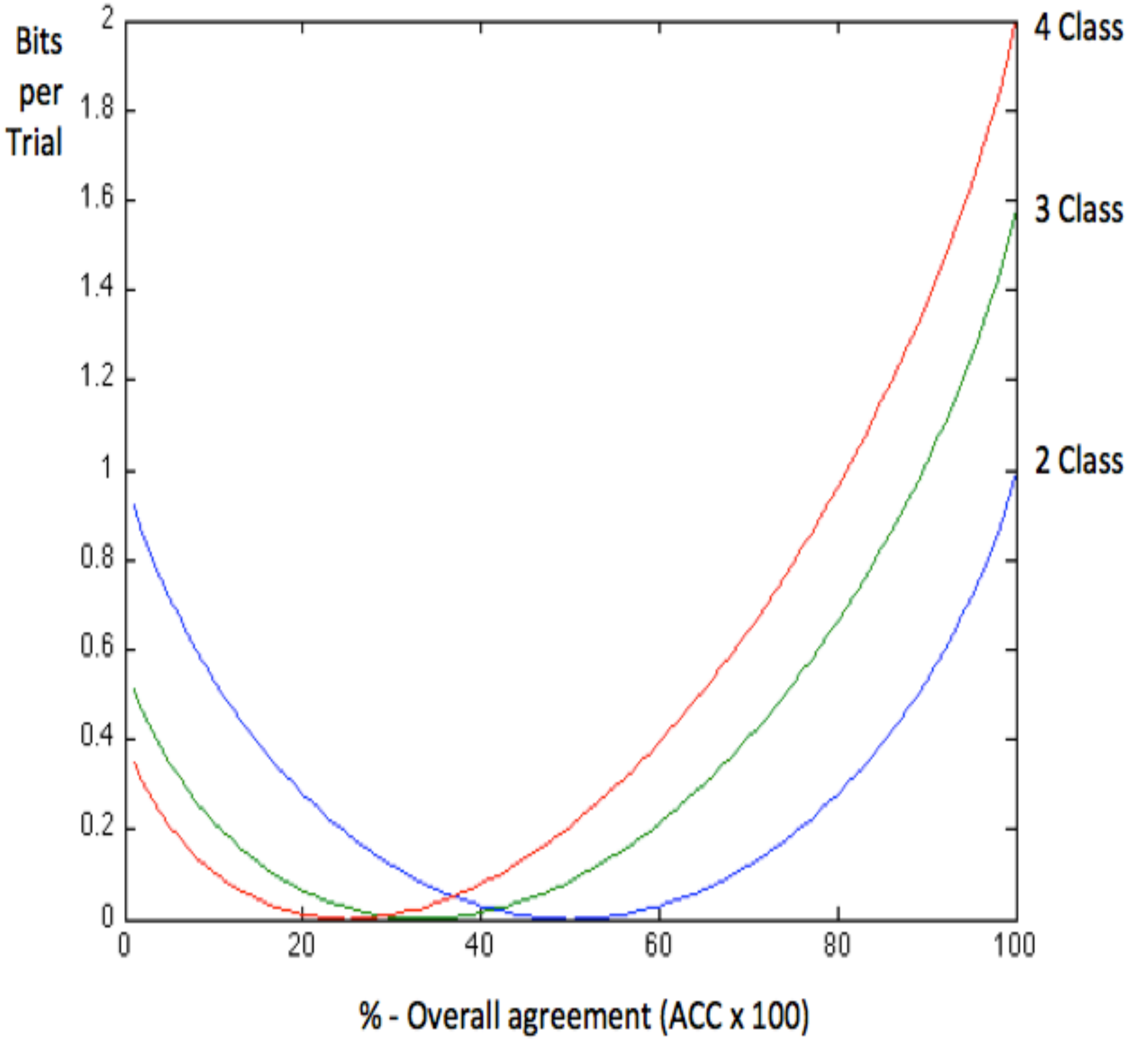
# CHAPTER 7.

# CONCLUSION

A subject independent brain computer interface based on features extracted from EEG signals cannot be expected to outperform its subject specific counterpart due to the intersubject variances present in EEG signals. However the level of performance achieved in a subject invariant implementation can be improved by the use of the techniques described in this study. The use of SFFS to find a common set of features, combined with the OVR SVM implementation produces satisfactory information transfer rates for a four class BCI. Additionally, by conducting a detailed analysis of the degenerate modes encountered in unbalanced SVMs we have been able to overcome one of the major obstacles toward achieving discrimination for large numbers of motor tasks. This technique has future implications as BCI implementations move beyond the detection of 3-4 tasks. The use of the Q-factor also has possible implications in other classification applications beyond BCI when separable features are difficult to obtain.

# APPENDIX A.

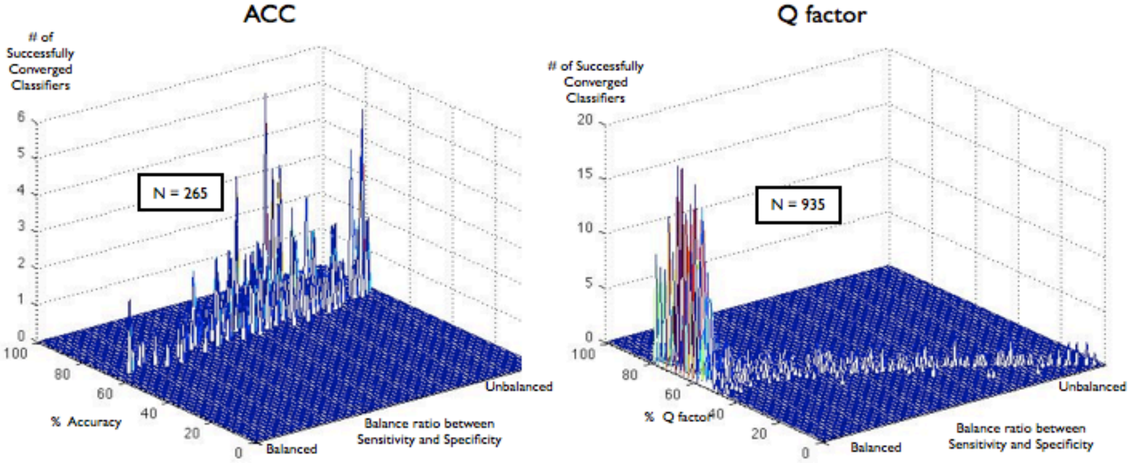
# Appendix D. Information Transfer Rate



# APPENDIX B.

## SIBCI with SFFS

Left Hand

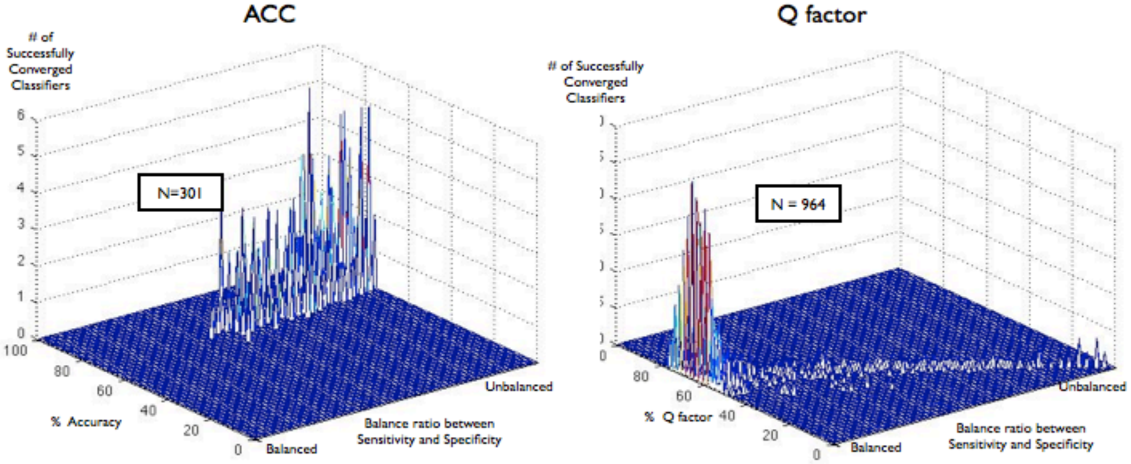


Plot of subject independent feature sets vs. accuracy or Q-factor

10 subjects; all modalities, cross validation results
N = total number of converged classifiers

## SIBCI with SFFS

Right Hand

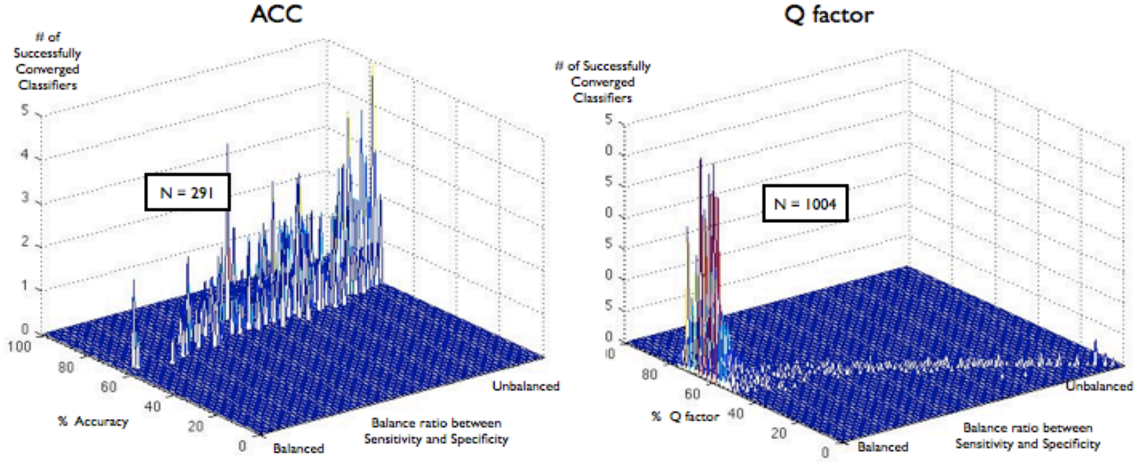


Plot of subject independent feature sets vs. accuracy or Q-factor

10 subjects; all modalities; cross-validation results
N = total number of converged classifiers

## SIBCI with SFFS

Left Foot

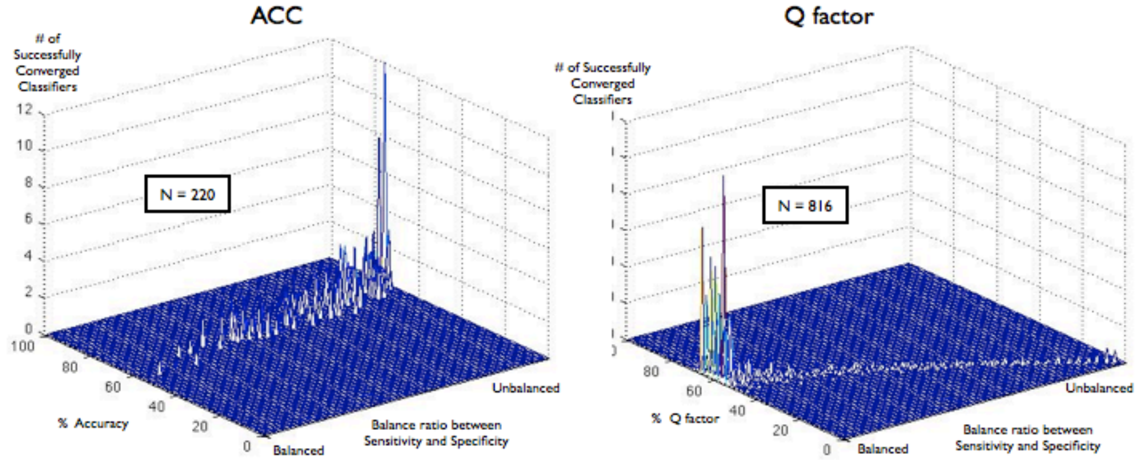


Plot of subject independent feature sets vs. accuracy or Q-factor

10 subjects; all modalities, cross validation results
N = total number of converged classifiers

## SIBCI with SFFS

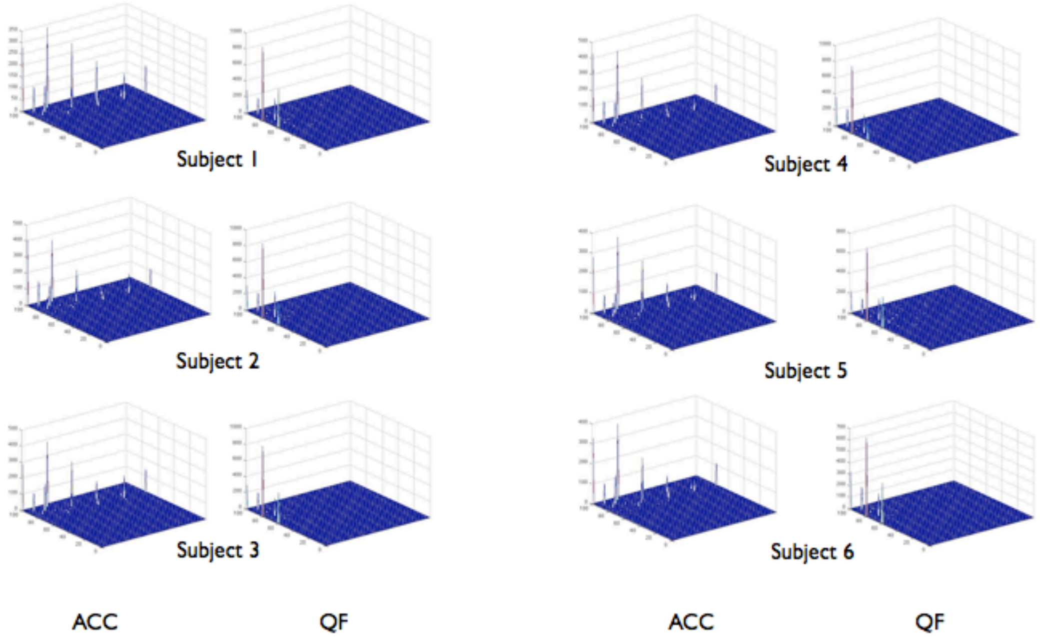
Right Foot



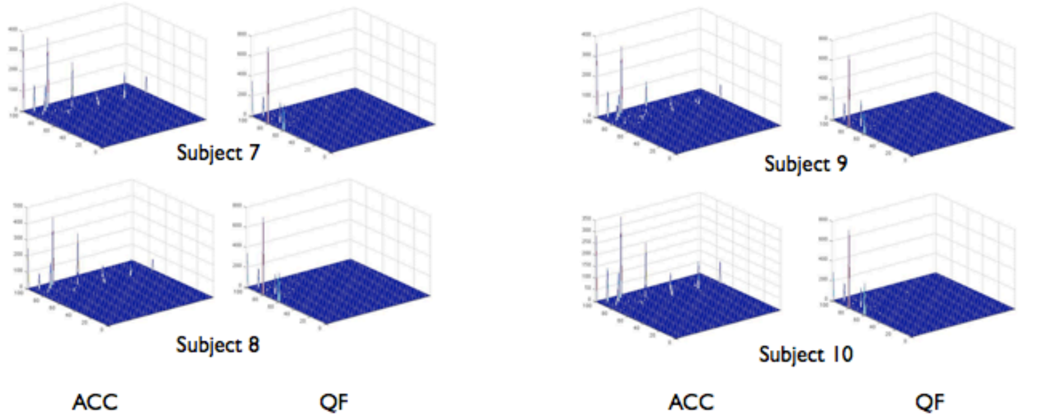
Plot of subject independent feature sets vs. accuracy or Q-factor

10 subjects; all modalities, cross validation results
N = total number of converged classifiers

# SS-BCI
## per subject results



# SS-BCI
## per subject results



For every case the total number of successfully converging classifiers is greater when QF is applied than for ACC.

| | | | |
|---|---|---|---|
| S1: | 1852 → 2403 | S6: | 1837 → 1881 |
| S2: | 1909 → 2224 | S7: | 1755 → 1901 |
| S3: | 1915 → 2118 | S8: | 1757 → 1937 |
| S4: | 1918 → 2224 | S9: | 1683 → 1849 |
| S5: | 1909 → 2224 | S10: | 1729 → 1891 |

# BIBLIOGRAPHY

1. Achuff, P. The lateralization of emotion. *Brain and Mind*. Nov 4, 2001 State University of Campinas, Center for Biomedical Informatics, Campinas Brazil [Online] http://www.cerebromente.org.br/n14/mente/lateralization.htm accessed on June 14, 2011

2. Alamgir, M., Grosse-Wentrup, M., and Altun, Y. Multitask learning for brain-computer interfaces. In *JMLR W & C Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Sardinia, Italy, May 13-15), JMLR, Cambridge, 2010, pp. 17-24.

3. Allwein, E. L., Schapire, R. E., and Singer, Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1 (2000), 113-141.

4. Alpaydin, E. *Introduction to Machine Learning*. MIT Press, Cambridge, 2004.

5. Andersen, P. and Andersson, S. *Physiological basis of the alpha rhythm*. Appleton-Century-Crofts, New York, 1968.

6. Anderson, C.W. Taxonomy of feature extraction and translation Methods for BCI, presented at the 3rd International Meeting on Brain-Computer Interface Technology, 2005, [Online] http://www.cs.colostate.edu/eeg/taxonomy.html, accessed on June 10, 2011

7. Anderson, C. W., Stolz, E. A., and Shamsunder, S. Discriminating mental tasks using EEG represented by AR models. *Engineering in Medical Biology Society Annual Conferences*, 2 (1995), 875-876.

8. Bashanti, A., Fatourechi, M., Ward, R. K., and Birch, G. E. A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. *Journal of Neural Engineering.* 4 (2007) R32-R57.

9. Bashashanti, A. *Towards Development of a 3-State Self-Paced Brain Computer Interface.* Ph. D. Dissertation, University of British Columbia, Department of Electrical and Computer Engineering (2007).

10. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer Science Business Media, LLC., New York, 2006.*System*. 2007.*Neural Engineering*, 4 (2007), R32-R57.

11. Blankertz, B., Dornhege, G., Krauledat, M., Kunzmann, V., Losch, F., Curio, G., and Müller, K.-R. The Berlin Brain-Computer Interface: Machine Learning-Based Detection of User Specific Brain States. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2004.

12. Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., and Curio, G. The non-invasive Berlin brain-computer interface: fast acquisition of effective performance in untrained subjects. *Neuroimage*, 37 (2007), 539-550.

13. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K.-R. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing*, 25, 1 (Jan 2008), 41-56.

14. Blankertz, B., Losch, F., Krauledat, M., Dornhege, G., Curio, G., and Müller, K.-R. The Berlin brain-computer interface: accurate performance from first-session in BCI-naive subject. *IEEE Trans. on Biomedical Engr.*, 55, 10 (2008), 2452-2462.

15. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K.-R. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing*, 25, 1 (Jan 2008), 41-56.

16. Bogert, B. P., Healy, M. J. R., and Tukey, J. W. The quefrency analysis of time series for echoes: cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis* (New York 1963), Wiley, New York, pp. 209-243.

17. Borisoff, J. B., Mason, S. G., and Birch, G. E. Brain Interface Design for Asynchronous Control. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2007.

18. Bostanov, V. BCI competition 2003-data sets Ib and IIb: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Trans. Biomedical Engineering*, 51 (2004), 1057-1061.

19. Bredensteiner, E. and Bennet, K. Multicategory classification by support vector machines. *Computational Optimizations and Applications*, 12 (1999), 53-79.

20. Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 2 (1998), 121-167.

21. Castleman, K. *Digital Image Processing*. Prentice Hall, Inc., Upper Saddle River, 1996.

22. Cantrell, C. D., *Modern Mathematical Methods for Physicist and Engineers*. Cambridge University Press, Cambridge, UK, 2000.

23. Caton, R. Electrical Currents of the Brain. *Chicago Journal of Nervous and Mental Disease*, 4 (Oct 1875), 610. available online through http://journals.lww.com. Accessed on June 10, 2011.

24. Christoforou, C., Haralick, R., Sajda, P., and Parra, L. C. The bilinear brain, subject-invariant bilinear discriminant. In *4th International Symposium on Communications, Control and Signal Processing (ISCCSP)* (Cyprus, Mar 3-5) IEEE, New York, 2010 [Onlineshttp://rkileaders.com/rkibussiness/ images/stories/documents/ ChristoforosChristoforouDissertation.pdf, accessed on June 10, 2011

25. Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1960), 37-46.

26. Cooley, J. W. and Tukey, J. W. An algorithm for the machine computation of the complex Fourier series. *Mathematics of Computation*, 19 (1965), 297-301.

27. Crammer, K. and Singer, Y. On the algorithmic implementation of multi-class kernel-based vector machines. *Journal Machine Learning Research*, 2 (2001), 265-292.

28. Curran, E., Skyacek, P., Stokes, M., Roberts, S. J., Penny, W., Johnsrude, I., and Owen, A. M. Cognitive tasks for driving a brain-computer interfacing system: a pilot study. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 12, 1 (Mar 2003), 48-54.

29. Curran, E. A. and Stokes, M. J. Learning to control brain activity: a review of the production and control of EEG components for

driving brain-computer interface (BCI) systems. *Brain Cogn.*, 51, (2003), 326-336.

30. Dal Sarno, B., Matteucci, M., and Mainardi, L. T. The utility metric: a novel method to assess the overall performance of discrete Brain-Computer Interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 18, 1 (2010), 20-28.

31. Debauchies, I. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

32. Decety, J. The neurophysical basis of motor imagery. *Behavioral Brain Research*, 77 (1996), 45-52.

33. Dornhege, G. *Increasing Information Transfer Rates for Brain-Computer Interfacing*. Ph.D. Dissertation, University of Potsdam, Germany. 2006.

34. Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, Inc., Hoboken, 2001.

35. Faradji, F., Ward, R. K., and Birch, G. A Self-Paced BCI Using Stationary Wavelet Packets. In *Proceedings of the 31st Annual International Conference of the IEEE EMBS* (Minneapolis, Mn., Sep 2-6 ), IEEE, New York, 2009, pp. 962-965.

36. Farwell, L. A. and Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70, 6 (1988), 510-523.

37. Fatourechi, M., Birch, G. E., and Ward, R. K. A self-paced brain interface system that uses movement related potentials and changes in the power of the brain rhythms. *Journal of Computational Neuroscience*, 23, 1 (2007), 21-37.

38. Fatourechi, M., Birch, G. E., and Ward, R. K. Applying a hybrid genetic algorithm in the design of a self-paced brain interface with a low false positive rate. *Proceedings of the IEEE International on Acoustics, Speech and Signal Processing*, 4 (2007), IV-1157, IV-1160.

39. Fatourechi, M., Mason, S. G., Birch, G. E., and Ward, R. K. Is information transfer rate a suitable performance measure for self-paced brain interface systems? In *International Symposium on Signal*

*Processing and Information Technology* (Vancouver, BC., Aug 27-30), IEEE, New York, 2006, pp. 212-216.

40. Ferri, F. J., Pudil, P., Hatef, M., and Kittler, J. Comparative Study of Techniques for Large-Scale Feature Selection. In Gelsema, E. S. and Kanal, L., eds., *Pattern Recognition in Practice IV*. Elsevier, 1994.

41. Franc, V. and Hlaváč, V. Multi-class Support Vector Machine. In *Proceedings 16th International Conference on Pattern Recognition* (Quebec City, Que., Aug. 11-15) IEEE Computer Press, Los Alamitos, CA, 2002, pp. 236-239.

42. Freeman, W. J. *Mass Action in the Nervous System*. Academic Press, New York, 1975.

43. Fukunaga, K. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, 1990.

44. Gabor, D. Theory of Communication. *Journal of Institute of Electrical Engineers*, 93 (1946), 429-457.

45. Geng, T., Gan, J. Q., Dyson, M., Tsui, C., and Sepulveda, F. A novel design of 4-class BCI using two binary classifiers and parallel mental tasks. *Computational Intelligence and Neuroscience* (2008). published online 2008 June 22. doi: 10.1155/2008437306. Accessed on June 10, 2011.

46. Gonzales, R. C. and Woods, R. E. *Digital Image Processing* Prentice-Hall, Upper Saddle River, NJ, 2008.

47. Graimann, B., Huggins, J. E., Levine, S. P., and Pfurtscheller, G. Toward a direct brain-computer interface based on human subdural recordings and wavelet-packet analysis. *IEEE Trans. Biomedical Eng.*, 51, 6 (2004), 954-962.

48. Graimann, G., Pfurtscheller, G., and Townsend, G. A comparison of common spatial patterns with complex band power features in a four-class BCI experiment. *IEEE Trans. Biomedical Eng.*, 53, 4 (2006), 642-651.

49. Grossman, A. and Morlet, J. Decomposition of Hardy functions into square integrable wavelets of constant shape. *Journal of Mathematical Analysis*, 15 (1984), 723-736.

50. Gwet, K. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliablity Assessment* , 2 (2002), 1-9.

51. Hamming, R. W. *Coding and Information Theory*. Prentice Hall, 1980. Upper Saddle River, NJ.

52. Hasan, B. A. S., Dyson, M., Balli, T., and Gan, J. Q. A study via feature selection on the separability of approximate entropy for brain-computer interfaces. In *The UK Workshop on Computational Intelligence* (Demontfort, UK Sep 10-12). UKWCI, 2008, pp. 189-194.

53. Hasan, B.A.S., Gan, J. Q., and Zhang, Q. Multi-objective evolutionary methods for channel selection in brain-computer interfaces: some preliminary experimental results. *IEEE World Congress on Evolutionary Computation* (Barcelona, Sp. July 18-23), IEEE, New York, 2010, pp. 1-6.

54. Hess-Nielsen, N. and Wickerhauser, M. V. Wavelets and Time-Frequency Analysis. *IEEE Proceedings*, 84, 4 (1996), 523-540.

55. Hinterberger, T., Mellinger, J., and Birbaumer, N. The Thought Translation Device: Structure of a multimodal brain-computer communication system. In *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering* (Capri Island, Italy, Mar 20-22), IEEE, New York, 2007, pp. 603-606.

56. Hinterberger, T., Nijboer, F., Kübler, A. et al. Brain Computer Interfaces for Communication in Paralysis: A Clinical Experimental Approach. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2004.

57. Hsu, C. and Lin, C. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Netwk.*, 13 (2002), 415-425.

58. Huggins, J. E., Levine, S. P., BeMent, S. L. et al. Detection of event-related potentials for development of a direct brain interface. *Journal of Clinical Neurophysiology*, 16, 5 (1999), 448-455.

59. Jaffard, S., Meyer, Y., and Ryan, R. D. *Wavelets: Tools for Science and Technology*. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

60. Jain, A. K., Duin, R. P. W., and Mao, J. Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22 (2000), 4-37.

61. Jain, A. and Zongker, D. Feature Selection: Evaluation, application and small sample performance. *IEEE Trans. Pattern Recog.*, 19, 2 (1997), 153-158.

62. Kübler, A. and Müller, K.-R. An Introduction to Brain-Computer Interfacing. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2007.

63. Keirn, Z. A. and Aunon, J. I. A new mode of communication between man and his surroundings. *IEEE Trans. Biomed. Eng.*, 37, 12 (Dec 1990), 1209-1214.

64. Klem, G. H., Lüders, H. O., Jasper, H. H., and Elger, C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*, 52 (1999), 3-6.

65. Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, 97, 1 (1997), 273-324.

66. Kuncheva, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, Inc., Hoboken, 2004.

67. Lee, Y., Lin, Y., and Wahba, G. *Multicategory support vector machines*. University of Wisconsin, Madison, WI, 2001.

68. Lodder, S. *Single-Trial Classification of an EEG-Based Brain Computer Interface Using the Wavelet Packet Decomposition and Cepstral Analysis*. 2009. Available online at http://hdl.handle.net/ 10019.1/2791. Accessed on June 10, 2011.

69. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, f., and Arnaldi, B. A review of the classification alogorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 4, 2 (2007), R1-R13.

70. Lotte, F. and Guan, C.T. Learning from other subjects helps reducing brain-computer interface calibration time. In *International Conference on Audio, Speech and Signal Processing (ICASSP)* (Dallas, Tx., Mar 14-19), IEEE, New York, 2010, pp. 614-617.

71. Lotte, F., Guan, C. T., and Ang, K. K. Comparison of designs towards a subject-independent brain computer interface based on motor imagery. In *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Minneapolis, Mn., Sep. 2-6), IEEE, New York, 2009, pp. 4543-4546.

72. Müller, K.-R., Anderson, C. W., and Birch, G. E. Linear and nonlinear methods for brain-computer Interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 11, 2 (2003), 165-169.

73. Müller-Gerking, J., Pfurtscheller, G., and Flyvbjerg, H. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, 110, 5 (1999), 787-798.

74. Mallat, S. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1999.

75. Mason, S. G., Bashashanti, A., Fatourechi, M., Navarro, K. F., and Birch, G. E. A comprehensive survey of brain interface technology designs. *Annals of Biomedical Engineering*, 35, 2 (2007), 137-169.

76. Mason, S. G. and Birch, G. E. A general framework for brain-computer Interface design. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 11 (March 2003), 72-87.

77. Mason, S. G., Kronegg, J., Huggins, J., Fatourechi, M., and Schlögl, A. *Evaluating the performance of self-paced BCI technology*. 2006. Technical Report, available online: http://www.bci-info.turgraz.at/Research Info/documents/articles/ self paced tech report-2006-05-19.pdf. Accessed on June 10, 2011.

78. McFarland, D. J., Anderson, C. W., Müller, K.-R., Schlögl, A., and Krusienski, D. J. BCI Meeting 2005 - Workshop on BCI Signal Processing: Feature Extraction and Translation. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 14, 2 (2006), 135-138.

79. Millán, J. del R., Ferrez, P., Buttfield, A., The IDIAP Brain-Computer Interface: An Asynchronous Multiclass Approach. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2004.

80. Millet, D. Hans Berger: from psychic engergy to EEG. *Perspectives in Biology and Medicine*, 44 (2001), 522-542.

81. Mix, D. F. and Olejniczak, K. J. *Elements of Wavelets for Engineers and Scientists*. Wiley-Interscience, Inc., Hoboken, 2003.

82. Ngatchou, P., Zarei, A., and El-Sharkawi, M. A. Pareto multi objective optimization. In *Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems* (Arlington, Va., Nov. 6-10), IEEE, New York, 2005, pp. 84-91.

83. Ngatchou, P. N., Zarei, A., Fox, W. L. J., and El-Sharkawi, M. A. Pareto multiobjective optimization. In *Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems*. John Wiley and Sons, Inc., Hoboken, 2007.

84. Niedermeyer, E. Historical Aspects. In *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott, Williams and Wilkins, Philadelphia, 2005.

85. Nikulin, V., Hohlefel, F. U., Jacobsc, A. M., and Curio, G. Quasi-movements: A novel motor-cognitive phenomenon. *Neuropsychologia*, 46, 2 (2008), 727-742.

86. Nolte, J. *The Human Brain: An Introduction to its Functional Anatomy*. Mosby, Inc., St. Louis, 2002.

87. Nykopp, T. Statistical modelling issues for the adaptive brain interface. 2001. Master's thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering.

88. Obermaier, B., Guger, C., Neuper, C., and Pfurtscheller, G. Hidden markov models for online classification of single trial EEG. *Pattern recognition letters* (2001), 1299-1309.

89. Obermaier, B., Neuper, C., Guger, C., and Pfurtscheller, G. Information Transfer rate in a five-classes brain-computer interface. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 9 (2000), 283-288.

90. Oppenheim, A. V. and Schafer, R. W. *Digital Signal Processing*. Prentice Hall, 1975, Upper Saddle River, NJ.

91. Pfurtscheller, G., Flotzinger, D., and Neuper, C. Differentiation between finger, toe and tongue movement in man based on 40 Hz EEG.

*Electroencephalography and clinical Neurophysiology*, 90 (1994), 456-460.

92. Pfurtscheller, G. and Lopes da Silva, F. H. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110 (1999), 1842-1857.

93. Pfurtscheller, G., Müller-Putz, G. R., Schlögl, A. et al. Graz-Brain-Computer Interface: State of Research. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2004.

94. Pfurtscheller, G. and Neuper, C. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89, 7 (July 2001), 1123-1134.

95. Pfurtscheller, G., Neuper, C., Flotzinger, D., and Pregenzer, M. EEG-based discrimination between imagination of right and left hand movement. *Electroencephalography and Clincal Neurophysiology* , 103 (1997), 642-651.

96. Pfurtscheller, G., Neuper, C., Schlögl, A., and Lugger, K. Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. *IEEE Trans. Rehabil. Eng.*, 6, 3 (1998), 316-325.

97. Picard, R., Vyzas, E., and Healey, J. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23, 10 (2001), 1175-1191.

98. Principe, J. C., Euliano, N. R., and Lefebvre, W. C. *Neural and Adaptive Systems: Fundamentals through Simulations*. John Wiley and Sons, Inc., New York, 2000.

99. Proakis, J. G. and Manolakis, D. G. *Digital Signal Processing. Principles, Algorithms, and Applications.* Prentice Hall, 1996

100. Pudil, P., Novovicovà, J., and Kittler, J. Floating search methods in feature selection. *Pattern Recognition Letters*, 15, 11 (1994), 1119-1125.

101. Ramoser, H., Müller-Gerking, J., and Pfurtscheller, G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8, 4 (2000), 441-446.

102. Rifkin, R. and Klatau, A. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5 (2004), 101-141.

103. Rioul, O. and Vitterli, M. Wavelets and Signal Processing. *IEEE Signal Processing Magazine* (Oct 1991), 14-38.

104. Rowan, A. J. and Tolunsky, E. *Primer of EEG: With a Mini-Atlas*. Elsevier Science, Philadelphia, 2003.

105. Sörnmo, L. and Laguna, P. *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Elsevier Academic Press, Burlington, 2005.

106. Salvetti, A. and Wiamowski, B. M. A brain-computer interface for recognizing brain activity. In *2008 Conference on Human System Interactions* (Krakow, May 25-27), IEEE, New York, 2008, pp. 714-719.

107. Sanei, S. and Chambers, J. A. *EEG Signal Processing*. John Wiley and Sons, Ltd., West Sussex, 2007.

108. Sarkar, T. K. and Su, C. A Tutorial on Wavelets from an Electrical Engineering Perspective, Part : Discrete Wavelet Techniques. *IEEE Trans. Antennas Propagat. Mag.*, 40, 5 (Oct 1991), 49-70.

109. Sarkar, T. K. and Su, C. A Tutorial on Wavelets from an Electrical Engineering Perspective, Part 2: The Continuous Case. *IEEE Antennas Propagat. Mag.*, 40, 6 (Dec 1998), 36-49.

110. Schlögl, A., Keinrath, C., Scherer, R., and Pfurtscheller, G. Information transfer of an EEG-based brain computer interface. In *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering* (Capri Island, Italy, Mar 20-22), IEEE EMBS, New York 2003, pp. 641-644.

111. Schlögl, A., Kronegg, J., Huggins, J. E., and Mason, S. G. Evaluation Criteria for BCI Research. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2007.

112. Schlögl, A., Lee, F., Bischoff, H., and Pfurtscheller, G. Characterization of the four class motor imagery EEG data for the BCI Competition., *J. Neural Eng.,*2, 4, (2005), L14-22.

113. Sellers, E. W., Krusienski, D. J., McFarland, D. J., and Wolpaw, J. R. Noninvasive Brain-Computer Interface Research at the

Wadsworth Center. In Dornhege, G. et al., eds., *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, 2007.

114. Seo, N. *A comparison of Multi-class support vector machine methods for Face Recognition*. University of Maryland, 2007. available online at http://note.sonots.com/. Accessed on June 10, 2011.

115. Serrien, D. J., Ivry, R. B., and Swinnen, S. P. Dynamics of hemispheric specialization and intergration in the context of motor control. *Nature*, 7 (Feb 2006), 160-167.

116. Shannon, C. E. and Weaver, W. *The mathematical theory of communication*. University of Illinois Press, Urbana, 1949.

117. Sheng, Y. Wavelet Transform. In Poularikas, A. D., ed., *The Transforms and Applications Handbook*. CRC Press, Boca Raton, 1996.

118. Sherwood, J. and Derakhshani, R. An ensemble method for quasi movement, subject-independent brain computer interfaces. *IST Transactions on Biomedical Sciences and Engineering* (2010). available at http://www.istpress.com/download/13-10-001_Sherwood.pdf.html. Accessed on June 10, 2011.

119. Sherwood, J. and Derakhshani, R. Derivation of a quality metric to improve convergence of wrapper method search for multiclass subject invariant brain computer interfaces, unpublished manuscript, 2011

120. Sherwood, J. and Derakhshani, R. On classifiability of wavelet features for EEG-based brain-computer interfaces. In *International Joint conference on Neural Networks* (Atlanta, Ga., Jun 14-19), INNS, Madison, Wi.,2009, pp. 2895-2902.

121. Steriade, M., Gloor, P., Llinàs, R. R., Lopes da Silva, F. H., and Mesulam, M.-M. Basic mechanisms of cerebral rhythmic activities. *Electroencephalography and clinical Neurophysiology*, 76 (1990), 481-508.

122. Strang, G. and Nguyen, T. *Wavelets and Filter Banks*. Wellesley College, Wellesley, 1996.

123. Swartz, B.E. and Goldenshon, E.S., Timeline of the history of EEG and associated fields, *Electroencephalography and Clinical Neurophysiology*, 106, 2 (1998), 173-176.

124. Theodoridis, S. and Koutroumbas, K. *Pattern Recognition*. Academic Press, Burlington, 2009.

125. Ting, W., Guo-zheng, Y., Bahg-hua, Y., and Hong, S. EEG feature extraction based on wavelet packet decomposition for brain computer interface. *Measurement*, 41 (2008), 618-625.

126. Townsend, G., Graimann, B., Pfurtscheller, G., A comparison of common spatial patterns with complex band power features in a four-class BCI experiment, IEEE Trans. Biomed. Eng., 53, 4, (2006), 642-651.

127. Vapnik, V. N. *Statistical Learning Theory*. Wiley-Interscience, New York, 1989.

128. Vapnik, V. M. The Nature of Statistical Learning. In *Statistics for Engineering and Information Science*. Springer-Verlag, New York, 2000.

129. Vaughn, T. M., McFarland, D. J., Schalk, G., Sarnacki, W. A., Krusienski, D. J., Sellers, E. W., and Wolpaw, J. R. The Wadsworth BCI Research and Development Program: At Home with BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 14, 2 (June 2006), 229-233.

130. Vetterli, M. and Herley, C. Wavelets and Filter Banks: Theory and Design. *IEEE Trans. Signal Process.*, 40, 9 (Sep 1992), 2207-2232.

131. Vidal, J. Toward direct brain-computer communication. *Annual Review of Biophysics and Bioengineering*, 2 (1973), 157-180.

132. Wallach, H. *Evaluation metrics for hard classifiers*. University of Cambridge, Cambridge, UK, 2006. Available online at http://www.inference.phy.cam.ac.uk/hmw26/papers /evaluation.ps accessed on July 14, 2011.

133. Wang, B., Jun, L., Bai, J., Li, G., and Li, Y. EEG recognition based on multiple types of information by using wavelet packet transform and neural networks. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* (Shanghai, China, Sep 1-4), IEEE, New York, 2005, pp. 5377-5380.

134. Welch, P. D. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short,

modified periodograms. *IEEE Trans. Audio Electroacoustics* (1967), 70-73.

135. Weston, J. and Watkins, C. *Multiclass support vector machines, Technical Report CSD-TR-98-04*. Royal Holloway, University of London, Department of Computer Science, 1999. available online at http://citeseerx.ist.psu.edu doi=10.1.1.50.9594.ps. accessed on June 10, 2011.

136. Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. Brain-computer interfaces for communication and control. *Clinical Neurophysicology*, 113, 5 (2002), 767-79

137. Zhiwei, L. and Minfen, S. Classification of mental task EEG signals using wavelet packet entropy and SVM. In *The Eighth International Conference on Electronic Measurement and Instruments* (Xian, China, Aug 16-18)IEEE, New York, 2007, pp. 3906-3909.

# VITA

Jesse Sherwood was born on October 15, 1954, in Saint Joseph, Missouri. He was educated in local public schools and graduated from Central High School in 1972. He received the Whitaker Foundation scholarship that year. He attended the University of Missouri at Kansas City, Missouri Western State College in St. Joseph, Missouri and the University of Missouri at Rolla, from which he graduated in 1978. His degree was a Bachelor of Science in Electrical Engineering.

Mr. Sherwood worked as an engineer and engineering manager in the radio and television broadcasting industry and as an engineering researcher and network planner for United Telecom and Sprint in Kansas City, Missouri, where he served as Principal Network Engineer. During this period, Mr. Sherwood began the Master's degree program in electrical and computer engineering at the University of Missouri - Columbia. He was awarded the Masters of Science degree in Electrical and Computer Engineering in December, 1985. He attended Rockhurst College in Kansas City, Missouri where he was awarded the Master of Business Administration degree in marketing and finance in December 1988.

In 1990, Mr. Sherwood was a founder of and Chief Technical Officer for ITN, later known as Illuminet, which became the largest independent telecommunications signaling network in the world. Mr. Sherwood, briefly worked for Agilent Technologies and Tekelec. He began work toward his interdisciplinary Ph.D. in Electrical and Computer Engineering and Telecommunications and Computer Networking at the University of Missouri-Kansas City in the Fall of 2007. He was a Chancellor's Doctoral Fellow for 2009-2011, and received the UMKC School of Computing and Engineering

Dean’s Outstanding interdisciplinary PhD student award for 2010-2011 school year. Upon completion of his degree requirements he plans to teach and pursue research interests.

Mr. Sherwood is a member of the Institute of Electrical and Electronics Engineers, Association of Computing Machinery, Missouri Society of Professional Engineers, Kansas Society of Professional Engineers, Omicron Delta Kappa, Eta Kappa Nu, Tau Beta Pi, Epsilon Phi Upsilon and the National Honor Society. He is a registered professional engineer in Missouri and Kansas.