ADDRESSING NONSTATIONARITY IN EEG APPLICATIONS

by

Matthew Sybeldon

B.S. in Electrical Engineering, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science in Electrical Engineering

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Matthew Sybeldon

It was defended on

November 8, 2016

and approved by

Murat Akcakaya, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Ervin Sejdic, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Zhi-Hong Mao, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Thesis Advisor: Murat Akcakaya, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

ADDRESSING NONSTATIONARITY IN EEG APPLICATIONS

Matthew Sybeldon, M.S.

University of Pittsburgh, 2016

Electroencephalography (EEG) is a measure of electrical activity from the brain that is used for numerous applications. EEG can be used for detecting the onset of epilepsy in a patient. It can also allow a disabled person to operate a computer using a brain computer interface (BCI). However, EEG measurements are usually interpreted as being generated from a random process. This random process is not stationary, meaning that the distribution governing the measurements can change over time. In seizure detection, this nonstationarity may indicate a seizure. In BCI applications, nonstationarity is problematic due to the invalidation of previously trained classifiers. This thesis provides techniques to both leverage and mitigate nonstationarity depending on the application. For seizure detection, a statistical detector requiring no previously labelled data or expert knowledge is outlined to detect seizures in the EEG. This detector achieved an average seizure prediction time of 70 seconds. In BCI applications, a combination of mutual information and ensemble learning is used to identify previously learned data most similar to incoming data and reduce calibration requirements. It was shown that this learning scheme provided adequate results for typical participants and outperformed state of the art techniques for steady state visual evoked potential BCIs for participants whose EEG violated typical assumptions.

Keywords: Electoencephalography, Nonstationarity, Seizure Detection, Brain Computer Interface, Transfer Learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my advisor, Dr. Murat Akcakaya, for giving me the opportunity and support necessary to conduct research as an undergraduate student, which eventually led to my graduate studies. I would also like to thank Dr. Ervin Sejdic and Dr. Zhi-Hong Mao for being on my Master's thesis committee.

I would also like to thank Dr. Arun Antony, Dr. Anton Bagic, and the rest of the UPMC epilepsy division for providing us with real epilepsy EEG data necessary to test our methods and for their continued support.

Last but not least, I would also like to thank my family back home in Chicago and my friends here in Pittsburgh for being there for me through thick and thin.

1.0 INTRODUCTION

1.1 BACKGROUND - ELECTROENCEPHALOGRAPHY

Electroencephalography (EEG) is a measurement of electrical activity from the brain. The foundations of EEG can be traced back to the late nineteenth century where Richard Caton discovered that animal brains exhibited electrical phenomena [1]. In 1890, Adolf Beck found that specific electrical rhythms could be induced by flickering light stimuli [1].

Research into the electrical activity of the brain eventually culminated in Hans Berger's invention and first recording of human EEG in 1924 [1]. Numerous applications have been devised in the time since. For example, epilepsy has been closely associated with abnormal spikes in EEG recordings, thus promoting the possibility of EEG as a diagnosis tool [1] [2] [3].

EEG has also been used as an interface for computers [4]. This is a particular kind of brain computer interface (BCI), a type of device used to operate a computer without the traditional physical requirements of a keyboard or mouse. EEG is a popular input modality for BCIs due to its relative cost compared to other input signals such as functional magnetic resonance imaging (fMRI) and higher relative speed due to its temporal resolution [1].

1.2 EEG AS A NONSTATIONARY RANDOM PROCESS

Like many biological measurements, EEG is governed by a random process. A random process governs the distribution of measurements. Assume there an EEG measurement at time t is denoted by x^t. A random process places a distribution over a collection of n measurements $f(x_1^t, \ldots, x_n^t)$. One key property of a random process is its stationarity. If a

random process is stationary, then the distribution of a series of measurements at one time does not change at a later time. This means for all T, the following statement holds true:

$$f(x_1^t, \ldots, x_n^t) = f(x_1^{t+T}, \ldots, x_n^{t+T}) \tag{1.1}$$

The EEG random process is assumed to be nonstationary [4]. This has implications about any previously learned distributions obtained from collected EEG data. For the case of seizure detection, the distributions of EEG measurements are almost assuredly different for a seizure than they are for typical EEG. In BCI applications, machine learning classifiers are often used to infer intent from collected EEG. If the distributions change, then any previously learned thresholds no longer hold and new data must be collected. In one of these applications, the nonstationarity of the EEG aids can aid in the task of finding when a patient is undergoing a seizure. However in BCIs, nonstationarity plays an adversarial role.

1.3 CONTRIBUTIONS

In this thesis, two techniques are proposed to take advantage of and mitigate nonstationarity respectively. In the first part, a statistical seizure detector is described that uses random graph theory and the extended CUmulative SUM (CUSUM) test to find abnormalities in the EEG that may indicate a seizure. In the second part, a technique for identifying a given distribution's similarity with previously learned data is also given. This effectively allows the identification of changing EEG distributions if they resemble previously collected data. This technique is demonstrated within a steady state visual evoked potential (SSVEP) BCI in combination with ensemble learning to take advantage of data sets collected from multiple people and multiple days toward incoming data.

2.0 SEIZURE DETECTION USING EEG AND RANDOM GRAPH THEORY

2.1 INTRODUCTION

Seizure detection is an application where EEG nonstationarity can be used in order to find when a seizure has occurred. Under typical circumstances, the distributions of EEG measurements can be assumed to be approximately stationary. By drawing on previously established techniques to detect changes in distributions, a seizure could be detected without previously observing the distributions corresponding to seizure. Such a paradigm is preferable over the traditional machine learning approaches where previously observed seizure instances are required.

About 22 million people are diagnosed with some form of epilepsy worldwide. Of this subset, approximately one third have seizures that are not addressable by antiepileptic drugs [2]. For these patients, surgery may be a viable treatment option. However, careful observation and diagnosis is required before proceeding with such a critical treatment. Seizure monitoring is a major step in identification and treatment formulation for surgical candidates [3]. This monitoring step can involve long hospitalization periods in which the patient is observed with simultaneous EEG and video recordings. The duration of monitoring often lasts days or weeks, which creates a large amount of data that needs to be reviewed manually by a neurologist in order to determine if and how further treatment should proceed [3]. The neurologist's difficulty in identifying seizures can increase as the amount of data to review grows. In addition, medical staff must remain vigilant in order to respond to a seizure. In these timely operations, an automated EEG seizure detector would improve the quality of treatment.

2.1.1 Technological Need

An EEG seizure detector needs several qualities in order to be robust to the individual EEG characteristics of each patient:

1.) *Low training requirements* - For any given patient undergoing seizure monitoring, it must be assumed that previous data is unavailable. This limits the application of many pattern recognition techniques such as quadratic discriminant analysis or support vector machine classifiers that require previous data and expert-supplied labels. Statistical methods for detecting outliers assumed to be associated with seizures would be more appropriate. In this case, if a short period of collected data can be assumed to correspond to typical EEG, the region of a feature space corresponding to a seizure can be derived based on statistical methods.

2.) *Minimal assumptions* - EEG data is highly nonstationary between patients, and the personal nature of epilepsy adds further complications. The method should make as few assumptions about the nature of the seizure as possible. Information such as seizure activity location or EEG frequency bands is assumed to be unavailable for new patients, and the detector must select a statistic to reflect the lack of assumptions.

3.) *Low detection delay* - The time available to respond to a seizure must be maximized to reduce the medical staff's burden in treatment and increase their chances of successful intervention.

4.) *High sensitivity and specificity* - The detector should miss as few seizures as possible. Conversely, the detector should refrain from throwing an alarm in the absence of a seizure, thus lessening the credibility of the detector. However if faced between a tradeoff of sensitivity and specificity, the detector should favor sensitivity due to the higher cost associated with a missed seizure [5].

5.) *Robust to artifacts* - The probability of EEG artifacts due to certain types of movement is high, especially over long stays. Other artifacts due to electrode impedance may be present. A seizure detector must not respond to these artifacts to maintain the integrity of the alarm.

2.1.2 Previous Work

Much work has already been done in the fields of seizure detection and prediction. Many efforts are focused toward the application of machine learning techniques [6] [7] [8] [9] [10] [11] [10] [12] [13] [14]. Support Vector Machines (SVM) have proven to be a popular approach with various frequency feature selections with reported accuracies reaching as high at 97-98.5 percent [6] [7] [8]. Neural networks have also been utilized with a sensitivity as high as 89 percent using a variety of features from the time domain and frequency domain [9] [11] [10] [12]. Other classifiers using quadratic disriminant analysis (QDA) and nearest-neighbor decision rules have been used to achieve high sensitivities (96.59 % and above) [13] [14].

Machine learning applications are difficult to translate into practice since seizures can vary wildly across patients or even within the same patient [2]. In general, machine learning violates the low training requirements needed by a clinical seizure detector. Sufficient amounts of labelled data are required, which is usually not the case for a given patient.

One study recognized the shortcomings of machine learning applications and attempted to use a statistical approach instead. This was accomplished through the use of the CUSUM test on entropy measures of the EEG signal [15]. This statistical approach yielded lower sensitivities than the pattern recognition studies, but at the benefit of requiring no training data in a formal sense. However, the frequency content of the ictal period is needed which is patient specific. There are also eight different parameters to set, several of which are patient specific. Reducing the amount of parameters needed to detect a seizure would be a necessary step for a practical detector. However, viewing the seizure detection problem as a statistical problem seems promising due to the data availability constraints. As such, this paper is focused on further developing the statistical interpretation of the seizure detection problem.

2.1.3 Contribution

In this section, a seizure detection method is proposed using a combination of partial directed coherence, random graph theory, and the CUSUM test to monitor changes in the EEG that may correspond to a seizure. This allows seizure detection to be done using no previous training data and fewer tuning parameters. In section 2.2, the details of the seizure detector

are given. Section 2.3 then explains the tests used to evaluate the detector. Section 2.4 details the results obtained on the patient data supplied. The last section evaluates the results and outlines the direction in which future work may head.

2.2 METHODS

Seizures are generally understood to be highly correlated between various channels, and the proposed seizure detector attempts to exploit this feature. Granger causality is a concept originating in economics that indicates whether one time series can effectively predict another time series [16]. This general concept is applied toward EEG, which has multiple time series corresponding to the multiple EEG channels. One of the earliest indicators of Granger causality is coherence $C(\omega)$, which is defined below [16]. Here, $Cr(\omega)$ is the power cross spectrum, and $f_x(\omega)$ and $f_y(\omega)$ are the power spectral densities of two time series x and y. The magnitude of the coherence can be interpreted as the square of the correlation coefficient between frequency components in x and y. The phase can be interpreted as the time delay. However, coherence lacks a directionality component that limits its application toward seizure detection purposes.

$$C(\omega) = \frac{|Cr(\omega)|^2}{f_x(\omega) f_y(\omega)} \tag{2.1}$$

Other measures of Granger causality have been proposed. Of particular interest is partial directed coherence (PDC), which is related but offers information about direction of causality [17]. The PDC factor from time series i to j is given by [17]:

$$\boldsymbol{\pi}(f) = \begin{bmatrix} \pi_{11}(f) & \dots & \pi_{1k}(f) \\ \pi_{21}(f) & \dots & \pi_{2k}(f) \\ \vdots & \ddots & \vdots \\ \pi_{k1}(f) & \dots & \pi_{kk}(f) \end{bmatrix} \tag{2.2}$$

$$\pi_{ij}(f) = \frac{\boldsymbol{A}_{ij}(f)}{\sqrt{\boldsymbol{a}_j^H(f)\Sigma^{-1}\boldsymbol{a}_j(f)}} \tag{2.3}$$

$$\boldsymbol{B}_r(n) = \begin{bmatrix} b_{11}(n) & \dots & b_{1k}(n) \\ b_{21}(n) & \dots & b_{2k}(n) \\ \vdots & \ddots & \vdots \\ b_{d1}(n) & \dots & b_{dk}(n) \end{bmatrix} \tag{2.4}$$

$$\boldsymbol{A}(f) = \boldsymbol{I} - \boldsymbol{B}(f) \tag{2.5}$$

$$\boldsymbol{A}(f) = \begin{bmatrix} \boldsymbol{a}_1(f) & \dots & \boldsymbol{a}_k(f) \end{bmatrix} \tag{2.6}$$

$$\boldsymbol{a}_j(f) = \begin{bmatrix} a_{1k}(f) \\ \vdots \\ a_{dk}(f) \end{bmatrix} \tag{2.7}$$

where $\boldsymbol{b}_{ij}(f)$ is the coefficient obtained from the Fourier transform of the multivariate autoregressive (MVAR) model between the two time series for a specific frequency f, Σ is the covariance of the cross spectral density matrix, and k is the number of EEG channels.

The strength of a measure such as PDC is apparent in its formulation because it is normalized according to the destination. This means that the amplitude effects of the seizure are mitigated.

Calculation of the partial directed coherence values leads to a large matrix in order to describe the connectivity between channels, as shown in (2.2). In addition, individual PDC values for specific frequencies, $\pi_{ij}(f)$, are random variables likely with a high variance. A seizure detector must consider a large collection of PDC matrices for a large range of frequency values in a manner robust to the random nature of the individual values.This can be accomplished using random graphs to model the collection of EEG channels.

2.2.1 Random Graphs

A graph C is a set of objects (nodes) $O_i, i = 1, ..., k$ $\{O_1, O_2, ..., O_k\} \in C$ and the connection (edges) $E_{ij} \in \{0, 1\}$ from node O_i to O_j. The degree d_j of node j indicates the amount of edges connected to the node. The degree can be calculated by a simple summation:

$$d_j = \sum_{i=1}^{k} E_{ij} \tag{2.8}$$

Graphs may be directed or undirected depending if an edge can exist from one node to another but possibly not in the reverse direction. In a directed graph, $E_{ij} \neq E_{ji}$ in general. A random graph is where an edge or node exists with a given probability p_{ij}. If p_{ij} is constant, then the degree distribution for a given node can be approximated by a Poisson distribution [18] [19]. Random graphs can grow or shrink, but for the purposes of seizure detection, as described below the number of nodes is static. Information can be obtained by analyzing the edges in the graph. An adaptive detector cannot use previous patient specific information, so assumptions about specific EEG channels are difficult to make. Instead, statistics obtained from the edges can be calculated and used for detection purposes. Next we describe how we obtain a graph model from the calculated PDC matrices.

For every node O_i in the random graph C, there is an associated degree d_i. A histogram of the node degrees can be constructed in order to generate statistics. Let $\boldsymbol{E}(f, t)$ be the edge matrix obtained from thresholding $\boldsymbol{\pi}(f, t)$ into a binary valued matrix. Summing across the rows of $\boldsymbol{E}(f, t)$ yields a degree vector $\boldsymbol{d}(f, t)$ with the i-th component being $d_i(f, t)$ as

$$\boldsymbol{d}(f, t) = \begin{bmatrix} d_1(f, t) \\ d_2(f, t) \\ \vdots \\ d_k(f, t) \end{bmatrix}. \tag{2.9}$$

With no prior information on the frequency content of the seizures, we make no assumptions as to which bands contain the relevant information. The average degree $\bar{\boldsymbol{d}}(t)$ across the frequency bands is calculated.

$$\bar{\boldsymbol{d}}(t) = \begin{bmatrix} \bar{d}_1(t) \\ \bar{d}_2(t) \\ \vdots \\ \bar{d}_k(t) \end{bmatrix} \tag{2.10}$$

If $\bar{\boldsymbol{d}}(t)$ is of a low enough dimension for the application, then $\bar{\boldsymbol{d}}(t)$ can be used directly for the detector, but this was not tested. A further averaging operation across all the channels can be done to obtain a scalar $\hat{d}(t)$. Regardless of the statistic used, it is assumed that the large amount of summations and averaging operations allows the distribution of the statistic to approach normality under the central limit theorem. This has several attractive properties that allow the statistic to be used in the detector. These are explored in the next section.

2.2.2 Detection Method - Cumulative Sum

Since the average degree is a random variable, it is governed by a random distribution. If an assumption is made that the distributions corresponding to normal EEG and seizure EEG are different, then detecting changes in the distribution allows seizures to be detected. Assume that the change in distribution occurs at a deterministic but unknown time. Minimizing the supremum of the detection delay while constraining the mean delay between false alarms yields an optimization problem whose solution will provide the foundation for the detector. Mathematically speaking, the problem can be stated as [20] [21]

$$\begin{aligned} \text{argmin}_{\tau} \sup_{n \geq 1} \text{ess sup} \quad & E_n \left[(\tau - n)^+ | \mathcal{D}_T \right] \\ \text{such that} \quad & E_\infty[\tau] \geq \alpha \end{aligned} \tag{2.11}$$

Here, $\mathcal{D}_T = \left[\hat{d}(1), \ \ldots, \ \hat{d}(T) \right]$ is the observed data, where $\hat{d}(t)$ is the calculated mean degree for time t; τ is the stopping time; $x^+ = \max(0, x)$; $E_\infty[\tau]$ is the mean time between false alarms assuming there is no change in the data distributions; and α is a predefined threshold. $E_n[\cdot]$ is the expectation taken with respect to a distribution p_n (such that under p_n, $\hat{d}(1), \ \ldots, \ \hat{d}(T)$ are independent and identically distributed (i.i.d.) with a fixed marginal

distribution. Moreover, ess sup stands for essential supremum. Essential supremum of a set of random variables $\mathcal{X}$ is a random variable Z with the following properties: (i) $P(Z \geq X) = 1 \; \forall X \in \mathcal{X}$; and (ii) $\{P(Y \geq X) = 1, \forall X \in \mathcal{X}\} \rightarrow P(Y \geq Z) = 1, \forall X \in \mathcal{X}$, where $P(\cdot)$ represents the probability [20]. In light of the definitions above, the solution to the constraint optimization problem in (2.11) minimizes the supremum of the average delay conditioned on the worst case realization of $\hat{d}(1), \; \ldots, \; \hat{d}(T)$ over all p_n, $n \geq 1$ [20], [21].

Assuming that the average degree data can be described by a parametric distribution, detecting changes in the parameters may help in finding seizures in the EEG. For this case, the optimum solution to (2.11) is known [22] [23]:

$$\tau = \inf \left\{ n \geq 1 : \left(g(\mathcal{D}_T) = R_t - \min_{1 \leq t \leq T} R_t \right) \geq b \right\}, \tag{2.12}$$

In (2.12), b is a predetermined threshold and $R_t = \sum_{n=1}^{t} \ln \dfrac{p_{\theta_1}(\mathcal{D}_n)}{p_{\theta_0}(\mathcal{D}_n)}$ is the log likelihood ratio between the alternate and null hypotheses. The null hypothesis is that the data distribution did not change while the alternate hypothesis is that the data distribution changed.

There is a practical problem where if the data is assumed to originate from a parametric distribution, but the family of distributions is not known. With adjustments to this technique, a suboptimal detector can be formulated by applying the central limit theorem in a way such that the data approaches normality.

2.2.3 Extended CUSUM detector

The extended CUSUM test is a type of detector that allows for changes in normal distributions to be detected [22]. The extended CUSUM test works by defining a window length N. That window is subdivided into M segments of length n, so $M = N/n$ and M is an integer. Feature vectors are input into the sliding window in the order the data is made available. At each point, the average of each of the M segments is calculated. Assuming that n is large enough, the central limit theorem applies and the sample mean vector of each segment is normally distributed. These steps are summarized below:

$$y(t) = \frac{1}{n} \sum_{\tau=(t-1)n+1}^{t_n} \hat{d}(\tau), t = 1, ..., \frac{N}{n} \tag{2.13}$$

Since each $y(t)$ is assumed to be defined by a Gaussian distribution, the distribution can be summarized by its mean and variance. A null hypothesis is formed that incoming data will be governed by the same distribution. The alternative hypothesis states that the data will come from some different distribution. It is possible to formulate bound distributions $p_{\Theta_1}(y(\tau))$ and utilize the log likelihood ratio between the bound and null hypothesis distributions with likelihood functions $p_{\Theta_1}(y(\tau))$ and $p_{\Theta_0}(y(\tau))$ respectively. The cumulative sum sequence $R(t)$ can be calculated from this log likelihood ratio:

$$R(t) = \sum_{\tau=1}^{t} ln \frac{p_{\Theta_1}(y(\tau))}{p_{\Theta_0}(y(\tau))} \tag{2.14}$$

From the training data that is assumed to originate from the null hypothesis, a sequence of $R(t)$ values is generated. The minimum of the sequence is defined as $m(t)$. Let $g(t)$ be defined as the difference between $R(t)$ and $m(t)$. Set a threshold h equal to the maximum $g(t)$ observed during the collection of the null hypothesis data. If a future $g(t)$ exceeds the threshold, then a change in distribution is detected at time t.

In order to calculate the sequence of $R(t)$ values, the bound distributions need to be calculated assuming some significance level α, test sensitivity γ, the sample mean M_0, and the sample covariance C_0. The two bound means and covariances are as follows [22]:

$$M_1 = M_0 \pm \gamma \sqrt{\frac{n}{N}} Z_{\alpha/2} \sqrt{diag(C_0)} \tag{2.15}$$

$$C_{1,min} = C_0 - \gamma \frac{\frac{N}{n} - 1}{\chi^2_{\alpha/2;N/n-1}} C_0 \tag{2.16}$$

$$C_{1,max} = C_0 - \gamma \frac{\frac{N}{n} - 1}{\chi^2_{1+\alpha/2;N/n-1}} C_0 \tag{2.17}$$

This yields four distributions due to the two bound means and two bound covariances. The log likelihood cumulative sum needs to be done with each of these four distributions. If the null hypothesis is rejected in any case, then a change is assumed to have occurred. A summary of the process can be found in Figure 1.

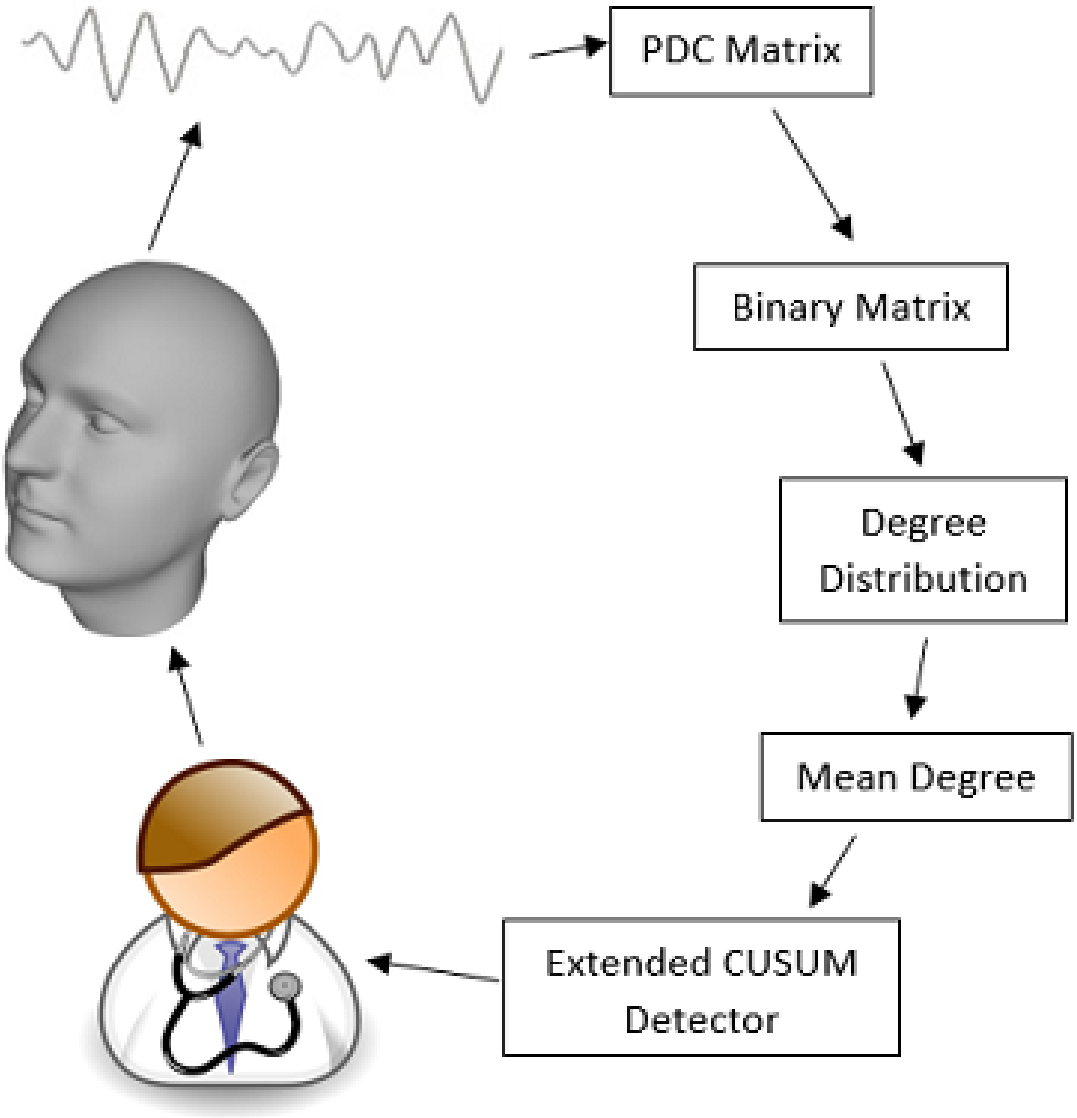

Figure 1: Process flow diagram for the proposed seizure detection routine.

2.3 EXPERIMENTAL PROCEDURES

Clinical transcranial EEG data was collected on four patients undergoing seizure monitoring (University of Pittsburgh IRB No. PRO15100311). Eleven EEG files were collected, each containing normal and seizure EEG. Files were approximately 50 minutes long split with divided approximately in half between normal EEG and seizure EEG. The EEG data was labelled by an expert closely familiar with the patients. The labels were separated into six different portions of "Awake", "Preictal", "Early Ictal", "Late Ictal", "Postictal", and "Sleep. Awake and sleep phases were considered to be normal EEG before and after the seizure while the others were associated with seizure itself. Both awake and sleep were included to determine if the detector alarm would fall silent during the transition from seizure EEG to normal EEG. This ratio of normal to seizure EEG is not representative of data that would be collected in an epilepsy monitoring unit, but it does provide a useful preliminary testbed for the detector.

2.3.1 Patient Dataset Descriptions

There were four patients utilized in this study. Each EEG recording is approximately 50 minutes total divided roughly in half between normal and seizure EEG. In total, about 9.16 hours of EEG data were analyzed. In each dataset, artifacts were observed in various channels. Since this is typical of EEG data collected under practical conditions, these channels were not omitted from the study.

Patient 1 provided two seizures which were recorded using 94 EEG intracranial EEG channels. Channel 89 displayed artifacts across various EEG stages.

Patient 2 had three recorded seizures recorded with 98 EEG channels. Various artifacts not corresponding to meaningful EEG were observed. During the seizure phase, artifacts were seen on channels 36 and 39. In the sleep phase, artifacts were seen on channel 8. Furthermore, a 60 Hz landline noise was identified on channels 8 and 44.

Patient 3 utilized 112 EEG channels for three EEG recordings. Artifacts were seen sporadically in channels 1, 7, 8, 55, 56, 75, and 76.

Patient 4 supplied three EEG recordings using 102 channels. Artifacts were seen on channels 28, 32, and 47.

Commercially available stereo EEG electrodes (Ad-Tech, Inc.) were placed by a neurosurgeon with 6-12 contacts (1.3 mm diameter; 8.88 sq. mm surface area). The placement of electrodes in the brain was decided by University of Pittsburgh Medical Center (UPMC) epilepsy and neurosurgery teams familiar with the patients. Intracranial EEG was recorded using an FDA approved 128-channel NATUS Xltech digital video-EEG system. Data was originally collected at a sampling frequency of 1000 Hz. Preprocessing was done by filtering between 2 and 200 Hz using a constrained least squared FIR filter [24]. An additional notch filter was applied at 60 Hz. The first filter is intended to mitigate high frequency EEG artifacts that likely do not correspond to a seizure but would impact detection. The 2 Hz lower bound is intended to reduce any possible DC drift that is common in EEG signals due to small variations between the prescribed and actual sampling rate. The 60 Hz notch filter is a precaution against noise contributed from land line power. The data was downsampled to 500 Hz to reduce computational requirements while minimizing aliasing effects.

2.3.2 Test Procedures

The offline test was then configured to mimic a seizure monitoring situation. The EEG file was made available to the seizure detection program in two second windows. A six second buffer was used to calculate the PDC values corresponding to the recent data. A four second portion was used to calculate the PDC values for a given window with a 50 percent time overlap used for each batch of PDC values. This overlap is often utilized for short time Fourier transforms and is typically used when calculating PDC values in toolboxes such as HERMES [25]. At each step, Otsu's method was used to select a threshold to turn the PDC matrix into a binary one to eventually obtain $\hat{d}(t)$, as described in Section 2.2.1 [26].

During preliminary investigations, the variance of the average degree distribution was found to be higher during ictal phases than in the awake or sleep phases. This can be seen in Figure 2 and Figure 3. These changes, along with possible shifts in the distribution means, are likely to be found by the extended CUSUM detector.

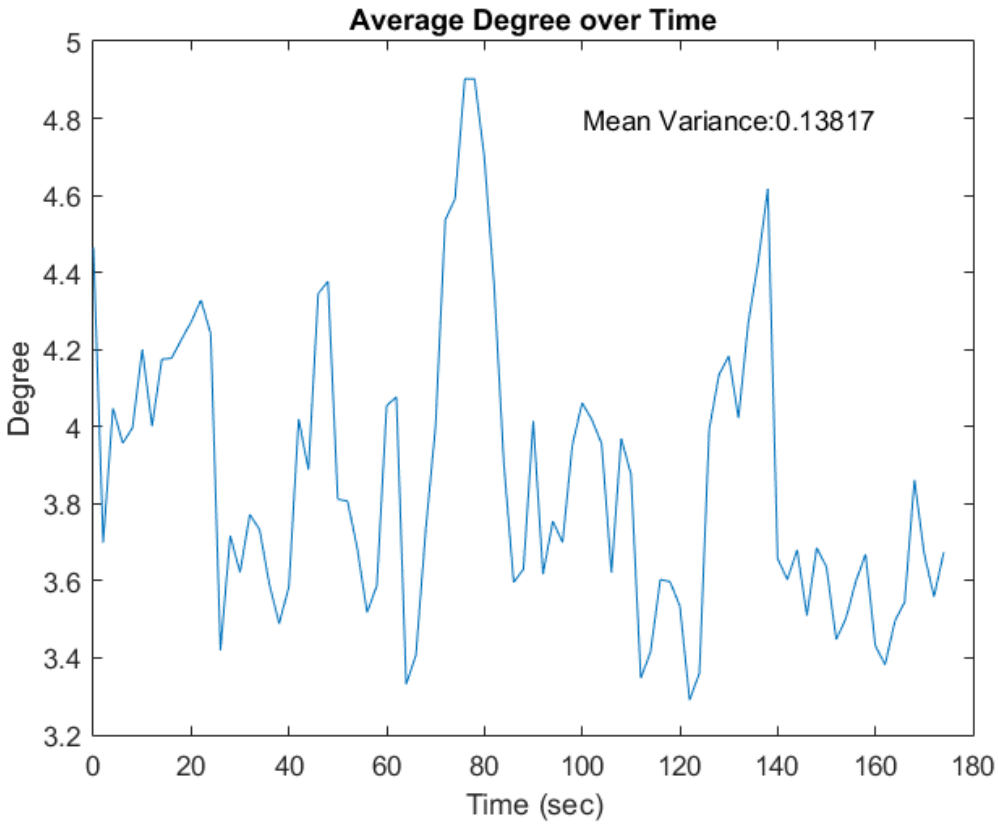

Figure 2: Plot showing the mean degree as a function of time during normal EEG.

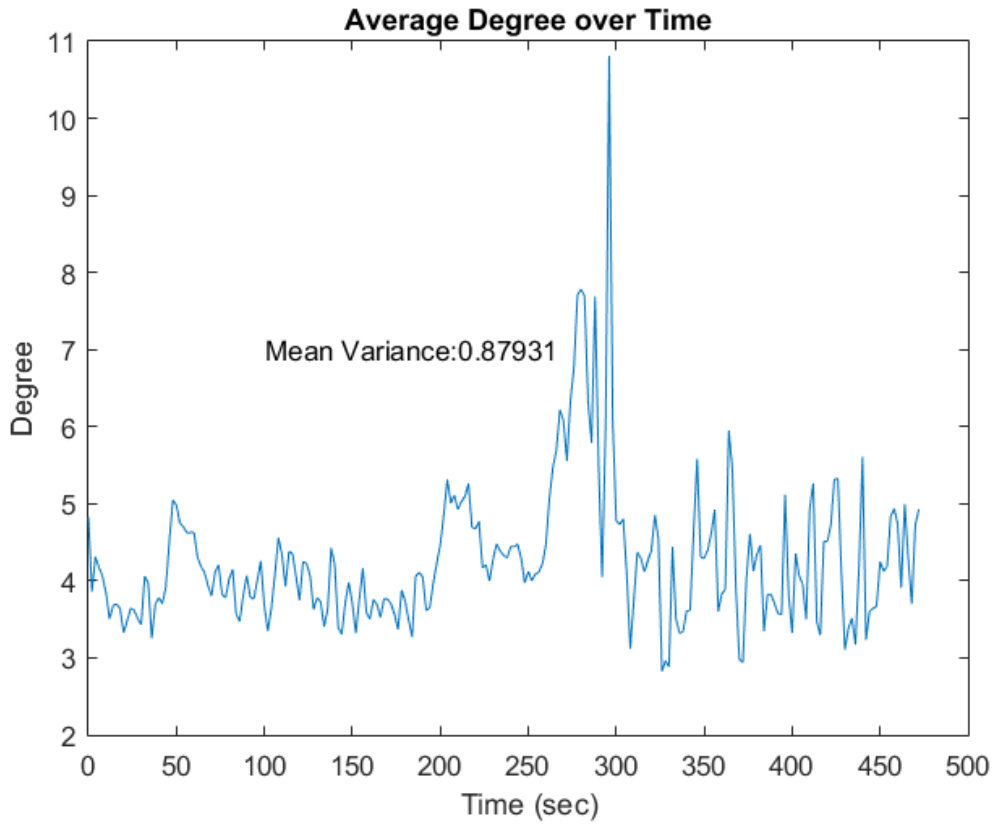

Figure 3: Mean degree during the seizure EEG. The variance of the means were normally higher by inspection. The mean degree also spiked in some cases.

The mean degree $\hat{d}(t)$ is the final statistic that is fed into the extended CUSUM detector. The CUSUM detector was defined to cover a ten minute moving window. With the PDC calculation settings previously mentioned, there was an opportunity for detection every two seconds; therefore, the CUSUM test has been applied every two seconds to detect a seizure. A .01 significance value α that is used in (2.15), (2.16) and (2.17) was used to set up the CUSUM detector. This established a wide null hypothesis distribution bounds to generalize over long periods of EEG data collection.

After the window was populated, the detection phase began. The EEG was processed in the same way as the learning period, and the mean degree was later streamed in the order corresponding to the chronological order of the EEG signal. At each point, the CUSUM detector has an opportunity to signal a change in the distribution. After five increments (corresponding to ten seconds) of detected seizure, a seizure marker was given. This behavior provided some robustness against short term artifacts at the expense of some detection delay.

2.4 RESULTS

In general, the seizure detector was able to achieve a mean sensitivity of .7094 and a mean specificity of .8786. This is more specific but less sensitive than previously established statistical detectors (.7708 sensitivity and .875 specificity) [15]. However, this method requires fewer parameters where other detectors require time series weighting parameters and forgetting parameters. Prior knowledge of the dominant seizure frequencies is also not needed in this detection method whereas previous efforts require this knowledge [15]. On average, the detector provided 74.18 seconds' advance notice before the display of seizure symptoms.

Results for the second patient's third seizure are displayed in Figure 4 as an example of the detector performing well. Labels corresponding to normal EEG are shown as a zero. Nonzero values from one to four are shown for preictal, early ictal, late ictal, and post ictal, respectively. In the displayed example, the CUSUM detector indicated that a statistical change occurred slightly before the labelled preictal phase.

In some cases, some portion of the sleep phase was labelled as a nonstationarity with respect to the awake phase. This can be seen in Figure 5. This indicates that false positives could arise during periods of normal EEG due to changes in mean degree distribution.

Figures 6 and 7 relate the detection results back to two channels of the preprocessed EEG. Due to the low spatial resolution of EEG, Figure 6 has some shared frequencies with significant power by inspection. Figure 7 shows the same two channels but during a seizure. The shared components are lower in frequency and greater in amplitude, which was correctly used by the seizure detector to identify the segment. This shows the capability of the detector to detect changes in the EEG without knowing specific characteristics of the EEG signal beforehand.

A summary of the results can be seen in Table 1. In nine of eleven cases, the seizure was detected before symptoms were shown. In two cases, detection occurred after the onset of seizure symptoms.

Sensitivity and specificity values were calculated using thirty second windows. If abnormal EEG activity was seen at any point in that window, the window is labelled as abnormal. Otherwise the window is considered typical EEG activity. The detection delay was calculated as the time difference in seconds between the first detection and the onset of the early ictal stage, which is when symptoms are present.

2.5 DISCUSSION AND CONCLUSIONS

There are special considerations that should be given to the true class labels that could affect sensitivity and specificity calculations. The EEG seizure data was broken up into neatly divisible segments corresponding to normal and seizure classes. This was done by human expertise, which is well suited to identifying these classes using informal rules involving amplitude, frequency, and waveform shape. Since the nature of the seizure is not fully understood, it is possible that there are small signals related to the seizure that are present before the identified time. In some cases, the CUSUM detector found statistical anomalies

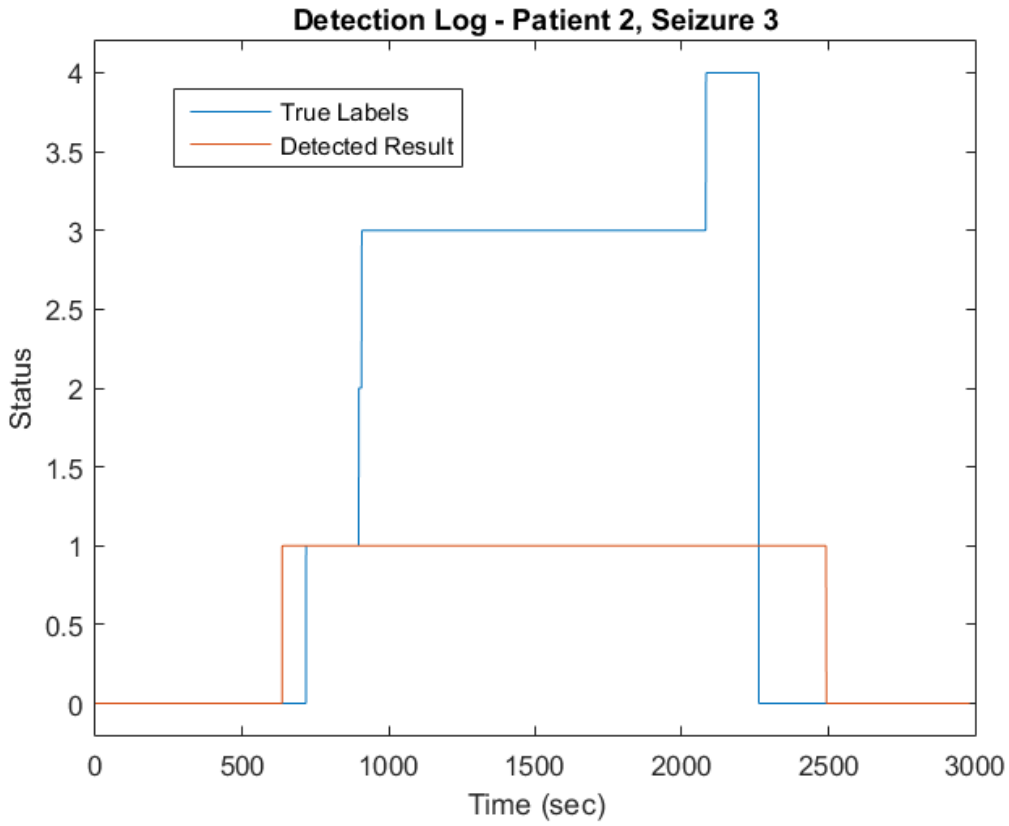

Figure 4: Detection results vs. the true class labels. The data points depict the awake, preictal, early ictal, late ictal, and postictal phases in order

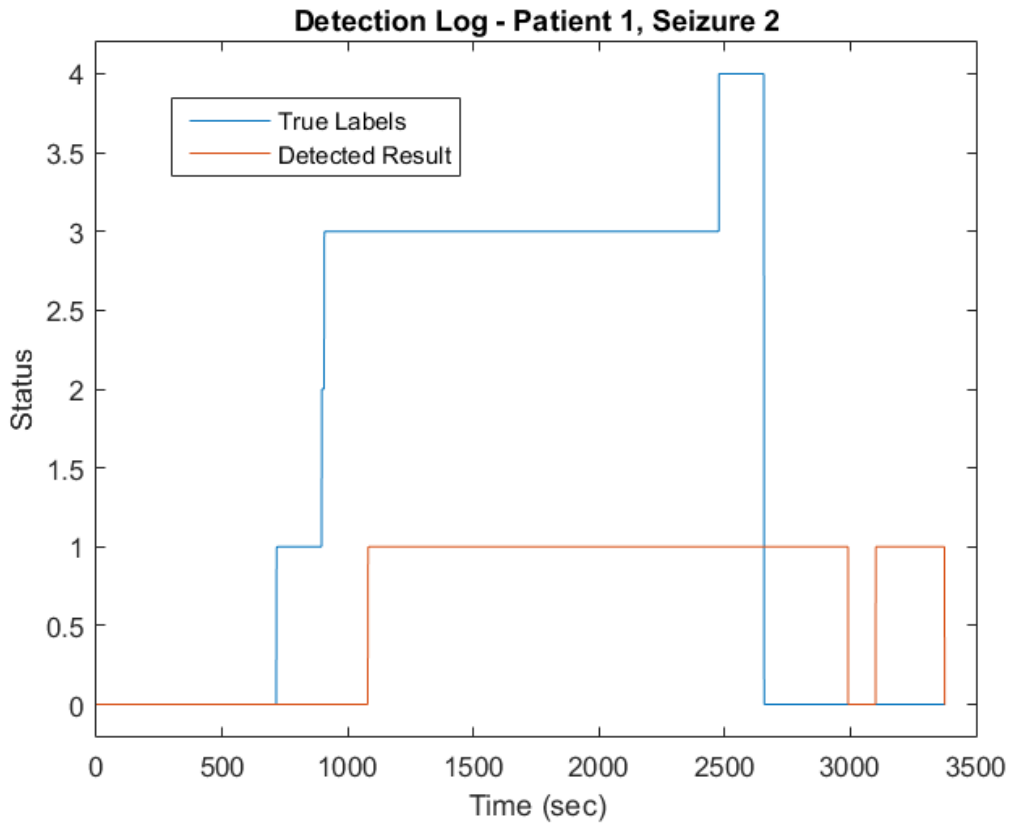

Figure 5: An example of one of the detector's lesser performances. There is a detection delay slightly over three minutes, which would be difficult to take advantage of in a practical setting.

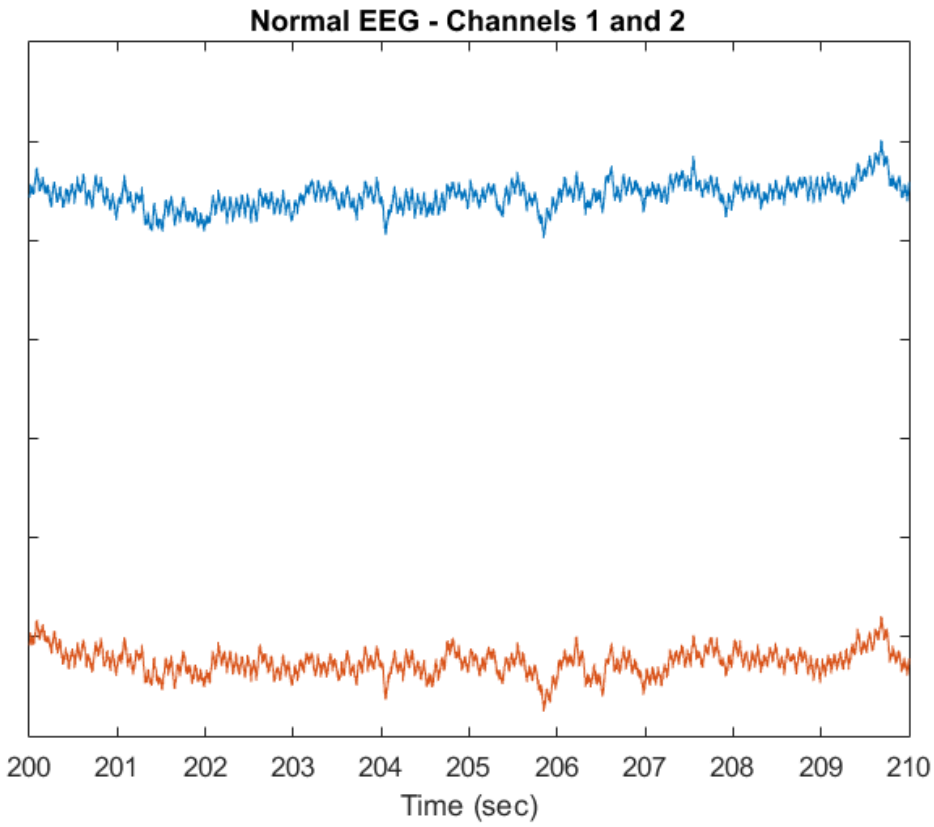

Figure 6: A short segment of two normal EEG channels as labelled by both the expert and the seizure detector. The EEG has some high frequency shared components due to EEG's low spatial resolution.

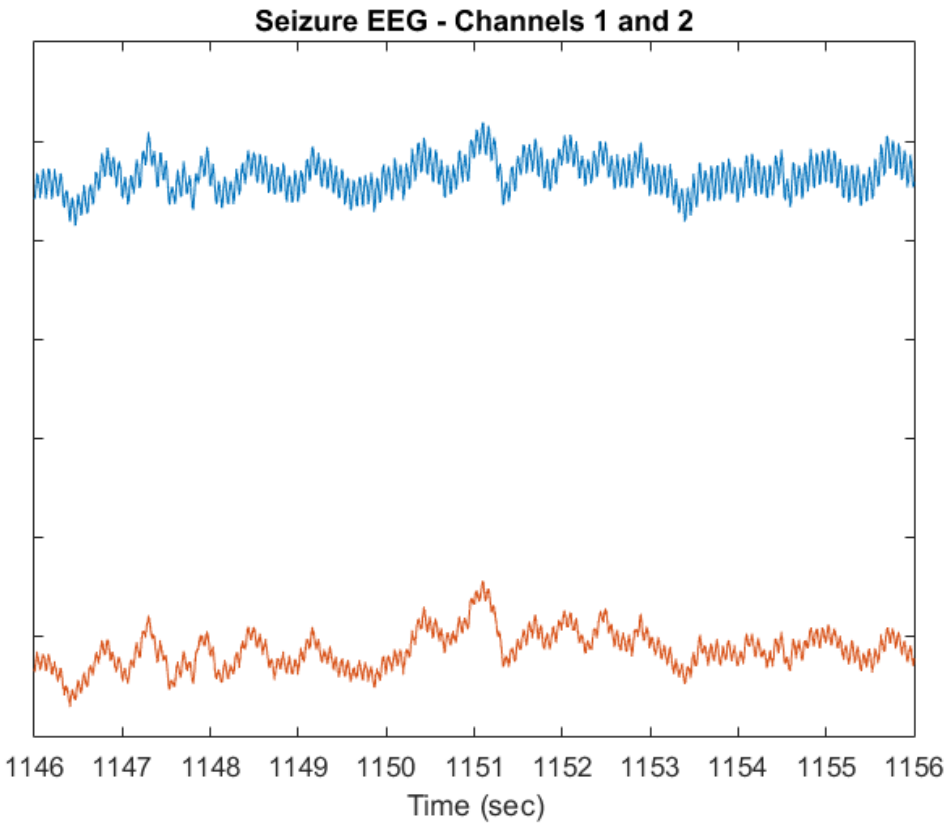

Figure 7: A short segment of two seizure EEG channels as labelled by both the expert and the seizure detector. The EEG channels share a different frequency with a higher power, which was identified without any previous knowledge of the seizure frequencies.

Table 1: Summary of results for the CUSUM detector. In nine of eleven cases, the seizure detector activated prior to the seizure. Performance appears worse for the first patient compared to the other patients

Seizure	**Sensitivity**	**Specificity**	**Delay**
1-1	.7579	.7599	111
1-2	.7216	.6947	186
2-1	.6788	1	-235
2-2	.6952	.8932	-94
2-3	.8319	1	-261
3-1	.6568	.8311	-33
3-2	.6720	1	-202
3-3	.7830	.9546	-132
4-1	.6192	.8069	-30
4-2	1	.8971	-97
4-3	.6909	.8269	-29
Mean	**.7094**	**.8786**	**-74.18**

before even the labelled preictal segments. These may in fact be indicators of a seizure. However, these were treated as false positives in the interest of fair results reporting.

In general, the activity of the detector was correlated with the seizure event. In some cases, the detector activated during the preictal phase. For these cases, medical staff would be given between thirty seconds and four minutes to respond to the beginning of visible seizure symptoms. In two cases, the detector activated after symptoms were shown. While this indicates that an underlying change in the degree distribution was found, medical staff would not be able to intervene based only on the detector's output. However, it would still expedite the offline analysis of EEG records, making it a useful result regardless.

The mean sensitivity of the detector was .7094 using the 30 second window definition. Seizures are typically reacted to within 1.5 to 2.5 minutes, meaning that the chance of a detection within that window is high. For example, a 1.5 minute window has on average a 97.6 percent chance of having a positive detection. Regardless, higher sensitivity would improve detector performance, and future development efforts should be focused on how to increase this sensitivity.

One interesting note is that the detector was unable to detect changes in the EEG for patient 1 prior to the seizure. The specificity for this patient was also considerably lower than the rest of the data set. The reason for this is unclear since the artifacts were not noted to be significantly different from the rest of the data set. Further investigation by a team familiar with the patient's EEG patterns is necessary to understand why the performance is different.

The main strength to this detection method is in the lack of assumptions needed. The large amount of summations allows the central limit theorem to take effect on the mean degree. For example, the recordings used for testing utilized between 94 and 112 channels. Averaging and adding the degrees over these channels and 64 frequency bands makes it likely that the central limit theorem applies. The mean degree was observed to have higher variance for the seizure phases during initial investigations, but the use of the CUSUM test detects changes in the distribution. If the mean degree changes in a different manner than what was seen during initial investigations, the detector still can find the seizure. All of this is done without needing to know any signal characteristics of the seizure.

Seizure detection using the extended CUSUM test appears to be a promising technique for clinical monitoring purposes. The proposed method has shown success in differentiating seizure EEG data from normal EEG without the need for highly customizable parameters, previously labelled EEG data, or expert knowledge regarding the nature of the seizures. After observing normal EEG data, assumptions about the multivariate autoregressive model allow for suitably fast computation of partial directed coherence values for online application, especially if implemented in a compiled language such as C++. Further testing is needed in order to assess the generalizability to a larger population, but with eleven seizure files, performance has been promising.

There is still room to improve the performance of this detection method. The seizure file that provided the poorest results had many artifacts. Artifacts were also present in the other files to different extents. Removal of common artifacts through a database of common EEG artifacts would likely increase the performance of the detector at the expense of maintaining the database. Statistical artifact removal would not be a valid option because it would conflict with the statistical methods used to mark a seizure.

3.0 TRANSFER LEARNING FOR SSVEP EEG BRAIN-COMPUTER INTERFACES USING LEARN++.NSE AND MUTUAL INFORMATION

3.1 INTRODUCTION

In the previous section, a detector was formulated to take advantage of EEG nonstationarity. Other EEG applications are posed in a way that EEG nonstationarity is a hindrance. One of these applications is the brain computer interface, which learns distributions conditional on the true class label corresponding to user intent. If these distributions change, then the learned decision thresholds no longer hold and user intent can no longer be determined. This means the brain computer interface must relearn the class conditional distributions, which takes time which would otherwise be used for useful operation. In this section, a technique for identifying previously learned class conditional distributions that convey the most information about current distributions is given, thus lowering the impact of EEG nonstationarity in classification scenarios.

Brain Computer Interfaces (BCI) are an emerging input modality for disabled users seeking to communicate by computer [4] [27]. BCI systems use brain signals directly as input to infer user intent. These systems are particularly useful for users capable of minimal movement who cannot rely on typical input modalities such as keyboards, mice, or joysticks.

Typical BCI systems require calibration to collect labelled data. These data points are used to estimate the distribution of the features calculated from the EEG. Unfortunately, nonstationarities exist in the EEG signal. For a single user, common nonstationarity sources include artifacts, equipment changes, environmental variables, and user fatigue [4]. This last issue is of interest since lengthy calibration can cause fatigue and limit the application of the learned data. Transfer learning may be a solution to reduce calibration requirements [28].

Steady state visual evoked potentials (SSVEP) are a type of BCI control signal elicited by the use of flickering stimuli [27]. Increased power can be seen in the frequency spectrum of the EEG directly corresponding to the stimuli frequencies. For this reason, SSVEP is considered as an easy to use control signal. As such, there has been little research in combating the transfer learning and nonstationarity problems in SSVEP systems.

3.1.1 Previous Work

The state of the art for SSVEP signals is canonical correlation analysis (CCA) where the EEG signals generate correlation scores with reference signals based upon the stimuli being shown [29] [30]. For straightforward CCA application where the stimuli are all different frequencies, there is no need for training since the maximum correlation score is picked on a trial-by-trial basis.

3.1.2 Cases Where CCA Assumptions are Violated

CCA is a suitable technique for healthy users who are able to adequately direct their covert attention to all the available stimuli [31] [32]. For patients with disabilities that require the use of a BCI, this may not be the case. We present an example of collected data from a healthy user who exhibited an EEG response similar to the described case. The user utilized an SSVEP system with two flickering checkerboard stimuli at 6 and 20 Hz. As Figure 8 shows, regardless of which stimultus the user was focusing on, the response to the 6 Hz stimulus is stronger. This can be seen in the first two harmonics. CCA was only able to achieve an accuracy of 67 percent on this user most likely due to the fact that choosing the maximum correlation score was inappropriate for a case where the stimuli responses were so uneven. Other examples like this have been seen in the literature, further emphasizing the need to develop algorithms for this population [33]. There have been a few attempts to rectify this by normalizing against the background EEG in neighboring frequency bands [34], but this assumes the signal to noise ratio is flat across the spectrum. Based on the large disparity in Figure 8, this assumption likely does not hold and further investigation is needed.

3.1.3 Machine Learning Alternatives and Transfer Learning

Where a maximum score selection method may fail, a machine learning approach may succeed by considering the stimuli harmonic responses as a whole. This reintroduces the problem of nonstationarity and high calibration requirements. As such, patients with disability end up neglected by the advances afforded by CCA and other state of the art tools whose assumptions are unfulfilled in the practical case. Since calibration is required, another technique to apply transfer learning for these patients is needed.

We provide two examples for the transfer learning and nonstationarity problems separately. Figure 9 shows a session of an SSVEP system with 6 and 20 Hz stimuli. Features were constructed using the first two harmonics of each stimuli. These are shown in the plot for each trial belonging to a particular stimulus. The top and bottom figures show the power in the harmonics when the visual attention is on the stimulus flickering with 6Hz and 20Hz, respectively. After the 15th data point, the variance for the principle harmonics increases. This is a clear example that nonstationarity exists in SSVEP signals despite their supposed simplicity.

The transfer learning issue is shown in Figure 10. Two participants' average EEG power in four frequency bands are shown. These correspond to a two stimulus system with 6 and 20 Hz flickering checkerboards. The figure shows the case when the visual attention is on 20 Hz stimulus. The two participants display very distinct average responses with the first one favoring the 6 Hz band even when directing attention to the 20 Hz stimuli. The second participant shows a more predictable response. Examples like these show that transfer learning is necessary even in the case of SSVEP.

Transfer learning applications in SSVEP are sparse due to the prevalence of CCA. One notable effort is Multiset CCA [35]. In this technique, the reference signal is not assumed to be a collection of sinusoids representing the stimulus frequencies and their harmonics. Instead, the reference signal is learned from a group of previously obtained data sets. The authors report increased accuracy over CCA, which is indicative of how transfer learning may improve performance in cases where traditional CCA assumptions do not apply. There are unanswered questions as to how well this method applies if the data sets are significantly

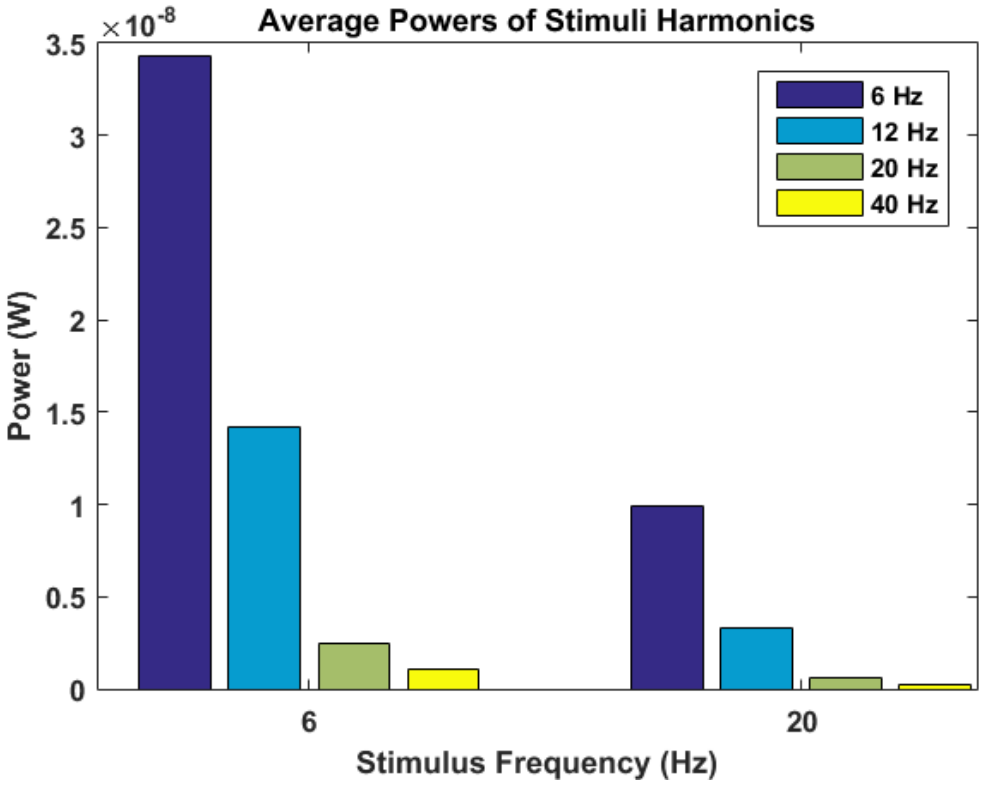

Figure 8: Example of a nonuniform stimulus response in a SSVEP system. Choosing the maximum correlation score favors the 6 Hz class, so machine learning with transfer learning is needed.

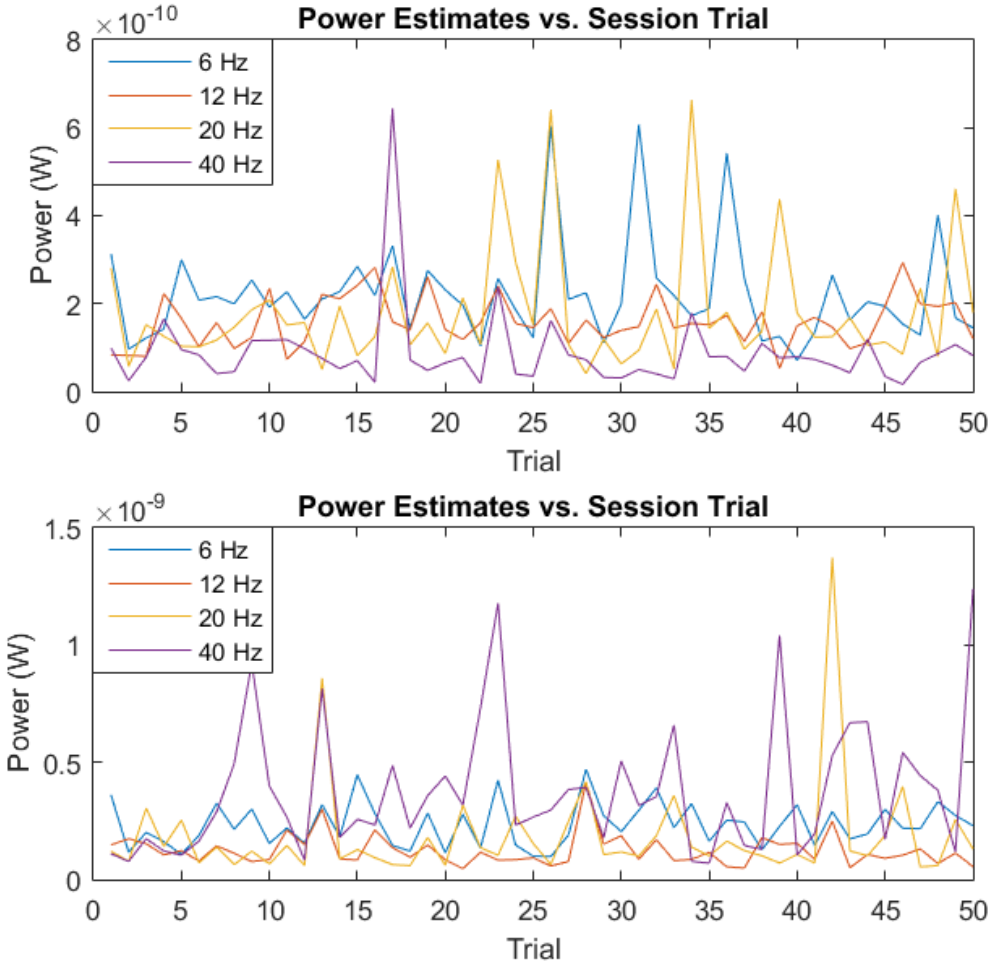

Figure 9: Example of a nonstationarity in an SSVEP BCI application. The variance of some features including the principal harmonics increases after the 15th trial.

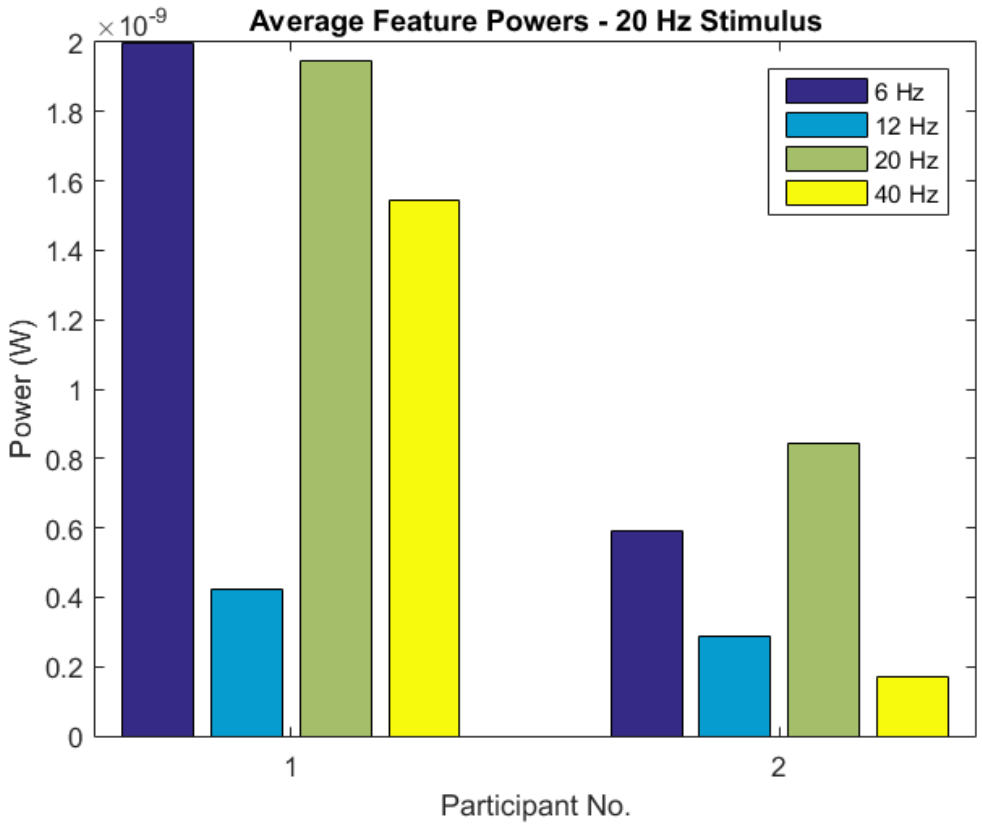

Figure 10: Example showing that where a single classifier would likely fail across multiple users.

different due to the nonstationary nature of EEG. This issue is investigated for comparison purposes to see if the current state of SSVEP BCI research is adequately prepared for transfer learning.

3.1.4 Contribution

We introduce an ensemble learning technique called Learn++.NSE combined with mutual information as a data set selection method. Learn++.NSE provides a framework for combining data where nonstationarity exists between data sets, and mutual information guides data set selection to populate the ensemble. This classification scheme gives the brain computer interface the ability to deal with inter-session nonstationarity while being able to select data sets that maximize performance on incoming data regardless of the current user. Results show that this method achieves higher accuracies for users where traditional classification schemes, particularly CCA, may fail while maintaining similar accuracies for typical users.

3.2 METHODS

3.2.1 Learn++.NSE

Learn++.NSE was chosen as the ensemble learning algorithm due to its ability to assign useful weights for ensembles of any size while keeping computational complexity low. The details of this algorithm are summarized in Algorithm 1 [36].

First, define an ensemble hypothesis for a given data point at a discrete time t as $H^t(x)$. Voting weights V^t for each of the k^t ensemble members must be found.

$$V^t = [V_1^t, V_2^t, \ldots, V_{k^t}^t] \tag{3.10}$$

Each of the k^t individual member hypotheses $h_{k^t}^t(x)$ will generate up to c candidate decisions for the entire ensemble. The final ensemble hypothesis $H^t(x)$ is chosen such that:

$$H^t(x) = \arg\max_c \left(\sum_{i=1}^{k^t} (h_i^t(x) == c) * V_i^t\right) \tag{3.11}$$

Data: Data set D^t of length m^t

A designated base classifier algorithm, Real valued (a,b) sigmoid parameters

Ensemble hypotheses $H^t(x_i^t) = \hat{y_i^t}$ with size k^t

Result: Trained ensemble H^t

for *t = 1, 2, 3, ...* **do**

if *t = 1* **then**

Initialize weight vector $w^t(i) = \frac{1}{m^t}$ and go to step 4

end

1. Determine ensemble error for current data set D^t

$$E^t = \frac{1}{m^t} \sum_{i=1}^{m^t} \hat{y}_i^t \neq y_i^t \tag{3.1}$$

2. Perform boosting step if correctly classified

$$w^t(i) = \frac{1}{m^t} * E^t \tag{3.2}$$

Otherwise

$$w^t(i) = \frac{1}{m^t} \tag{3.3}$$

3. Normalize w^t so that w^t is a distribution. Train new base classifier on D^t

4. Compute individual classifier errors on D^t for k=1:k^t

$$\epsilon_k^t = \sum_{i=1}^{m^t} w^t(i) * (h(x_i^t) \neq y_i^t) \tag{3.4}$$

If $\epsilon_k^t > \frac{1}{2}$ for $k < k^t$, set $\epsilon_k^t = \frac{1}{2}$. If $\epsilon_k^t = \frac{1}{2}$ for $k = k^t$, retrain latest classifier

5. Normalize individual classifier errors

$$\beta_k^t = \frac{\epsilon_k^t}{1 - \epsilon_k^t} \tag{3.5}$$

6. Compute weighted average of all classifier errors using sigmoidal curve

$$\omega_k^t = \frac{1}{1 + e^{-a(t-k-b)}} \tag{3.6}$$

$$\omega_k^t = \frac{\omega_k^t}{\sum_{j=t-k}^{t} \omega_k^t} \tag{3.7}$$

$$\bar{\beta}_k^t = \sum_{j=0}^{t-k} \omega_k^{t-j} * \beta_k^{t-j} \tag{3.8}$$

7. Calculate voting weights

$$V_k^t = \log(\frac{1}{\bar{\beta}_k^t}) \tag{3.9}$$

end

Algorithm 1: Outline of the Learn++.NSE algorithm

Next, a data weight distribution w^t for the incoming training data set is defined. The distribution is initialized uniformly, so $w^t(i) = \frac{1}{m^t}$, where m^t is the amount of training data points in newly available data set D^t.

First, the ensemble error rate, E^t, is assessed on the data set D^t. This is done using the previous ensemble hypothesis $H^{t-1}(x)$ from k^{t-1} member classifiers on each data point x in D^t. Each of the i data points in D^t is assigned a new weight $w^t(i)$ by multiplying its current weight by the ensemble error rate E^t if the data point was classified correctly by the ensemble. Since $E^t \leq 1$, correctly classified points will always have a lower weight in the distribution. These steps are represented by steps 3.2 and 3.3 in Algorithm 1.

Learn++.NSE handles nonstationarities by adding a new hypothesis $h^t_{k^t}(x)$ on the most recent training data set D^t and by calculating voting weights V^t of the resulting ensemble. The voting weights are found by evaluating the individual classifier error rate ϵ^t_k for each of the k^t classifiers as shown by step 5. These error rates are also affected by the data weight distribution D^t. Note that the age of the ensemble members is not directly taken into account.

A sigmoidal error weighting scheme is included in step 7 to prevent overfitting to the data [36]. The sigmoid curve weight before normalization, $\omega(t)$, is defined by:

$$\omega^t_k = \frac{1}{1 + e^{-a(t-k-b)}} \tag{3.12}$$

In this formula, k is the classifier position within the ensemble. The quantity $t-k$ is the time difference between the current time and the classifier creation time. Two parameters a and b are also introduced. These control the slope and horizontal offset respectively. These hyperparameters need to be tuned according to the data [37]. This was accomplished using a grid search in the hyperparameter space using ten fold cross validation for testing, with nine fold internal cross validation for every point in the hyperparameter space.

The final classifier voting weight of classifier k, V^t_k, is based on a combination of the errors from the current and past data sets. The weighted error rate $\bar{\beta}^t_k$ is calculated based on the procedures shown in step 7.

The final voting weights for the k-th classifier, are obtained by taking the log reciprocal of the $\bar{\beta}^t_k$ [2].

3.2.2 Incorporating Mutual Information

Mutual information is a measure of how much information one random variable provides about another. In this experiment, mutual information was used as a method for finding which previously collected data sets best represented the incoming data set. In general, the mutual information between vector random variables X and Y is defined as [38]:

$$\int_X \int_Y p(X,Y) \log(\frac{p(X,Y)}{p(X)p(Y)}) dY dX \tag{3.13}$$

Applying a Gaussian distribution assumption, the mutual information between X and Y of equal dimension with covariance matrices C_X and C_Y respectively can be calculated as:

$$I(X:Y) = \frac{1}{2} \log(\frac{det(C_X)det(C_Y)}{det(C)}) \tag{3.14}$$

The covariance matrix C is the full covariance matrix obtained by concatenating X and Y. If a data set contains n vectors for each variable, then there are n^2 combinations that will yield their own unique estimates of C. Averaging these estimates will reduce the overall estimate variance of C.

The mutual information can be incorporated into Learn++. NSE by receiving a new training data set. From that data set, the true class labels can be used to calculate the posterior probability distributions for each class. The total mutual information between every pre-existing data set and the incoming data set is found. From there, the m highest ranking data sets are chosen for training in the Learn++.NSE framework where the lowest ranking data set is introduced first, thereby making it the oldest classifier in the ensemble.

3.2.3 Statistical Measurement Comparisons to Mutual Information

There are other methods with which the similarity of distributions can be assessed. Two of these are considered for this paper. Mahalanobis distance allows the distance of a vector from a distribution given its mean vector μ and covariance matrix C [39]:

$$M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T C^{-1} (\vec{x} - \vec{\mu})} \tag{3.15}$$

Bhattacharyya distance is similar in its goals but instead measures the distance between two distributions. For two distributions P and Q, the Bhattacharya distance is [39]:

$$B(P,Q) = \int \sqrt{P(\vec{x})Q(\vec{x})}d\vec{x} \tag{3.16}$$

Applying the same normality assumptions as the mutual information, the Bhattacharyya distance can be calculated from the means and covariances of P and Q:

$$B(P,Q) = \frac{1}{8}(\vec{\mu_1} - \vec{\mu_2})^T C^{-1}(\vec{\mu_1} - \vec{\mu_2}) + \frac{1}{2}\ln(\frac{det\frac{C_1+C_2}{2}}{\sqrt{det(C_1)det(C_2)}}) \tag{3.17}$$

These distance metrics were used in the same way as mutual information to populate the ensemble except that data sets with minimal distance were chosen whereas maximum mutual information was used previously. This comparison exists to check if an information theoretic or statistical approach provides any differences in performance in the collected data sets.

3.3 EXPERIMENTAL PROCEDURES

3.3.1 System Description

We developed an SSVEP-based BCI for binary selection that employed two flickering checkerboards at 6 and 20 Hz. The system was realized on a Lenovo ThinkPad laptop running 64-bit Windows 7. MATLAB 2015a was used for data acquisition, signal processing, feature extraction and classification; and Psychtoolbox (a freely available toolbox for creating time-accurate stimuli for experiments) was used for presentation. A general flowchart for system operation is shown in Figure 11.

The system was connected to a g.Tec g.USBamp via USB for data acquisition [40]. The amplifier was connected to a g.gammaBox which was directly connected to the electrodes. Single channel EEG was used over the visual cortex (OZ on the 10-20 system) with a butterfly

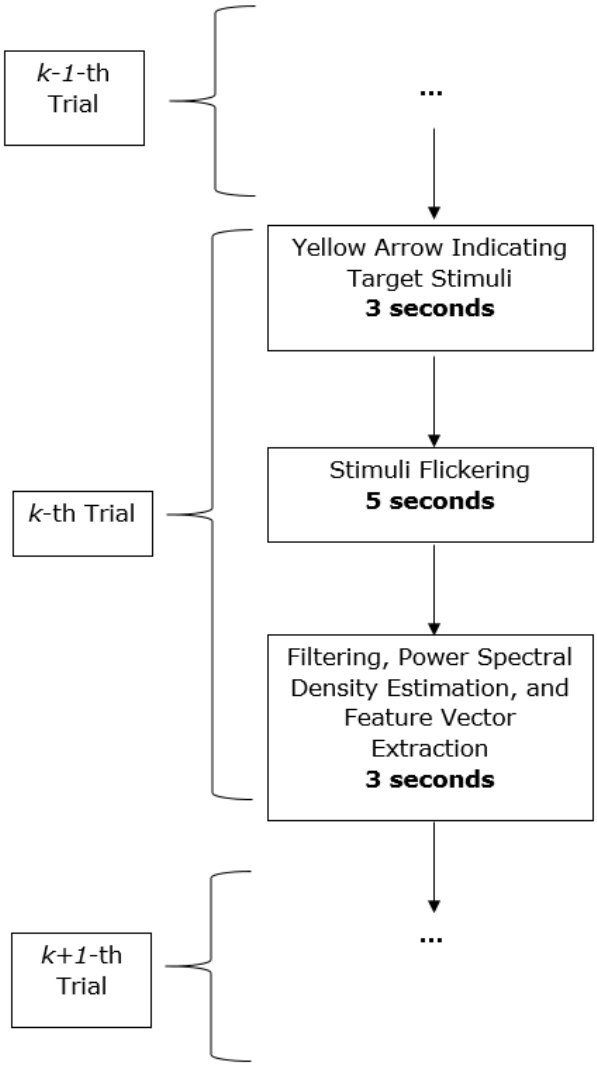

Figure 11: Flowchart of system operation. The above process is repeated one hundred times for a given calibration or test phase.

electrode. A ground electrode was placed over the forehead (FPZ on the 10-20 system). A reference electrode was clipped to the earlobe. A parallel port cable was also used to output digital values to the amplifier depending on the system's current state. This digital value was sampled alongside the EEG data to easily separate the EEG data of interest.

3.3.2 Participant Description and Experimental Procedures

Ten healthy participants (8 males and 2 females) were enrolled in this study according to the University of Pittsburgh IRB No. PRO15060140. All participants were required to be at least 18 years of age and have no history of epilepsy.

All participants were asked to direct their covert attention randomly at one of the two checkerboards at the start of every trial. Each trial consisted of flickering of the checkerboards for five seconds. In one usage session one hundred trials were presented. There were three usage sessions. In all sessions, a calibration phase was taken to collect training data. On the final session, a test phase of equal length was used to collect testing data where the ensemble would be evaluated.

3.3.3 Signal Processing and Feature Extraction

The EEG segments collected during each trial were sampled at 256 Hz and filtered using a 150th order constrained least squares FIR filter from 2-45 Hz [41]. A power spectral density estimate was obtained using Welch's method [42]. Features were made using the first two harmonics of the stimuli frequencies to obtain a four dimensional feature vector.

3.3.4 Classification

A linear discriminant classifier was used in the Learn++.NSE ensemble to reduce computational complexity. For each participant, ensembles were formed using groups of three individual classifiers. Two groups of ensembles were generated. The first group consisted of the mutual information-based Learn++.NSE ensembles (designated LPP-MI). Here, the

mutual information between the latest data set recorded from a certain participant and all other data sets from all the participants were computed. Latest data set and the two data sets with the most mutual information were used for the training of LPP-MI for that specific participant. Specifically, the data sets were added to the Learn ++ algorithm in the following order: (1) the set with second most mutual information, (2) the set with most mutual information, and (3) most recent data set. The second group contained the standard Learn++.NSE ensembles (denoted as LPP-S). For each participant, LPP-S was formed using the three training data sets corresponding to that specific participant. An LDA classifier was also trained for each participant on their last session's calibration phase in order to compare performance under typical calibration procedures. The three classifiers were then compared by examining their accuracies over the test data which was not used for training.

3.4 RESULTS

Figure 12 shows the accuracies for six different methods. The first one is a standard Learn++.NSE ensemble (LPP-S) while the second is for the Learn++.NSE ensemble augmented with mutual information (LPP-MI). This comparison exists to show if the mutual information calculation aids classification. The third method is standard CCA where the reference signals are sinusoids with frequencies corresponding to the first two harmonics of the presentation stimuli. The fourth method is MultiSet CCA with reference signals learned from previous data sets corresponding to a particular user. This gives CCA a transfer learning component that makes the comparison fairer. The fifth and sixth methods utilize Learn++.NSE augmented with Bhattacharyya distance (LPP-Bhatt) and Mahalanobis distance (LPP-Mahal) instead of mutual information. These additional transfer learning methods serve to compare mutual information to statistical measures. The average accuracies are shown in Table 2.

LPP-MI saw a large increase in accuracy over LPP-S and a small increase over other LPP ensembles. However, CCA performed best across all users. On the other hand MultiSet CCA performed worse than both LPP-MI and CCA when using data sets from previous sessions.

Participant 8 should be noted where LPP-MI performed best by a wide margin. This participant had a much stronger response to the 6 Hz stimulus than the 20 Hz one and is the source of Figure 8 in the Introduction. An LDA classifier was also attempted on this participant with an accuracy of 74 percent.

3.5 DISCUSSION AND CONCLUSIONS

The participant pool in this study consisted of healthy users who were able to fully divert their covert attention toward the targeted checkerboards. For this reason, the relative success of CCA over other techniques is expected given its current position as state of the art for SSVEP BCI systems.

However, participant 8's unusual responses to the two stimuli frequencies makes for a good test case of someone who may not have full control of their covert attention. In this case, CCA would fail since the first target's frequencies would always be chosen. This requires machine learning techniques and transfer learning if incorporating support for multiple data sets. Mutual information and Learn++.NSE allows for this transfer learning. In this case, the algorithm used a combination of participant 8's own data and two other data sets to achieve its accuracy.

The ensembles using mutual information appear to perform better than the ones using Bhattacharyya or Mahalanobis distance. This indicates that using information theoretic measures outperforms purely statistical ones. The reasons for this are uncertain, but the estimation of the augmented covariance matrix may provide robustness against variance in the EEG.

Multiset CCA appears to perform worse than CCA in every case. This may be due to the nonstationarity between data sets reducing the effectiveness of the generated reference signal. This indicates that variants of CCA are currently not equipped to handle nonstationarity.

The results of the investigation show that further effort is needed to reduce calibration requirements for non ideal usage cases. This includes users without full gaze control or unusual SSVEP responses. Mutual information and ensemble learning are possible venues

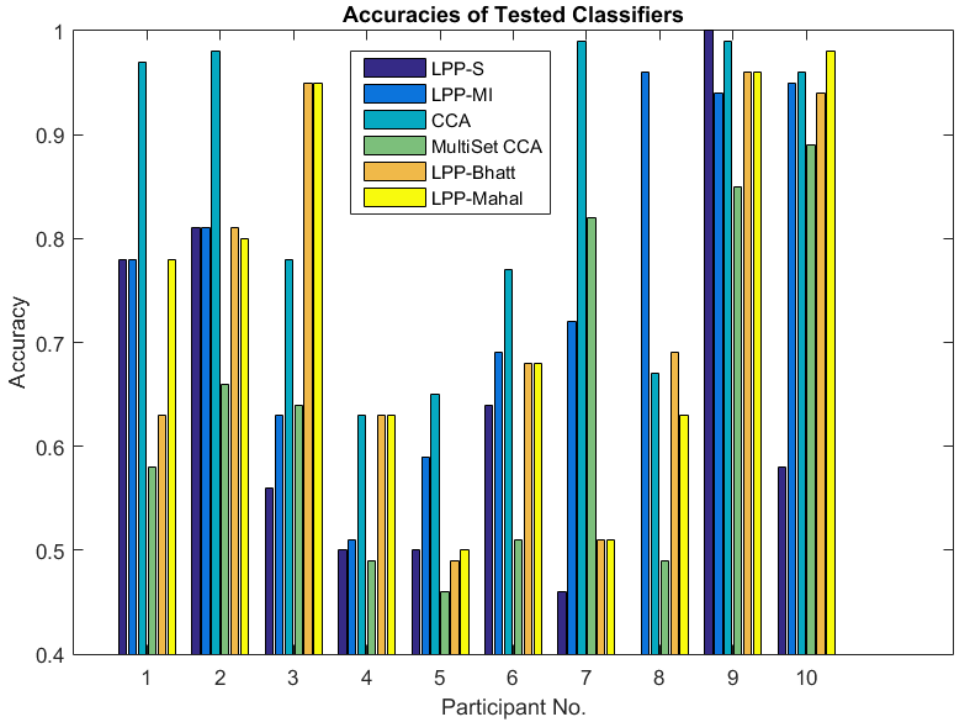

Figure 12: Accuracy results comparing the LPP-S, LPP-MI, CCA, MultiSet CCA, LPP-Bhatt, and LPP-Mahal. CCA appears to perform best across all users, but seems to perform much worse than LPP-MI for participant 8.

Table 2: Average accuracy for each method tested in this investigation.

LPP-S	**LPP-MI**	**CCA**	**MultiSet CCA**	**LPP-Bhatt**	**LPP-Mahal**
62.3%	75.8%	83.9%	63.9%	72.9%	74.2%

to explore toward this end. Further modifications are possible. For example, using mutual information as a score within the Learn++.NSE algorithm may improve results.

4.0 CONCLUSIONS

As demonstrated above, EEG nonstationarity can be either a hindering or useful feature depending on the application. In the first half of this thesis, the seizure detection problem was framed in a way such that nonstationarity could be indicative of a seizure onset. In this case, the problem essentially becomes one of outlier detection. This has some attractive properties due to the ability to use statistical methods to define when a change in distribution has occurred. These statistical methods are better suited to nonstationarity than the machine learning techniques often applied in seizure detection due to the lack of labelled training data needed to train a classifier. This statistical approach was realized by posing the EEG channels as a random graph with the edges determined by partial directed coherence. The distributions during normal EEG were learned so that the extended CUSUM detector could derive alternative distributions that could be used to invalidate the null hypothesis of unchanging EEG distributions.

In other applications such as brain computer interfaces, machine learning techniques are more appropriate since decisions need to be made on a case by case basis. Here, the nonstationarity cannot be leveraged as in seizure detection applications. Instead it must be ameliorated. By using mutual information to rate the similarity of data sets to incoming data, a brain computer interface could identify data sets that would best train the classifier to the current distribution. Stacking this on top of an ensemble learning algorithm provides further robustness by taking a vote among several classifiers instead of relying on the decision of a single one.

4.1 FUTURE WORK

The techniques described in this thesis have further room for improvement. The seizure detection method can utilize other features of the random graph that might provide better results. For example, cluster coefficients may reveal more information about an oncoming seizure. However, a different detection test beside the extended CUSUM test would likely need to be employed in order to observe when the ensuing distributions have changed. The extended CUSUM test relies on a normality assumption through the Central Limit Theorem. The distribution of the cluster coefficients would need to be derived and applied to a standard CUSUM test, assuming that said distributions are parametric. Theoretical advances aside, the detector also could be implemented in a real time system placed in an epilepsy monitoring unit. This would allow the detector's performance to be assessed in an online practical situation.

For BCI applications, the next step would be to improve on the ability to select beneficial training data. Currently, the classification scheme outlined is limited to selecting entire data sets that best apply to observed circumstances. However it may be better to select subsets of individual data sets. Specifically, drawing upon the different true class labels within each data set may provide additional performance. Selecting individual data points would provide the classifier even more ability to adapt to new distributions with old data.

BIBLIOGRAPHY

[1] E. Niedermeyer and d. S. Lopes, *Electroencepahlography: basic prinicples, clinical applications, and related fields*. LWW, 2005.

[2] S. Ramgopal, S. Thome-Souza, M. Jackson, N. E. Kadish, I. ánchez Fernández, J. Klehm, W. Bosl, C. Reinsberger, S. Schachter, and T. Loddenkemper, "Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy," *Epilepsy and Behavior*, vol. 37, pp. 291–307, 2014.

[3] Y.-Y. Lee, M.-Y. Lee, I.-A. Chen, Y.-T. Tsai, C.-Y. Sung, H.-Y. Hsieh, S.-N. Lim, P. W. Hung, and T. Wu, "Long-term video-EEG monitoring for paroxysmal events.," *Chang Gung medical journal*, vol. 32, no. 3, pp. 305–312, 2009.

[4] M. Akcakaya, B. Peters, M. Moghadamfalahi, S. Member, A. R. Mooney, U. Orhan, S. Member, B. Oken, D. Erdogmus, S. Member, and M. Fried-oken, "Noninvasive Brain Computer Interfaces for Augmentative and Alternative Communication," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 31–49, 2014.

[5] D. Takeshita and S. Bahar, "Synchronization analysis of voltage-sensitive dye imaging during focal seizures in the rat neocortex," *Chaos*, vol. 21, no. 4, 2011.

[6] U. R. Acharya, "Autoamtic Detection of Epileptic EEG SIgnals Using Higher Order Cumulant Features," *International Journal of Neural Systems*, vol. 21, no. 5, pp. 403–414, 2011.

[7] A. Kharbouch, A. Shoeb, J. Guttag, and S. S. Cash, "An algorithm for seizure onset detection using intracranial EEG," *Epilepsy and Behavior*, vol. 22, no. SUPPL. 1, pp. S29—-S35, 2011.

[8] A. M. Chan, F. T. Sun, E. H. Boto, and B. M. Wingeier, "Automated seizure onset detection for accurate onset time determination in intracranial EEG," *Clinical Neurophysiology*, vol. 119, no. 12, pp. 2687–2696, 2008.

[9] S. B. Wilson, "A neural network method for automatic and incremental learning applied to patient-dependent seizure detection," *Clinical Neurophysiology*, vol. 116, no. 8, pp. 1785–1795, 2005.

[10] W. R. S. Webber, R. P. Lesser, R. T. Richardson, and K. Wilson, "An approach to seizure detection using an artificial neural network (ANN)," *Electroencephalography and Clinical Neurophysiology*, vol. 98, no. 4, pp. 250–272, 1996.

[11] R. Esteller, J. Echauz, T. Tcheng, B. Litt, and B. Pless, "Line length: an efficient feature for seizure onset detection," *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, no. 3, pp. 1707–1710, 2001.

[12] A. T. Tzallas, M. G. Tsipouras, and D. I. Fotiadis, "Automatic seizure detection based on time-frequency analysis and artificial neural networks," *Computational Intelligence and Neuroscience*, vol. 2007, 2007.

[13] A. T. Khan, I. Husain, and Y. U. Khan, "Seizure Onset Patterns in EEG and their Detection using Statistical Measures," *31st International COnference of the IEEE EMBS*, 2015.

[14] G. Chen, "Automatic EEG seizure detection using dual-tree complex wavelet-Fourier features," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2391–2394, 2014.

[15] A. S. Zandi, G. A. Dumont, M. Javidan, and R. Tafreshi, "An entropy-based approach to predict seizures in temporal lobe epilepsy using scalp EEG," *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, no. September 2009, pp. 228–231, 2009.

[16] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[17] L. A. Baccala and K. Sameshima, "Partial Directed Coherence: a new Concept in Neural Structure Determination," *Biological Cybernetics*, vol. 84, pp. 463–474, 2001.

[18] M. E. J. E. J. Newman, D. J. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences*, vol. 99 Suppl 1, pp. 2566–2572, 2002.

[19] P. Erdös and a. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[20] H. Poor and O. Hadjiliadis, *Quickest Detection*. 2008.

[21] V. V. Veeravalli and T. Banerjee, "Quickest Change Detection," *arXiv preprint arXiv:1210.5552*, pp. 1–53, 2012.

[22] C. Alippi and M. Roveri, "Just-in-time adaptive classifiers - Part I: Detecting nonstationary changes," *IEEE Transactions on Neural Networks*, 2008.

[23] M. Pollak, "Optimal Detection of a Change in Distribution," *The Annals of Statistics*, vol. 13, no. 1, pp. 206–227, 1985.

[24] I. Selesnick, "Constrained Least Square Design of FIR Filters without Specific Transition Bands," *IEEE Transactions on Signal processing*, vol. 44, no. 8, pp. 1879–1892, 1996.

[25] G. Niso, R. Bruña, E. Pereda, R. Gutiérrez, R. Bajo, F. Maesto, and F. Del-Pozo, "HERMES: Towards an integrated toolbox to characterize functional and effective brain connectivity," *Neuroinformatics*, vol. 11, no. 4, pp. 405–434, 2013.

[26] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[27] S. Amiri, A. Rabbi, L. Azinfar, and R. Fazel-Rezai, "A Review of P300, SSVEP, and Hybrid P300/SSVEP Brain- Computer Interface Systems," *InTech Open*, pp. 195–213, 2013.

[28] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[29] C. Jia, X. Gao, B. Hong, and S. Gao, "Frequency and Phase Mixed Coding in SSVEP-Based Brain Computer Interface," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 200–206, 2011.

[30] Q. Liu, K. Chen, Q. Ai, and S. Q. Xie, "Review: Recent development of signal processing algorithms for SSVEP-based brain computer interfaces," *Journal of Medical and Biological Engineering*, vol. 34, no. 4, pp. 299–309, 2014.

[31] B. Allison, T. Lüth, D. Valbuena, A. Teymourian, and I. Volosyak, "BCI Demographics : How Many (and What Kinds of) People Can Use an SSVEP BCI ?," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 2, pp. 107–116, 2010.

[32] B. Z. Allison, D. J. McFarland, S. Gerwin, S. D. Zheng, M. Moore Jackson, and J. R. Wolpaw, "Towards an Independent Brain-Computer Interface Using Steady State Visual Evoked Potentials," *Clinical Neurophysiology*, vol. 119, no. 2, pp. 399–408, 2008.

[33] M. Higger, M. Akcakaya, H. Nezamfar, G. Lamountain, U. Orhan, and D. Erdogmus, "A Bayesian Framework for Intent Detection and Stimulation Selection in SSVEP BCIs," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 743–747, 2015.

[34] M. Nakanishi, S. Member, Y. Wang, Y.-t. Wang, S. Member, Y. Mitsukura, T.-p. Jung, and S. Member, "Enhancing Unsupervised Canonical Correlation Analysis-Based Frequency Detection of SSVEPs by Incorporating Background EEG," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3053–3056, 2014.

[35] Y. U. Zhang, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Frequency Recognition in SSVEP-Based BCI Using Multiset Canonical Correlation Analysis," *International Journal of Neural Systems*, 2013.

[36] R. Elwell, R. Polikar, and S. Member, "Incremental Learning of Concept Drift in Non-stationary Environments," *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1517–1531, 2011.

[37] L. Schmidt-thieme and F. Hutter, "Beyond Manual Tuning of Hyperparameters," *Kunstliche Intelligen*, vol. 29, no. 4, pp. 329–337, 2015.

[38] M. S. Alencar, *Information Theory*. New York: Momentum Press, 2015.

[39] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley, 2000.

[40] G.Tec, "g.USBamp USB Biosignal Amplifier."

[41] I. W. Selesnick, L. Markus, and C. S. Bums, "Filters without Specified Transition Bands," *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 44, no. 8, pp. 1879–1892, 1996.

[42] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power SPectra: A Method Based on Time Averaging over Short, Modified Periodograms," *IEE Transaction on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.