PREDICTION OF CANINE EPILEPSY

BY

YOGATHEESAN VARATHARAJAH

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Advisers:

Professor Ravishankar K. Iyer
Professor Zbigniew T. Kalbarczyk

ABSTRACT

Seizure prediction is a problem in biomedical science which now is possible to solve with machine learning methods. A seizure prediction system has the power to assist those affected by epilepsy in better managing their medication, daily activities and improving the quality of life. Usage of machine learning algorithms and the availability of long term Intracranial Electroencephalographic (iEEG) recordings have tremendously reduced the complications involved in the challenging seizure prediction problem. Data, in the form of iEEG was collected from canines with naturally occurring epilepsy for the analysis and a seizure prediction system consisting of a machine learning based pipeline was implemented to generate seizure warnings when potential preictal activity is observed in the iEEG recording. A comparison between the different extracted features, dimensionality reduction techniques, and machine learning techniques was performed to investigate the relative effectiveness of the different techniques in the application of seizure prediction. The machine learning protocol performed significantly better than a chance prediction algorithm in all the analyzed subjects. Moreover, the analysis revealed subject-specific neurophysiological changes in the extracted features prior to lead seizures suggesting the existence of a distinct, identifiable preictal state.

To my parents, for their love and support.

ACKNOWLEDGMENTS

This thesis was built upon the foundation laid by many people, without whom I would not have been successful in my efforts. I am grateful for the support and guidance offered by my advisers, Ravishankar K. Iyer and Zbigniew T. Kalbarczyk. This effort would not have been possible without the advice and guidance of Benjamin Brinkmann (Mayo Clinic) and Gregory Worrell (Mayo Clinic), who were instrumental in the development of this research project. I sincerely appreciate the efforts of my colleagues, Homa Alemzadeh and Brent Berry (Mayo Clinic), who played pivotal roles during various stages of this project's development. Additionally, I thank my family and friends for supporting me during this process.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS vii

CHAPTER 1 INTRODUCTION 1
1.1 Motivation 1
1.2 Approach 2
1.3 Contributions/Results 2
1.4 Organization 3

CHAPTER 2 BACKGROUND 4
2.1 Epilepsy 4
2.2 Seizures 4
2.3 Electrocorticography (ECoG)/Intracranial Electroencephalography (iEEG) 6
2.4 Normal EEG Brain Rhythms 7
2.5 Abnormal EEG Brain Rhythms 8

CHAPTER 3 RELATED WORK 9

CHAPTER 4 SUBJECTS AND DATA COLLECTION 14
4.1 Subjects 14
4.2 Device 14
4.3 Data 15

CHAPTER 5 THE PREDICTION PIPELINE 19
5.1 Feature Extraction 19
5.2 Dimensionality Reduction 22
5.3 Machine Learning 26
5.4 Forecasting and Assessment 35
5.5 Execution of the Prediction Pipeline 38

CHAPTER 6 EVALUATION 40

CHAPTER 7 RESULTS AND DISCUSSION 43
7.1 Forecasting Results 45
7.2 Preictal Changes 48

CHAPTER 8 CONCLUSION 54
8.1 Future Directions 54

REFERENCES 56

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
AUC	Area Under Curve
BCI	Brain Computer Interface
CO	Convex Optimization
CSF	Cerebral Spinal Fluid
DCES	Direct Current Electrical Stimulation
DR	Dimensionality Reduction
ECoG	Electrocortigography
EEG	Electroencephalogram
FFT	Fast Fourier Transform
FLE	Frontal Lobe Epilepsy
fMRI	Functional Magnetic Resonance Imaging
FPR	False Positive Rate
iEEG	Intracranial Electroencephalogram
ILA	Implantable Lead Assembly
ITU	Implantable Telemetry Unit
KKT	Karush-Kuhn-Tucker
MCD	Malformations of Cortical Development
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MST	Multiple Subpial Tran-sections
MTS	Mesial Temporal Sclerosis
PAD	Personal Advisory Device
PCA	Principal Component Analysis
PCR	Principal Component Regression
PET	Positron Emission Tomography
PIB	Power In Bands
PLS	Partial Least Squares

PLSR	Partial Least Squares Regression
QP	Quadratic Programming
RF	Random Forests
RFC	Random Forests Classifier
ROC	Receiver Operating Characteristics
TLE	Temporal Lobe Epilepsy
TMCO	Time Domain Correlation
SP	Specificity
SPCO	Spectral Coherence
SPECT	Single Photon Emission Tomography
SS	Sensitivity
SVM	Support Vector Machines

CHAPTER 1

INTRODUCTION

1.1 Motivation

Approximately 0.5-1% of the world's population are affected by epilepsy [1] and World Health Organization (WHO) reports that epilepsy accounts for nearly 1% of the entire global burden of diseases [2]. Regardless of the pharmacotheraphy using anti-epileptic drugs, 20-40% of the people affected by epilepsy continue to suffer from seizures [3]. The uncertainty in the occurrence of seizures is seen as the most significant cause of epilepsy related disability [4, 5, 6]. Consistent anxiety about when the next seizure will occur has been expressed by even patients with infrequent seizures [6]. Individualized epilepsy treatment could be made possible by the ability to predict seizures in a timely manner so that the patients could be warned and take medications only when required. Seizure forecasting has become a major research interest due to the potential clinical impacts [7].

Machine learning based approaches of seizure forecasting have the following steps [8, 9, 10, 11]. Measurements from the brain are taken in some form (different measurements include scalp EEG, iEEG, FMRI, etc.). Since raw measurements are usually very noisy and less revealing, they are transformed into features which summarize the important changes in the raw signals. A machine learning algorithm incorporating these features is then used to make predictions on seizure occurrence. The features to which the raw signals are converted are usually chosen in such a way that the features are active in seizure related activity. However, what these selected works fail to consider is the fact that the features can be subject dependent. A feature which is found to contribute to seizure related activity of one subject may not contribute to that of another. In this thesis, the possibility of these features being subject dependent has been considered and a framework to analyze which features

contribute in seizure related activities was devised.

1.2 Approach

In this study, iEEG recordings obtained from canine subjects were analyzed. Five dogs with naturally occurring epilepsy were instrumented with an implanted device recording 16 channels of continuous iEEG. Spontaneously occurring seizures were automatically detected and visually verified to create accurate long-term seizure catalogs. Long-term, continuous, iEEG records (ranging from 6.5 to 15 months) containing multiple seizures were evaluated (data freely available on the iEEG portal [12, 13]). Three features, namely powers in frequency bands (PIB), time domain correlation between channels (TMCO) and spectral coherence (SPCO) between selected pairs of channels were extracted from iEEG. Dimensionality reduction was used to extract a small number of useful correlates from each of these features. These correlates were then visualized around the seizures to identify preictal signatures. As a second step, machine learning was used to map the extracted features, including the signatures to predict the onset of future seizures. Predictions of the machine learning classifiers were statistically analyzed and compared using standard metrics such as Area Under the Curve (AUC, [14]) and p-value test. In the visual analysis, different features were found to be active in seizure related activities of different subjects. Machine learning based analysis provided better outcomes (in terms of AUC metric) when the subject specific features were used in developing the predictor. This result confirms that the features which are active in seizure related activities are subject dependent.

1.3 Contributions/Results

1. An analytic technique to visually identify preictal signatures: A subset of the features, PIB, TMCO and SPCO showed preictal changes in all the subjects analyzed. Although, the specific features which showed preictal signatures as well as the signatures are different for each subject.

2. A machine learning based pipeline for analyzing different features and forecasting seizures: The framework contains isolated blocks yielding the ability to apply different techniques in each block. The pipeline produced predictions which were significantly better than a chance (50-50) predicting algorithm for all the subjects analyzed.

1.4 Organization

The rest of the thesis is organized as follows. Chapter 2 provides a background on epilepsy, seizures and EEG. Chapter 3 describes previous works on seizure prediction and explains how the objectives of this work were drawn from those. Chapter 4 describes the subjects used in this study and the data collection methods. Chapter 5 explains the technical and implementation details behind the development of the seizure prediction pipeline. Chapter 6 explains the approach used to evaluate the performance of the seizure prediction pipeline. Chapter 7 describes the results obtained using the experiments and compares the performance of the different techniques used in the pipeline. Chapter 7 also discusses the observations in the obtained results suggesting distinct preictal states. Chapter 8 presents the conclusions drawn from the thesis and discusses the future directions for this work.

CHAPTER 2

BACKGROUND

2.1 Epilepsy[1]

Epilepsy has been considered as a superficial mental disorder for centuries. However, today, epilepsy is identified as a neurological disorder of the central nervous system. The underlying physiological phenomena that cause epilepsy still remain unknown. However, it is widely observed that epilepsy is common among those who had undergone brain related injuries or diseases. In children and young adults, genetic disorders, congenital abnormalities, and birth trauma affecting the brain are commonly considered as the causes of epileptic symptoms. On the other hand, in mature adults and the elderly, strokes, tumors, and cerebrovascular disease are considered the causes.

Although epilepsy is an increasingly worsening disorder (i.e., each seizure damages the brain), those affected by epilepsy are capable of a standard career and family lives. However, they are not advised to engage in activities such as driving, swimming, etc., during which the occurrence of a seizure episode could lead to death. Apart from these, the side effects of anti-epileptic medication, recurring episodes of loss of consciousness and motor control, and the general misconception about the disorder create clinical and psychological barriers.

2.2 Seizures

A seizure can be described as a combination of unintentional changes in behavior, movement, sensation and consciousness as a result of abnormal brain activity. Seizures can be epileptic seizures or non-epileptic seizures. Epilep-

[1] Adapted from [15].

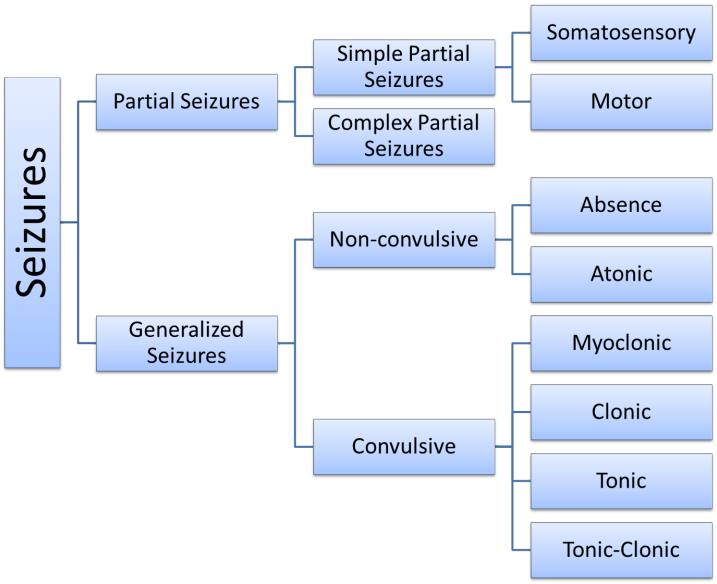

Figure 2.1: Different types of seizures

tic seizures occur as a result of an abnormal brain activity characterized by synchronized abnormal and excessive electrical activity. On the other hand, non-epileptic seizures occur in response to an external disturbance to the central nervous system such as alcohol withdrawal, drug abuse, acute illness, sleep deprivation or in the context of psychological trauma.

Different types of seizures are listed in Figure 2.1. Different treatments are needed for each type of seizure and thus the ability to distinguish among them is crucial. Two major types of seizures are partial seizures and generalized seizures. While partial seizures are localized to a part of the brain, generalized seizures involve the entire brain.

2.2.1 Partial Seizures

In a partial seizure, epileptic activity is contained in one part of the brain. Partial seizures that do not affect consciousness are classified as simple partial seizures, while those that do are classified as complex partial seizures. A simple partial seizure that originates in the somatosensory area of the brain is called a simple partial sensory seizure, while one that originates from the motor cortex is called a simple partial motor seizure.

2.2.2 Generalized Seizures

In a generalized seizure, epileptic activity involves the entire brain from the onset. Generalized seizures which lead to irregular muscular movements are classified as generalized convulsive seizures, while those that do not, are classified as generalized nonconvulsive seizures. Depending on the state of consciousness after the seizure, we can further categorize convulsive seizures into the myoclonic, clonic, tonic, and tonic-clonic types.

Nonconclusive seizures that result in the loss of consciousness, eye blinking, staring, and other minor facial movements are called absence seizures. Generalized nonconclusive seizures that do not lead to a loss of consciousness are called atonic seizures.

2.2.3 Treatment of Epilepsy

Epilepsy affects individuals with variable degrees of severity. Between 70-80% of epilepsy patients suffer from seizures whose severity and frequency can be limited with the use of antiepileptic drugs, each of which essentially limits the capacity of neurons to fire at excessive rates. The correct classification of these patients' seizures is crucial since different seizure types require specific drug regiments. In fact, the use of the wrong antiepileptic drug may exacerbate certain types of seizures. The remaining 20-30% of epilepsy patients suffer from seizures that are refractory to medication. These patients seek alternative treatment options that include surgery, vagus nerve stimulation, and ketogenic diets.

2.3 Electrocorticography (ECoG)/Intracranial Electroencephalography (iEEG)[2]

The electro-physiological monitoring which uses electrodes directly implanted on the exposed region of the brain to record electrical activity from the cerebral cortex, is called Electrocorticography (ECoG), or Intracranial Electroencephalography (iEEG). Conventional Electroencephalogram (EEG) on the

[2]Adapted from [16, 17].

other hand, monitors this activity from outside the skull. ECoG can be performed in two ways, (1) in the operating room during surgery or (2) outside of surgery. ECoG is an invasive procedure as it requires a surgical incision into the skull. ECoG/iEEG possesses a clear advantage over neuro imaging methods because of its high spatial and temporal resolution. Further, the contaminations due to muscle movement, eye blinks which regularly impair the quality of scalp EEG, is minimal in ECoG/iEEG. However, the typical characteristics of EEG and ECoG/iEEG are comparable, i.e. the typical constituents of an EEG recording can also be observed in ECoG/iEEG.

2.4 Normal EEG Brain Rhythms

A typical EEG recording contains the following different rhythms.

- Alpha rhythm - EEG activity with frequency between 8-13 Hz that is prominent in the occipital regions of normal, relaxed adults whose eyes are closed
- Beta rhythm - EEG activity with frequency exceeding 13 Hz that is most prominently observed in the frontal and central regions in adults, but may also be generalized
- Theta rhythm - EEG activity with frequency between 4-7 Hz; this activity is abnormal in awake adults, but commonly observed in sleep and children below the age of 13 years
- Delta rhythm - The delta rhythm exhibits a frequency below 3 Hz and amplitudes that exceed those of all other rhythms; it is most prominent frontally in adults and posteriorly in children in the third and fourth stages of sleep
- Mu rhythm - The mu rhythm refers to EEG activity with frequency between 7-11 Hz that is most prominently observed in the central region; mu activity is suppressed by movement (fist clenching), imagined movement, or tactile stimulation; in contrast, it is enhanced by immobility and heightened attention

- Lambda waves - Transient sharp waves lasting for a duration of approximately 0.25 seconds that occur in the occipital region whenever an adult scans a visual field with horizontal eye movement

- Sleep-spindles, K-complexes, and vertex Waves - These are unique waveforms observed only during the different stages of sleep

2.5 Abnormal EEG Brain Rhythms

Abnormal EEG activity is any activity that is prevalent in the EEG of groups of people with neurological or other disease complaints, and absent from that of normal individuals. Abnormal EEG may be an unusual waveform as well as the absence or deviation of normal EEG from well-documented limits on frequency, amplitude, morphology, localization, and reactivity. The following can be considered as the constituents of an abnormal EEG.

- Spike and sharp waves - Spike waves are transients with pointed peaks exhibiting durations between 20-70 milliseconds; sharp waves are similar to spike waves, but exhibit longer durations typically between 70-200 milliseconds

- Periodic discharges - Periodic discharges refer to time-limited bursts that are repeated at a certain rate; bursts may exhibit a variety of durations, frequencies, amplitudes, morphologies, and localizations

- Rhythmic hyper synchrony - Rhythmic hyper synchrony refers to rhythmic activity emerging from a quiescent background and exhibiting unusual frequency, amplitude, morphology and localization of any degree; rhythmic activity may either be continuous or intermittent

- Electro cerebral inactivity - Electro cerebral inactivity refers to a variable length period not caused by instrumental or physiological artifacts that exhibits extreme attenuation of the EEG relative to a patient-specific baseline

CHAPTER 3

RELATED WORK

With the availability of well established signal processing and statistical methods, techniques used on seizure prediction vary broadly. Understandably, features which characterize the changes in EEG are being given the most importance from the very beginning of seizure prediction research. These features could be linear, non-linear, univariate (single channel) or multivariate (multiple channels) and the changes in these features could be characterized using thresholding or machine learning.

Early work on seizure prediction dating back to the 1970s was performed based on surface EEG recordings in the absence of seizures using linear approaches to extract seizure precursors [18]. Rogowski et al. [19] and Salant et al. [20] have used an autoregressive modeling to identify preictal changes within 6 s prior to seizure onset. Siegel et al. [21] identified characteristic changes between the one-minute periods prior to a seizure and similar baseline periods for individual patients. In this study, statistical confidence of the findings was assessed and the influence of different vigilance states was discussed.

Spike occurrence rate in the EEG was evaluated as potential preictal change in a few studies. While Lange et al. [22] identified a decreased focal spiking rate and an increased rate of bilateral spikes prior to seizures; other studies showed no visible changes in spike rates before seizures [23, 24, 25].

Advances in the mathematical theories of non-linear systems in the 1980s opened up a lot of new approaches in modeling dynamically changing complex systems. Time series analytics became applicable in seizure prediction with the availability of long term EEG recordings. In the early 1990s, Iasemidis et al. [26] found that the largest Lyapunov exponent could be a seizure precursor as its behavior was observed to change during the preictal stages of intracranial EEG recordings. Martinerie et al. [27] reported decrease in the correlation density before the seizures preictally. They also developed a mea-

sure called dynamic similarity index which quantified changes in dynamics relative to a constant preictal reference window. Their works identified decreased dynamic similarity in preictal regions of intracranial and scalp EEG recordings [28, 29, 30]. However, these studies focused only on the preictal regions of the recordings and neglected the baseline characteristics on the identified measures. Therefore the evaluation of the investigated measures' applicability in seizure prediction was incomplete due to the unknown specificities.

Later works analyzed the specificities by comparing the measures in both interictal and preictal regions. Navarro et al. [31] observed that in selected examples of their subjects, the similarity measures showed more frequent drops before seizures than during interictal regions. Phase synchronization between different areas in brain was evaluated as a seizure precursor by Mormann et al. [32].These results were confirmed by two works of Le Van Quyen et al. [33, 34] on neocortical epilepsy. Chavez et al. [35] reported that preictal changes in phase synchronization occur predominantly in beta band based on their analysis after bandpass filtering of EEG.

Several measures including the correlation dimension [36, 37] (as a measure for dynamical complexity), dynamical entrainment [38] (defined as the convergence of largest Lyapunov exponents in certain selected channels), accumulated signal energy [39, 40], simulated neuronal cell models [41] or phase synchronization [42] were shown to be suitable for differentiating interictal from preictal data.

A number of studies published starting from 2003 raised skepticism in seizure prediction as the earlier optimistic studies could not be reproduced. De Clercq et al. [43] and Winterhalder et al. [44] questioned the optimistic results obtained using similarity index [30]. Correlation dimension was reevaluated in [45, 46] and the previous work on it [37] was challenged. Similarly, the work on accumulated energy [39] could not be reproduced in [47, 48]. Studies by Lai et al. [49, 50] raised doubts about the suitability of the Lyapunov exponent [51] for seizure prediction.

McSharry et al. [52] questioned the performances of nonlinear features like correlation density [27]. Upon reevaluation, the studies showed that this measure was more or less a reflection of the variance in EEG signals. As a suggestion for further studies on nonlinear measures, the authors pointed out that usage of nonlinear or complicated features should not be taken into

account unless it can be shown that these measures indeed outperform simple linear measures.

Starting from 2004, machine learning algorithms were being used for generation of seizure warnings instead of manual labeling. Machine learning methods are able to map the complex relationships between the features extracted from the EEG recordings to seizure annotations. This ability of machine learning techniques remarkably increased the prediction capability when used with the features which showed preictal changes in earlier studies. Table 3.1 and Table 3.2 list the different attempts on seizure prediction starting from 2004. Features used in each study, characteristics of the prediction method and its validation and some comments on the work are summarized in the tables.

Feature engineering has been explored multiple times from the very beginning of seizure prediction research. Many linear, non-linear features and their combinations have been tried as bio-markers for epilepsy with no specific attempt providing extraordinary results. Therefore, we believe that the basic assumption of a single or combination of features being the bio-markers for all individuals is wrong. Also, usage of machine learning for warning generation is very important to achieve the best trade-off between sensitivity and specificity. Therefore, in this work, we have used a machine learning based approach to identify patient specific bio-markers and used them for seizure prediction to achieve best results. In doing that, we also performed a preictal analysis to assert that the bio-markers identified from the machine learning based analysis do indeed show preictal changes leading to seizures.

Table 3.1: Seizure prediction studies with statistically validated results, published between 2014 and 2013 (Based on [53])

Year	Authors	Features	Prediction	Result	Comments
2014	Bandarabadi et al. [54]	Bivariate spectral power	SVM	iEEG: SS: 0.736±0.25, FPR: 0.28±0.28/hr, Scalp EEG: SS: 0.677±0.22, FPR: 0.39±0.37 /hr	Patient specific predictors; no preictal analysis
2014	Moghim et al. [55]	43 univariate features	SVM	SS: 0.91±0.13, SP: 0.99±0.0008	Patient specific analysis was not performed; preictal changes were reported
2014	Eftekhar et al. [56]	Repeating EEG patterns	Thresholding	Temporal lobe: SS: 0.67, FPR: 0.04/hr, Frontal lobe: SS: 0.72, FPR: 0.61/hr	Patient specific analysis was not performed; no analysis on preictal duration or changes
2014	Aarabi et al. [57]	Univariate computational model parameters	Rule based decision	SS: 0.87 - 0.926, FPR: 0.2 - 0.15/hr	Machine learning was not used for decision making; model based seizure prediction, patient specific analysis
2014	Zheng et al. [58]	Bivariate mean phase coherence	Thresholding	SS: 0.70, FPR: 0.15/hr	Machine learning was not used; independent test data was not used to test the predictor, preictal changes were observed 30 minutes before the seizures
2013	Li et al. [59]	Spike rate	Thresholding	SS: 0.76, FPR: 0.09/hr	Machine learning was not used for decision making; patient specific threshold was used
2013	Gadhoumi et al. [60]	Wavelet energy and entropy	Linear discriminant analysis	SS: >0.85, FPR: <0.35/hr	Time in seizure warnings was high; worked only for 41% of the patients

FPR - False Positive Rate (per hour), SP - Specificity, SS - Sensitivity

Table 3.2: Seizure prediction studies with statistically validated results, published between 2013 and 2004 (Based on [53])

Year	Authors	Features	Prediction	Result	Comments
2013	Cook et al. [61]	Average energy, Teager-Kaiser energy, line-length	Decision trees and K-Nearest Neighbors	SS: 0.65 - 1.00	The largest iEEG database; machine learning was not used for decision making, preictal analysis was not performed
2012	Williamson et al. [62]	Time delayed univariate and multivariate correlations	SVM	AUC: 0.973	Patient specific machine learning was used; 30 minutes prediction window was used, no clear separation between training and testing sets
2012	Aarabi et al. [63]	6 non-linear features	Thresholding and Rule based decisions	SS: 0.799, FPR: 0.17/hr	Machine learning was not used for decision making; patient specific analysis was not done, many features were explored
2012	Park et al. [11]	Univariate spectral power	SVM	SS: 0.975, FPR: 0.27/hr	Smoothing of the seizure warnings was done using Kalman filter; very high false positive rate
2011	Kuhlman et al. [64]	Bivariate mean phase coherence	Fixed and dynamic thresholding	SS: 0.5 - 0.88, FPR: 0.64 - 4.691/hr	Machine learning was not used for decision making; very high false positive rate, results were not better than chance prediction
2009	Mirowski et al. [65]	6 linear and non-linear features	Logistic regression, convolutional networks and SVM	SS: 0.71, FPR: 0	Patient specific feature selection; comparison was done with a random predictor
2006	Sackellares et al. [66]	Short-term maximum Lyapunov exponent	Dynamic thresholding	SS: 0.80, FPR: 0.56/hr	Statistical validation method used was wrong; not clear if the predictions are better than chance

AUC - Area under ROC curve, FPR - False Positive Rate (per hour), SP - Specificity, SS - Sensitivity

CHAPTER 4

SUBJECTS AND DATA COLLECTION

4.1 Subjects

The NeuroVista seizure advisory system [67, 68] was implanted in canine subjects with naturally occurring epilepsy and spontaneous seizures. The dogs were housed in the University of Minnesota canine epilepsy monitoring unit and the University of Pennsylvania canine epilepsy monitoring unit. The subjects were continuously monitored (24 hours/day) with iEEG and video. Anti-epileptic medications were provided to the dogs during this study. Five dogs had an adequate number of seizures and prolonged interictal recordings suitable for analysis. Table 4.1 lists the different subjects and the number of seizures recorded on those subjects.[1]

4.2 Device

An implantable acquisition system (Figure 4.1) was utilized to record and store long-term, continuous iEEG recordings [67, 68]. This system consists of three components: (1) Implantable Lead Assembly (ILA); (2) Implantable Telemetry Unit (ITU); and (3) External Personal Advisory Device (PAD). Intracranial EEG signals are recorded in sixteen channels using the ILA

Table 4.1: Subjects and recorded seizures

Subject	1	2	3	4	5
Seizures	23	105	29	132	22

[1] Approval of the Institutional Animal Care and Use Committee (IACUC) at Mayo Clinic, the University of Minnesota, and University of Pennsylvania was obtained for acquisition of the data analyzed in this work.

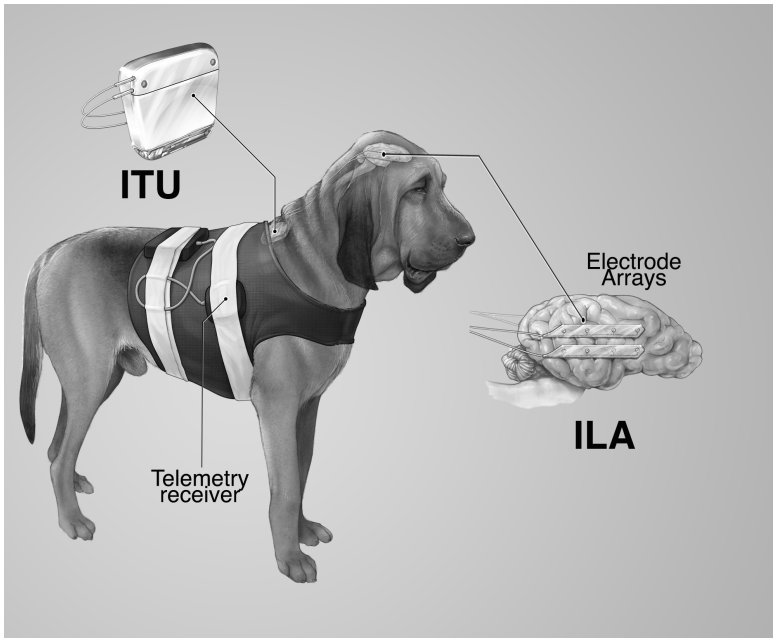

Figure 4.1: NeuroVista seizure advisory system

contacts. The ITU filters, amplifies, digitizes and wirelessly transmits the recordings to the external PAD. Approximately 1-hour is spent everyday on charging ITU using an external power supply. The PAD, which collects the iEEG data wirelessly from ITU, was fixed on the subject's back using a harness. The PAD was equipped with high accuracy seizure detection mechanisms [69] and alarming capabilities.

The recordings stored on the PAD were uploaded to a data center weekly. All detected seizures were verified via expert annotations and correlation with continuous video. Canine data recorded for this study are publicly available on the IEEG-portal [13].

4.3 Data

Continuous iEEG was recorded in each canine subject using 16 electrodes sampling at 400 Hz. Each subject's iEEG was recorded for different periods. Table 4.2 displays the time durations for which iEEG was recorded on each subject.

Due to reasons such as providing medication, disconnections between the different devices, human intervention, etc. a considerable part of the collected

Table 4.2: Summary of collected iEEG data

Subject	Data		
	Total (days)	Discontinuous (%)	Effective (days)
Subject 1	475.7	27.99%	342.5
Subject 2	451.8	52.77%	213.4
Subject 3	460	22.21%	357.8
Subject 4	287.4	41.47%	168.2
Subject 5	294.1	72.48%	80.9

data is discontinuous. Figure 4.2 shows a typical recording with discontinuities and seizures. The seizures tend to occur in clusters, i.e. a lead seizure followed by a number of follow-up seizures. C1, C2, C3 and C4 are such clustered seizures. Table 4.2 also displays the percentage of data which is discontinuous and the effective recording days for each subject.

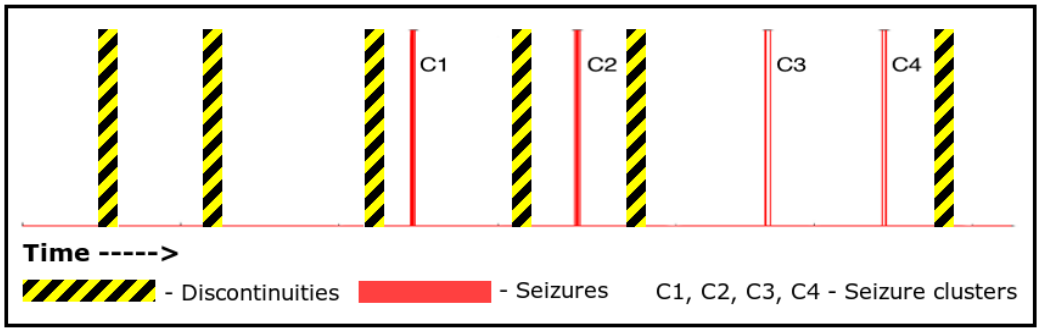

Figure 4.2: iEEG recording with discontinuities and clustered seizures

As the first step, discontinuous regions were eliminated from the recordings. Since there were too many small discontinuities, only the discontinuities which were longer than 20 seconds were eliminated. Smaller discontinuities were kept in the data assuming that the effect of them will be negligible in the analysis. The rest of the iEEG recording after eliminating the long discontinuities was used for analysis. The entire iEEG recording with seizures can be divided in to five different periods. The period where there is no seizure related activity is called the interictal period. The period before the seizure is called the preictal period. The duration of preictal effects is still unknown and can be varying. The period of seizure is called the ictal period. The period after the seizure is called the postictal period. The duration of postictal effects is also unknown. The period between interictal period and perictal

period is not of interest to us. The interictal activities are usually observed very far from the seizures, typically a week away from the seizure. Figure 4.3 explains the four important periods pictorially. There is hardly any difference between the iEEG of the preictal period beyond the 5-minute horizon of seizure and the iEEG of the interictal period. In order to distinguish preictal iEEG from interictal iEEG further analysis is required. The ability to distinguish preictal iEEG from interictal iEEG will enable the seizure warning system to generate seizure warnings when the transitions from interictal activity to preictal activity occurs.

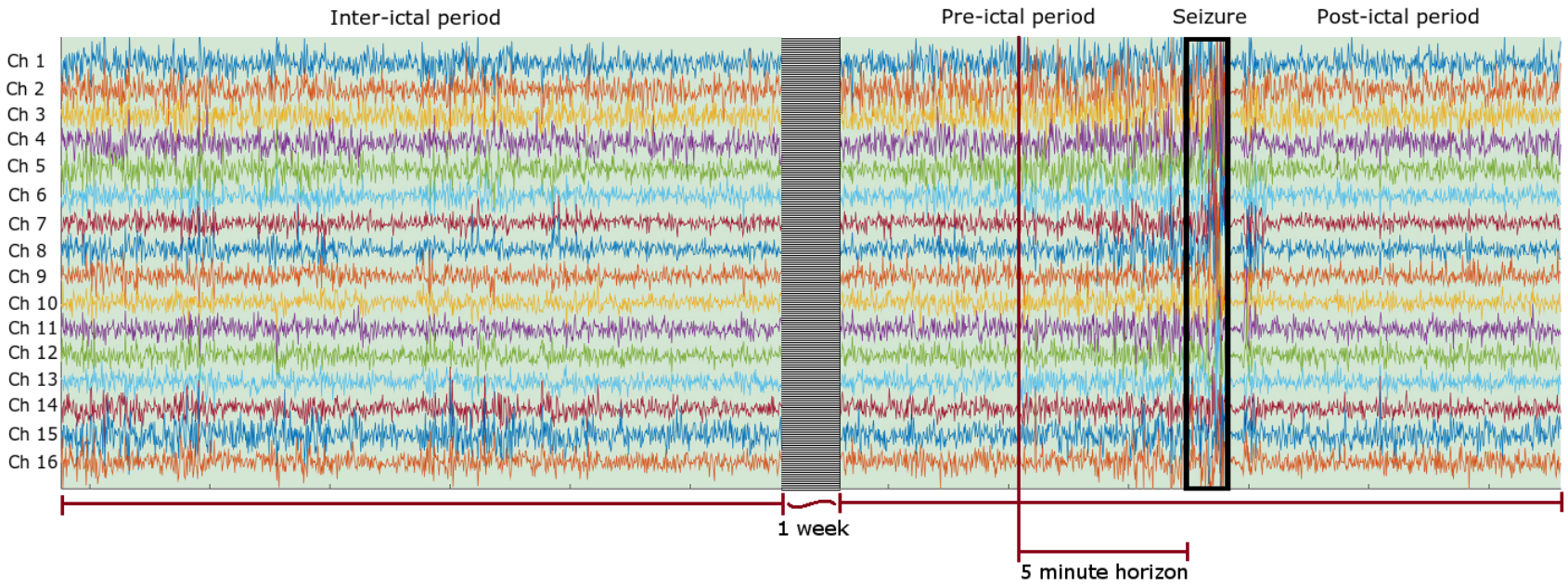

Figure 4.3: Different periods of an iEEG recording

CHAPTER 5

THE PREDICTION PIPELINE

For the purpose of analyzing the iEEG data in a sequential and structured fashion, we developed a pipeline consisting of different functionalities in each of its stages. The pipeline consists of four isolated components: feature extraction, dimensionality reduction, machine learning classifier and forecasting and assessment. The four components are individually isolated from each other so that different techniques can be employed on each component without altering the other components. Figure 5.1 shows the different components in the pipeline.

Following subsections explain each component in the pipeline in detail.

5.1 Feature Extraction

From Figure 4.3 we understand that characterizing the changes in the raw iEEG signals alone is not enough to identify the preictal signatures. Therefore, a transformation of the raw iEEG into some features is necessary. As seen from Chapter 3, researchers have used a large number of features to describe the changes in the EEG recordings. However, with the increasing usage of machine learning techniques for decision making, the quest for identifying complex feature descriptors have become less important. Machine learning techniques are capable of expressing the complex relationships between the

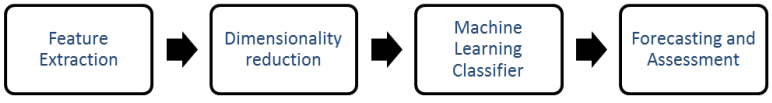

Figure 5.1: Predictive pipeline: Top-level flow diagram

simple features which can be extracted very easily. For that reason, we have opted to use simple features in our approach. For the purpose of comparing the capabilities of different features, we have extracted three simple features.

Worrell et al. [70] found that high-frequency activity is prominently observed in iEEG at seizure onsets. In the same work, they also found that there is significant correlation in the iEEG of some channels (electrodes) during seizure onsets, depending on the seizure origin. Since these features are significant in the seizure onset zones, we believe that there must be significant changes in these features in the preictal time of the same seizures as well. Therefore, we have decided to extract these features in the iEEG of the data segments that were extracted around the lead seizures. To account for the possible high-frequency activity, we have extracted powers in bands (PIB) as features. To account for the signal correlation between different channels, we extracted time domain correlations (TMCO) and spectral coherence (SPCO) between different pairs of channels.

5.1.1 Power in Bands (PIB)

PIB features were calculated for each channel, as power in each frequency in 0-50 Hz (51 PIB), power in 5 Hz bands in 50 - 100 Hz (10 PIB) and power in 10 Hz bands in 100 - 180 Hz (8 PIB). The higher resolution used in calculating the powers in the lower part of the spectrum reflects the power law in EEG [71]. Power of the signal was calculated from the Discrete Fourier Transform (DFT) obtained by FFT. If the time domain signal of length N is represented as x, DFT of x, X is obtained from Equation 5.1. Power of the signal within a frequency range $[f1 - f2]$, P_r is obtained from Equation 5.2.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-jkn2\pi}{N}} \tag{5.1}$$

$$P_r = \frac{1}{N} \sum_{i=f1}^{f2} X(i)^2 \tag{5.2}$$

This produces 69 PIB features for a channel and therefore 16 * 69 = 1104 PIB features for a one-minute clip.

5.1.2 Time Domain Correlation (TMCO)

TMCO features were calculated as the linear absolute correlation coefficient between different pairs of channels in a one-minute clip. The correlation coefficient r between two channels x and y is calculated from Equation 5.3, where μ and s represent the sample mean and sample standard deviation respectively.

$$P_r = \frac{1}{N-1} \sum_{i=1}^{N} \frac{x_i - \mu_x}{s_x} \frac{y_i - \mu_y}{s_y} \tag{5.3}$$

TMCO features were calculated for each distinct pair of the 16 channels, i.e. 120 pairs. This produces 120 TMCO features for a one-minute clip.

5.1.3 Spectral Coherence

SPCO features were calculated as the magnitude squared coherence between different pairs of channels in a one-minute clip. The coherence function (at frequency w) of two signals x and y is given by Equation 5.4.

$$C_{xy}(w) = \frac{P_{xy}(w)}{\sqrt{P_{xx}(w) P_{yy}(w)}} \tag{5.4}$$

where P_{xx}, P_{yy} are power spectral densities of x, y and P_{xy} is the cross power spectral density of x and y.

Power spectral density (PSD) of x is given by Equation 5.5, where X is the DFT of x and X^* is the complex conjugate of X. Cross power spectral density of x and y is calculated by simple replacing one of the X's by Y where Y is the DFT of y.

$$P_{xx} = XX^* \tag{5.5}$$

For an FFT length of N, since we only consider real frequencies (which is just half of the spectra) and DC (0Hz), we will get a PSD of length $\frac{N}{2} + 1$. Since PSDs are of length $\frac{N}{2}+1$, spectral coherence will also be of length $\frac{N}{2}+1$. Magnitude squared coherences of pairs of channels at each real frequency is taken as the SPCO features.

In this thesis, we have used an FFT length of 128, which produces 65 SPCO features per channel pair. If the spectral coherences of every pairs of

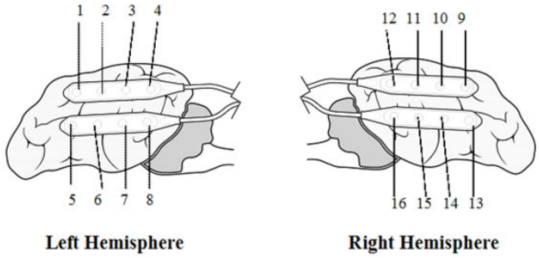

Figure 5.2: Electrode positions as implanted

channels are considered, there will be 120 * 65 = 7800 SPCO features per one-minute clip, which is too much to handle. For this reason, we considered selected pairs of channels to extract SPCO features. Figure 5.2 describes the electrode placement in the brains of canine subjects. We considered the adjacent pairs (12), vertical pairs (8) and left-right pairs (8) to calculate SPCO features according to this map. This produces 28 potential pairs and therefore 28 * 65 = 1820 SPCO features.

5.2 Dimensionality Reduction

Since analyzing all the features in the predictive pipeline seemed infeasible due to the size of the dataset, applying dimensionality reduction on these features was considered. To account for the differences in the dimensionality reduction techniques, we used two dimensionality reduction techniques, one unsupervised and one supervised technique. Principal component analysis (PCA) is an unsupervised dimensionality reduction technique used extensively in many applications [72]. On the other hand, partial least squares (PLS) regression is a supervised dimensionality reduction technique, which finds those constituents in the features which largely contribute to the discriminability of the different classes in the dataset [73].

PLSR and PCA are both methods to model a response variable when there are a large number of predictor variables, and those predictors are highly correlated or even collinear. Both methods construct new predictor variables, known as components, as linear combinations of the original predictor variables, but they construct those components in different ways. PCA creates components to explain the observed variability in the predictor variables,

without considering the response variable at all. On the other hand, PLSR does take the response variable into account, and therefore often leads to models that are able to fit the response variable with fewer components. Whether or not that ultimately translates into a more parsimonious model, in terms of its practical use, depends on the context.

5.2.1 Principal Component Analysis (PCA)

In summary, PCA reduces the dimension of the data by finding a few orthogonal linear combinations (the Principal Components or PCs) of the original variables accounting for the largest variance. Consider a unit vector u and a point x. The length of the projection of x onto u is given by $x^T u$, i.e., when $x^{(i)}$ is a point in our experimental dataset, its projection onto the direction vector u is distance $x^T u$ from the origin. Thus, the unit vector u which maximizes the variance of the projections of all $x^{(i)}$'s, is obtained by maximizing:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x^{(i)^T}u)^2 &= \frac{1}{n}\sum_{i=1}^{n}u^T x^{(i)}x^{(i)^T}u \\ &= u^T(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}x^{(i)^T})u\end{aligned}$$

It is easily identifiable that maximizing this criteria subject to $\| u \|_2 = 1$ gives the principal eigenvector of $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}x^{(i)^T}$, which is just the empirical covariance matrix of the experimental data.

To summarize, we have found that the one-dimensional approximation of the data could be obtained by choosing u to be the principal eigenvector of Σ. More generally, a q-dimensional ($q < p$) approximation of the data could be obtained by choosing a q-dimensional subspace spanned by $u_1, \cdots, u_q$ to be the top q eigenvectors of Σ. The u_is now form a new, orthogonal basis for the approximated data.

Then, the representation of $x^{(i)}$ in this basis is obtained as $y^{(i)}$s by,

$$y^{(i)} = \begin{pmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ \vdots \\ u_q^T x^{(i)} \end{pmatrix} \in \mathcal{R}^q \tag{5.6}$$

Thus, whereas $x^{(i)} \in \mathcal{R}^p$, the vector $y^{(i)}$ now gives a lower, q-dimensional, approximation/representation for $x^{(i)}$.

5.2.2 Partial Least Squares (PLS) regression[1]

Notations

The observations are denoted by $Y_{I\times K}$, which contains I independent observations each of K variables. The matrix $X_{I\times J}$ contains the values of J predictors collected on these I observations.

Goal

The objective here is to predict Y from X and to be able to describe how they are related. The very common case is when Y is a vector and X is a full rank matrix. In that case, an ordinary multivariate regression will suffice. However, when the number predictors (J) is larger than the number of observations (I), X will very likely be singular and therefore the previous solution is not feasible. There can be several solutions to this problem. The very obvious solution is to eliminate some predictors using a stepwise method. Another solution could be to perform PCA on the matrix X and then use small number of principal components as regressors. Although the second approach looks feasible, the decision on the optimum number of predictors has to be made. It is possible to keep the first n principal components, however they are chosen so that they explain X rather than Y and they may not be relevant for Y.

In contrast, PLS regression identifies components from X that are relevant to Y as linear combinations of the predictors. Specifically, PLS regression performs a simultaneous decomposition of X and Y with the requirement

[1] Adapted from [74].

that the components explain the covariance between X and Y as much as possible. This is followed by a regression step where this decomposition is used to predict Y.

Simultaneous decomposition of predictors and dependent variables

In PLS regression, both X and Y are decomposed as a product of common orthogonal factors and specific loadings. Thus, the independent variables are decomposed as $X = TP'$ with T consisted of orthogonal set of vectors (i.e. $T' \times T = I$, some variations of the technique do not require T to have unit norms). In PCA, the corresponding T is called the score matrix, and P the loading matrix (in PLS regression the loadings are not orthogonal). Likewise, Y is estimated as $\hat{Y} = TBC'$ where B is a diagonal matrix with the "regression weights" as diagonal elements (see below for more details on these weights). The columns of T are the latent vectors. When their number is equal to the rank of X, they perform an exact decomposition of X. Note, however, that they only estimate Y (i.e., in general $\hat{Y}$ is not equal to Y).

PLS regression and covariance

The latent vectors could be chosen in a variety of ways. According to the previous formulation, any set of orthogonal vectors spanning the column space of X could be used to play the role of T. In order to specify T, additional conditions are required. For PLS regression this amounts to finding two sets of weights w and c in order to create (respectively) a linear combination of the columns of X and Y such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors $t = Xw$ and $u = Yc$ with the constraints that $w'w = 1$, $t't = 1$ and $t'u$ be maximal. When the first latent vector is found, it is subtracted from both X and Y and the procedure is re-iterated until X becomes a null matrix.

A PLS regression algorithm

Here, we provide a sketch of the PLS regression algorithm. Two matrices $E = X$ and $F = Y$ are initially created. These matrices are then column centered and normalized. SS_X and SS_Y denote the sum of squares of these matrices.

Initially, the vector u is initialized with random values. (In what follows the symbol := means "assign and normalize the result of the operation".)

1. $w := E'u$ (estimate X weights).
2. $t := Ew$ (estimate X factor scores).
3. $c := F't$ (estimate Y weights).
4. $u := Fc$ (estimate Y scores).

Upon convergence of t, the value of b is computed as $b = t'u$ and then b is used to predict Y. The factor loadings for X are calculated as $p = E't$. Further, the effect of t is subtracted out from E and S as $E = E - tp'$ and $F = F - btc'$. The vectors t, u, w, c, and p are then stored in the corresponding matrices, and the scalar b is stored as a diagonal element of B. The sum of squares of X (respectively Y) explained by the latent vector is computed as $p'p$ (respectively b^2), and the proportion of variance explained is obtained by dividing the explained sum of squares by the corresponding total sum of squares (i.e., SS_X and SS_Y). The stopping criterion is met when E is a null matrix, in which case the whole set of latent vectors has been found. Otherwise the procedure is reiterated from step 1 until the stopping criterion is met.

5.3 Machine Learning

For the purpose of comparing the performance of the different classifiers, this block was implemented using three different Machine Learning (ML) classifiers: Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Random Forests Classifier (RFC). The decision to use these three classifiers for analysis was influenced by existing literature and the results of the Kaggle seizure prediction challenge [75].

5.3.1 Support Vector Machines[2]

Let us assume that our dataset is $(x_1, y_1), \ldots, (x_n, y_n)$, with given labels $y_i \in \{1, -1\}$. Our goal is to find the hyperplane $w^T x + b = 0$ (where x is any

[2]Based on Machine Learning course notes from SUNY at Stony Brook [76].

of $(x_1, \ldots, x_n)$), characterized by parameters (w, b), which separates the two classes in the dataset. This hyperplane should satisfy the conditions below:

1. $\min_{i \leq n} |w^T x_i + b| = 1$
2. $y_i(w^T x_i + b) \geq 0$ for all $i \leq n$
3. $|w|^2$ is minimum

It is possible that there is no hyperplane that separates the two classes in the dataset. However, for simplicity, let us assume that the dataset is linearly separable. Combining conditions 1 and 2 results in

$$y_i(w^T x_i + b) \geq 1 \text{ for all } i \leq n$$

Hence the problem can be formulated as

$$\text{minimize } \frac{1}{2}|w|^2$$

over all $w \in R^d$ and $b \in R$ subject to the conditions

$$y_i(w^T x_i + b) - 1 \geq 0 \text{ for all } i \leq n$$

This turns out to be a simple quadratic programming problem and there are algorithms of complexity $O(n^3)$ that can be used for solving this problem. However, when n and d are large even the best QP methods will fail. The SVM solution is obtained by applying KKT optimality conditions to the dual of the above QP problem.

KKT conditions for SVM

Lagrangean of the SVM optimization problem is given by

$$L(w, b, \lambda) = \frac{1}{2} \sum_{i=1}^{d} w_i^2 - \sum_{j=1}^{n} \lambda_j \left\{ y_j(w^T x_j + b) - 1 \right\}$$

And the KKT-conditions for optimality are

$$\begin{aligned}
\nabla_w L &= 0, \quad \text{i.e., } w = \sum_{j=1}^{n} \lambda_j y_j x_j \\
\nabla_b L &= 0, \quad \text{i.e., } \sum_{j=1}^{n} \lambda_j y_j = 0 \\
\lambda_j \left\{ y_j (w^T x_j + b) - 1 \right\} &= 0, \text{ for all } j \leq n
\end{aligned}$$

These conditions completely characterize the optimal plane. The first condition asserts that w must be a linear combination of the observed vectors x_j. The second condition asserts that, the coefficients of this linear combination must add up to 0. The complementary condition asserts that the only non-zero Lagrange multipliers λ_j are those associated to the vectors x_j right on the margin, i.e.

$$y_j (w^T x_j + b) = 1$$

The corresponding x_j are called *support vectors* and they are the only ones needed because

$$w = \sum_{j \in J_0} \lambda_j y_j x_j$$

where $J_0 = \{j : x_j \text{ is a support vector}\}$.
The support vectors also satisfy that their distance from the separating plane is $\rho = 1/|w|$. Typically, the number n of such vectors is very small compared to the size of the dataset.

The dual problem

According to the definition of duals, maximizing $W(\lambda)$, subject to the constraint that $\lambda \geq 0$, where

$$W(\lambda) = \min_x L(x, \lambda)$$

is equivalent to minimizing $f(x)$ subject to the $g_j(x) \geq 0$.

SVM solution

The λ (which are the Lagrange multipliers), which minimizes $L(w, b, \lambda)$ w.r.t. (w, b), must satisfy the condition that w is a linear combination of the x_j's with coefficients $\lambda_j y_j$ that must add up to zero. Thus, the dual formulation is obtained by replacing the conditions into $L(w, b, \lambda)$ as,

$$\text{maximize } W(\lambda)$$

where

$$W(\lambda) = \sum_{j=1}^{n} \lambda_j - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j (x_i^T x_j)$$

and the $\lambda \geq 0$ satisfying

$$\sum_{j=1}^{n} \lambda_j y_j = 0$$

With the optimal λ, w can be obtained as linear combination of x_j as given above. b can be obtained by using the fact that the plane must be in canonical position.

$$\min_{i \leq n} y_i (w^T x_i + b) = 1 = y_j (w^T x_j + b) \text{ for all } j \in J_0$$

which results in

$$b = y_j - < w, x_j >$$

Multiplying this by λ_j and adding j yields

$$b = \frac{- \sum_{i,j} \lambda_i \lambda_j y_i y_j < x_i, x_j >}{\sum_j \lambda_j}$$

5.3.2 Random Forests[3]

Random forests is a classification approach that is especially well suited for problems with many classes when large datasets are available for training. The random forests classifier is obtained by randomly creating a number of decision trees and deriving an ensemble decision from those decision trees. A pictorial depiction of this is showed in Figure 5.3.

[3]Based on Machine Learning course slides from University of British Columbia, [77].

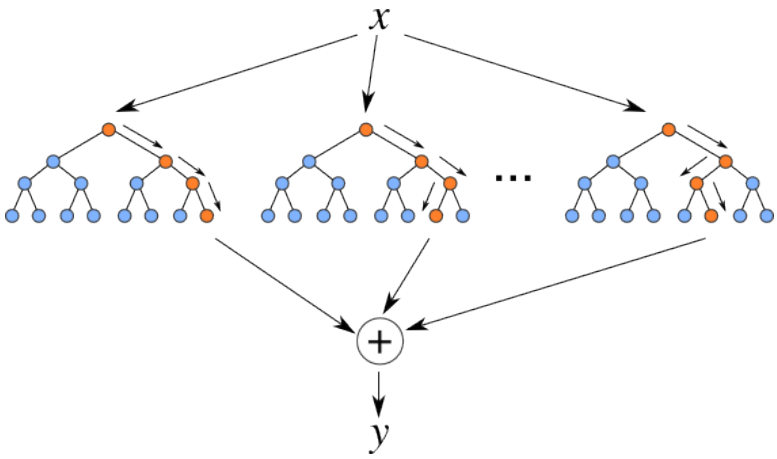

Figure 5.3: Random forests classifier obtained as an ensemble of decision trees

Suppose that we have a dataset $X_1, X_2, ..., X_N$ and corresponding labels $Y_1, Y_2, ..., Y_N$. Each X_i has p attributes and each Y_i is one of q different classes. A decision tree is made of a root node and several decision nodes (split points), where each node splits the dataset into a number of subtrees. The number of subtrees is typically the number of classes in the output labels. Further, the height of the decision tree is typically the number of attributes in the input dataset. At each node the splitting of the dataset is done using one of the attributes. The attribute that is used for splitting at a node is determined by the maximum information gain property. The information gain obtained by splitting using an attribute j is calculated using Equation 5.7.

$$I_j = H(S_j) - \sum_{i \in 1...q} \frac{|S_j^i|}{|S_j|} H(S_j^i) \tag{5.7}$$

where $H(X)$ is the entropy of the random variable X.

Random forest is built by using multiples of such decision trees accommodating randomness two different ways, (a) in the data used to build the trees and (b) in the attributes used for splitting. Following is the textual description of the random forest algorithm. Suppose that we want to generate B number of random decision trees.

1. For $b = 1 : B$

(a) Draw a bootstrap sample Z^* of size N from training data.

(b) Grow a random decision tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree until the minimum node size is reached.

i. Select m attributes at random from the p attributes.

ii. Pick the best attribute/split-point among them.

iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}^B$.

When a new data sample arrives, a decision on its class will be made by all the decision trees in the random forest. The likelihood probability for that data sample belonging to an output class is obtained by averaging the likelihood probabilities of the individual decision trees. The average can be obtained by taking the arithmetic or geometric mean.

5.3.3 Artificial Neural Networks[4]

Preliminaries

A complex system could be decomposed into simple elements for better understanding. Also simple elements could be merged to produce a complex system. Networks could be used for this purpose. The networks are typically characterized by the following components: a set of nodes, and connections between nodes.

The nodes operate computational units. They receive inputs, and process them to produce an output. This processing might be very simple (such as summing the inputs), or quite complex (a node might contain another network). The connections determine the information flow between nodes. They can be unidirectional, when the information flows only in one sense, and bidirectional, when the information flows in either sense.

One type of network sees the nodes as "artificial neurons". These are called Artificial Neural Networks (ANNs). An artificial neuron is a computational model inspired in the natural neurons (Figure 5.4). Natural neurons receive

[4]Adapted from [78] and Machine Learning course notes from Stanford University [79].

signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons.

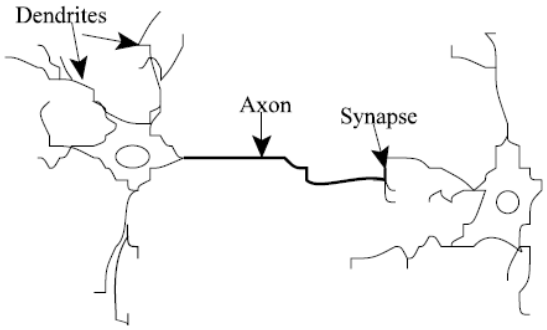

Figure 5.4: Natural neurons

Suppose that we have a set of labeled training examples $(x^{(i)}, y^{(i)})$ and we want to train a supervised learning algorithm to predict y using x. Neural networks can be used model the complex relationship using a non-linear hypothesis $h_{W,b}(x)$ with parameters W, b. Let's consider the simplest possible neural network, which only contains a single "neuron" (Figure 5.5).

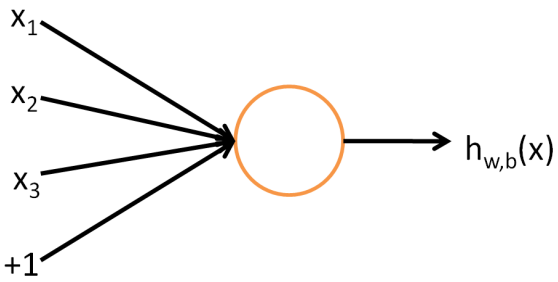

Figure 5.5: Single neuron neural network

This neuron is a computational unit which takes x_1, x_2, x_3 and a $+1$ bias term as inputs and outputs $h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{3} W_i x_i + b)$, where function f is called the activation function. Sigmoid (Equation 5.8) and hyperbolic tangent (tanh, Equation 5.9) functions are typically used as activation functions.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{5.8}$$

$$f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{5.9}$$

A neural network is built using many simple neurons such that the output of a neuron can be the input of another neuron. For instance, Figure 5.6 shows a sample neural network.

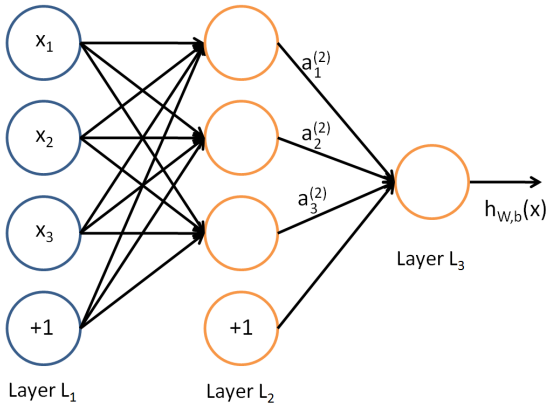

Figure 5.6: A 3-layer neural network

In Figure 5.6, circles denote inputs to the network. The leftmost, rightmost and the middle layers are called *input layer*, *output layer* and *hidden layer* respectively. There can be many hidden layers in a neural network. This neural network is called a three-layer, three-input, single output neural network.

Notation:

- n_l - Number of layers in the network.
- L_l - Layer l.
- $W_{ij}^{(l)}$ - Weight associated with the connection between unit j in layer l and unit i in layer $l+1$.
- $W^{(l)}$ - The matrix of weights connecting layer l and layer $l+1$.
- $b^{(l)}$ - Bias terms at layer l, s_l - number of units in layer l.

In general, a computational unit performs a weighted sum of all the inputs to that particular unit using the connecting weights and applies the activation function on this value to obtain the output value. This output value is passed to all the computational nodes in the next layer for their computation. If we

let the vector $z^{(l+1)}$ denote the weighted sums of all inputs to layer $l+1$ from layer l, and $a^{(l+1)}$ denote value of the activation function applied on $z^{(l+1)}$, the two quantities can be iteratively calculated from what is often called as the feed-forward propagation.

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)} \tag{5.10}$$

$$a^{(l+1)} = f(z^{(l+1)}) \tag{5.11}$$

Training the neural network

Suppose we have a training set $(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})$ of m training examples. Training of the network is done by minimizing a cost function based on the training labels and the outputs of the neural network. A typical cost function is the squared-error cost function, given by Equation 5.12

$$J(W, b) = \frac{1}{2} \left\| h_{W,b}(x) - y \right\|^2 \tag{5.12}$$

Our objective is to minimize $J(W, b)$ as a function of W and b. The backpropagation algorithm [80] with a batch gradient descent optimization is typically used for this purpose. The backpropagation algorithm performs a feed-forward pass, backward error propagation pass and gradient descent updates iteratively to reach the optimum configuration of the neural network. The backpropagation algorithm is as follows:

1. Perform a feed-forward pass using Equations 5.10 and 5.11.
2. For each output unit i in L_l, an intermediate quantity $\delta_i^{(n_l)}$ is calculated as

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} \frac{1}{2} \left\| h_{W,b}(x) - y \right\|^2 = -(y_i - a_i^{n_l}).f'(z_i^{(n_l)}) \tag{5.13}$$

3. Similar quantities for each lower layers are calculated as

$$\delta_i^{(n_l)} = (\sum_{j=1}^{s_l+1} W_{ij}^{(l)} \delta_j^{(l+1)} f'(z_i^{(n_l)})) \tag{5.14}$$

4. Partial derivatives for the gradient descent algorithm are calculated as

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) = a_j^{(l)} \delta_i^{(l+1)}$$
$$\frac{\partial}{\partial b_i^{(l)}} J(W,b) = \delta_i^{(l+1)}$$

5. Finally the gradient updates are performed as

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b)$$
$$b_i^{(l)} = b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W,b)$$

   where α is the learning rate.

This algorithm is repeated several times until the change in the gradient update becomes negligible (very close to zero). Once the algorithm has converged, the weights and the bias terms obtained at the last iteration will be the optimal weights and bias terms.

5.4 Forecasting and Assessment

The assessment block employs a forecasting algorithm to generate the warnings and generates fitness metrics to quantify the correctness of prediction. The triggering of the seizure prediction algorithm depends on the classifier's output, i.e. the likelihood probability of a data sample exceeding a predefined threshold. This threshold is chosen such that the total time spent on warning is maintained under the tolerance levels. The warning is kept for a period of 90 minutes once it is triggered. The warnings which occur within this 90 minute period are combined and the warning time is extended accordingly. The warning period essentially means that, seizures could occur at any time within this period with high probability.

A 5-minute seizure horizon was used to make sure that the prediction method is not simply a detection method. The warning is considered valid only if the warning occurs before the 5-minute horizon. If one or more seizures occurred during the warning, then it was considered as a true positive. If

seizures do not occur during a warning period, then that warning was considered as a false positive. If a seizure occurred during a period where there was no warning triggered, then it was considered a false negative. If seizures didn't occur during a period and no warnings were triggered during that period, then it was considered as a true negative. Based on this, the Receiver Operating Characteristics (ROC) curve was also plotted to evaluate the performance of the pipeline.

To evaluate the ability of the seizure forecasting mechanism, it was compared to a chance seizure prediction algorithm previously proposed by Snyder et al. [81]. A candidate seizure prediction algorithm must perform significantly ($p < 0.05$) better than the chance seizure prediction method in order to be considered for a potential seizure predictor.

5.4.1 Receiver Operating Characteristics (ROC) Curve[5]

An ROC curve is typically used to visualize, organize and select classifiers based on their performances. Let us consider a classification problem with two classes. Therefore, each test data point is mapped to one of the labels in set p, n of positive and negative labels. The labels are usually provided based on thresholding. For each data point x, the classifier generally generates a likelihood probability between 0 and 1. A threshold between 0 and 1 is used to assign the data point to a particular. For instance, if the likelihood probability is 0.7 and the threshold is 0.6, then this data point will be assigned to class 1 (positive).

When the data point has a true label and a predicted label, there are four possibilities. If the instance is positive and it is classified as positive, it is counted as a *true positive*; if it is classified as negative, it is counted as a *false negative*. If the instance is negative and it is classified as negative, it is counted as a *true negative*; if it is classified as positive, it is counted as a *false positive*. This is explained in Figure 5.7.

Denoting the total number of true positives as TP, positives as P, false positives as FP and negatives as N, True Positive Rate (TPR) and False

[5]Adapted from [14].

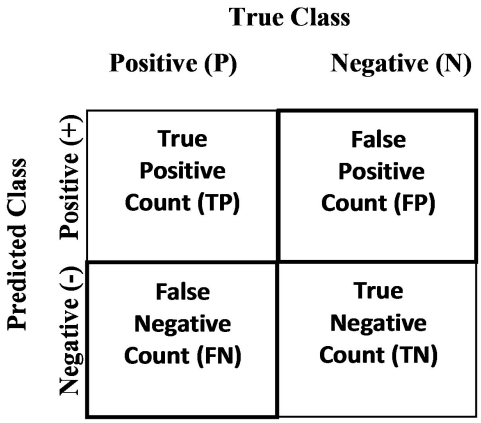

Figure 5.7: Confusion matrix

Positive Rate (FPR) can be calculated as follows

$$TPR = \frac{TP}{P}$$

$$FPR = \frac{FP}{N}$$

TPR and FPR are calculated for a particular threshold. The threshold can be varied between 0 and 1 to generate multiple TPR and FPR values for that particular classifier. These multiple values of TPR and FPR are plotted against each other to generate the ROC curve. Figure 5.8 shows two sample ROC curves.

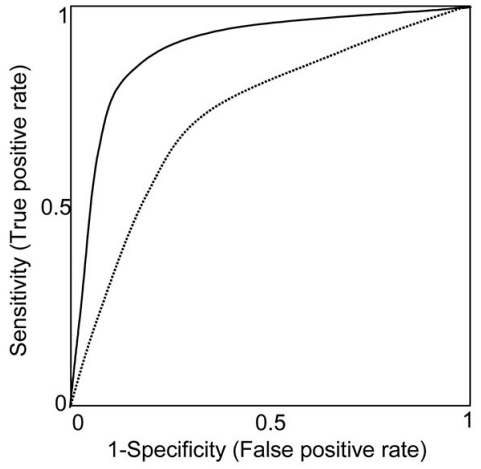

Figure 5.8: Receiver Operating Characteristics (ROC) curve

Area under the ROC curve (AUC) is considered as a measure for the

predictive capability of a classifier. AUC should be close to 1 for the classifier to be considered as a good predictor. When the AUC is close to 0.5, the classifier is as good as a predictor which predicts the label for a data point randomly (50-50 chance). It is practically impossible to achieve a perfect AUC value, i.e. 1. However, AUC values very close to 1 could be achieved and those classifiers are considered as good predictors.

5.5 Execution of the Prediction Pipeline

Data samples representing different classes are required to train the machine learning classifier. So we extracted data samples representing interictal activity and preictal activity. We considered the entire recording of a subject and excluded discontinuous regions as explained in Chapter 4. Annotated seizures were used to mark the interictal and preictal activities. Interictal periods were very conservatively marked as any part of the recording that is at least one week away from a seizure. Preictal periods were marked five to ninety-five minutes before a seizure, accounting for the possibilities of preictal signatures persisting for different durations before the seizure. Preictal and interictal regions were extracted from the raw recordings, and were divided into non-overlapping one-minute clips. These one-minute clips were extracted from the raw data prior to the execution of the pipeline and saved in storage to expedite further processing. Table 5.1 lists the number of interictal and preictal clips extracted for each subject.

Table 5.1: Summary of extracted 1-min clips

Subject	Interictal clips	Preictal clips
Subject 1	166170	328
Subject 2	159808	345
Subject 3	145788	606
Subject 4	286	261
Subject 5	33649	264

Here, we explain how the pipeline would operate with a combination of the techniques employed in each block of the pipeline. The feature extraction block takes the saved one-minute clips as input and extracts the features explained in Section 5.1. The extracted features are fed to the dimensionality

reduction block for reduction of dimensionality. The dimensionality reduction block takes the extracted features and reduces the dimensionality by using one of the two methods described in Section 5.2. The method in which the dimensionality is reduced can be specified as an input. The dimensionality of the output can also be specified as an input to this block. At this point in the pipeline, the feature set of each one-minute clip is assigned a label depending on whether that one-minute clip was in the interictal region or in the preictal region. Then the dataset and the corresponding labels are divided into two parts, training and testing sets. Machine learning block takes a set of training features, corresponding labels and a set of test features as input. It trains one of the classifiers explained in Section 5.3 (which can be specified as an input to the block) and produces the labels and likelihood probabilities for the given test features suing the trained classifier. The assessment block takes the true labels, predicted labels and the likelihood probabilities of the test dataset as input and produces goodness of fit metrics for the combination of techniques used in the pipeline. Chapter 6 further explains the execution of the pipeline using the cross validation method to account for the randomness in different executions.

CHAPTER 6

EVALUATION

A top-level software wrapper was implemented to execute the individual blocks varying the technique employed in each block. A flow diagram of the top-level wrapper is shown in Figure 6.1.

The wrapper selects a test subject and extracts all the features for the interictal and preictal one-minute clips of that test subject. For the purpose of testing the whole iEEG recording of the subject, the entire recording is divided into one-minute clips and the features for each one-minute clip are also extracted. One of PIB, TMCO and SPCO is selected from the extracted features of that subject. The selected features for all the one-minute clips (training and test sets) are provided to the dimensionality reduction block as input. The top-level wrapper then selects the dimensionality reduction method as one of PCA and PLS and selects the dimensionality of the output. Then it performs dimensionality reduction on the input features and provides the features with reduced dimensionality as input to the cross-validation block. The cross-validation block contains both machine learning and forecasting and assessment blocks. The machine learning block takes the dimensionality reduced features as input. The features are first shuffled to distribute the preictal and interictal features evenly. Two-thirds of the features are chosen as train features and one of SVM, RFC and ANN is chosen as the classifier and is trained using the train data.

Box constraint, misclassification costs for SVM, number of neurons, number of layers, misclassification costs for ANN and input fraction used for training, misclassification costs for RFC were chosen by trial and error method, achieving the best classifier performance for each subject.

The classifier which was trained using the training set is called the predictor. Predictor is used to predict the labels and likelihood probabilities of the labels for the testing set. Predicted labels, likelihood probabilities and the actual labels for the testing set are passed to the forecasting and assessment

block as inputs. To test the trained classifier, dimensionality reduced features of the whole dataset are used as the test dataset. The machine learning block then uses the trained classifier (i.e. the predictor) to produce the labels and the corresponding likelihood probabilities for the test dataset. Predicted labels, likelihood probabilities and actual labels for the test dataset are fed into the forecasting and assessment block. Forecasting and assessment block generates the warnings and goodness of fit metrics for the classifier. The process of training the classifier using the training dataset, testing the predictor using testing dataset and generating goodness of fit metrics using the forecasting and assessment block is repeated five times to account for the randomness in the training of machine learning classifier and the bias in the selection of training data. The goodness of fit metrics (Area Under the ROC Curve (AUC) and p-value obtained by comparing the classifier with a random predictor) generated in the five runs of the cross-validation stage are averaged to produce the average performance metrics of the classifier.

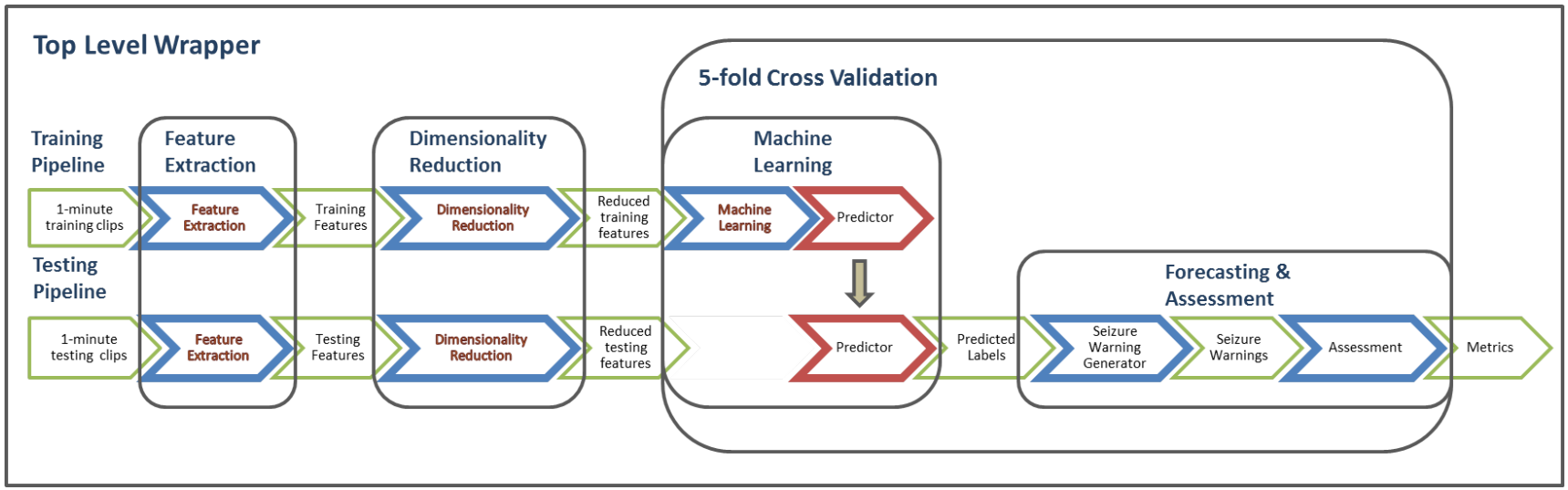

Figure 6.1: Top-level wrapper

CHAPTER 7

RESULTS AND DISCUSSION

Each canine subject was evaluated separately by using all the different combinations of features and techniques. Table 7.1 shows the different features and techniques employed in each block.

Table 7.1: Breakdown of the techniques

Feature Extraction	Dimensionality Reduction	ML Classifier
Power in Bands (PIB)	Principal Component Analysis (PCA)	Support Vector Machine (SVM)
Time Correlation (TMCO)	Partial Least Squares (PLS) regression	Random Forests Classifier (RF)
Spectral Coherence (SPCO)		Artificial Neural Networks (ANN)

Based on this, there are 18 combinations of features and techniques possible for each subject. Each combination of the features and techniques was cross-validated five times to ensure that the results are generalizable. The AUC values of the 18 combinations were compared to obtain the best performing combination and corresponding AUC value for a particular subject. Table 7.2 shows the best performing combinations and the corresponding AUC values for each subject. Figure 7.1 shows the ROC curves that resulted in the best AUC value for each subject.

Table 7.2: Best performing combinations of techniques for each subject and the corresponding AUC values

Subject	Feature	DR	ML	AUC
1	SPCO	PLS	SVM	0.8341
2	PIB	PCA	SVM	0.7795
3	SPCO	PLS	SVM	0.9246
4	PIB	PLS	RF	0.7349
5	PIB	PCA	SVM	0.8899

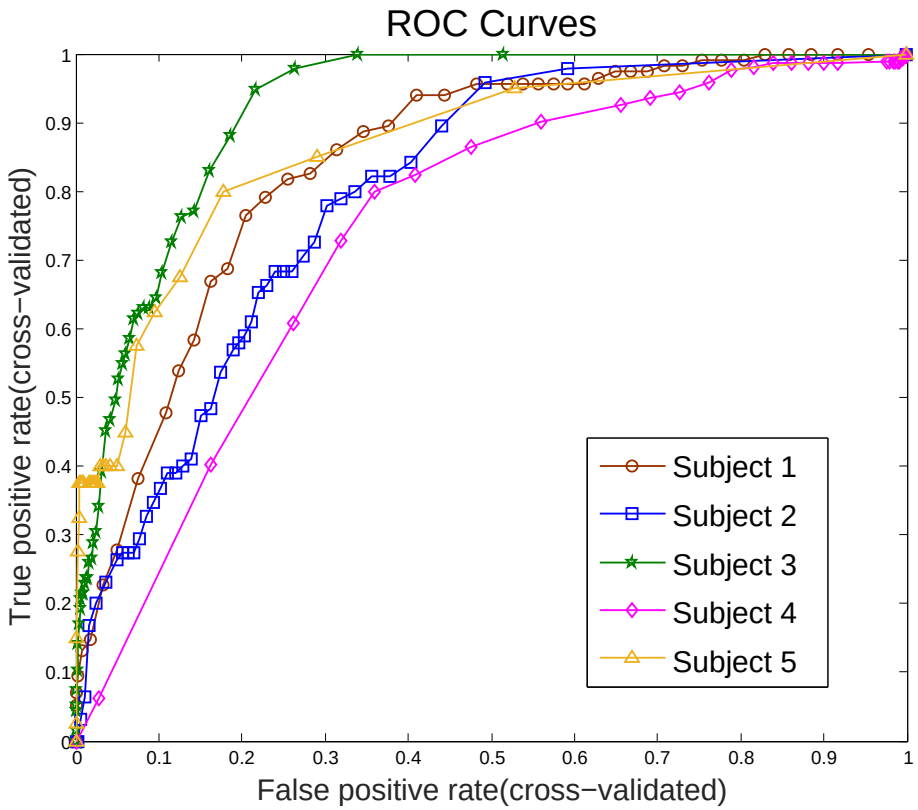

Figure 7.1: ROC curves which resulted in best AUC values

Based on the best AUC achieved for each subject, seizures on Subjects 1, 3 and 5 (AUC > 0.8) could be predicted relatively better than Subjects 2 and 4.

To evaluate the different techniques used in the individual blocks, many trials were performed varying the technique employed in each block. A combination of the different techniques was used for the blocks feature extraction, dimensionality reduction and machine learning in each of the eighteen possible trials. For each trial, the average AUC metrics and p-values generated from the forecasting and assessment block were recorded. Table 7.3 shows the forecasting results using different features and analytical methods. AUC metrics recorded for each combination of techniques were averaged over all the canines to produce the means (μ) and standard errors (SE) listed in the table. The number of canines for which a given combination of features and

Table 7.3: Mean forecasting results for the different combinations of the techniques

Feature	**Dimensionality Reduction**	**Machine Learning**	**AUC - mean (SE)**	**Number of Canines with p-value < 0.05**
PIB	PCA	SVM	0.7943 (0.0373)	4
		RF	0.7740 (0.0396)	5
		ANN	0.7402 (0.0481)	3
	PLS	SVM	0.7520 (0.0595)	4
		RF	0.7552 (0.0339)	4
		ANN	0.6285 (0.0335)	1
TMCO	PCA	SVM	0.6968 (0.0378)	4
		RF	0.7524 (0.0312)	5
		ANN	0.6926 (0.0468)	3
	PLS	SVM	0.7559 (0.0333)	4
		RF	0.7559 (0.0304)	5
		ANN	0.6498 (0.0331)	2
SPCO	PCA	SVM	0.6735 (0.0455)	3
		RF	0.7442 (0.0326)	5
		ANN	0.6921 (0.0457)	3
	PLS	SVM	0.7730 (0.0510)	4
		RF	0.7287 (0.0327)	5
		ANN	0.5954 (0.0184)	1

methods generated a forecasting method with $p < 0.05$ is reported in the final column.

7.1 Forecasting Results

The best performing features in this dataset appear to be the PIB with the PCA dimensionality reduction, as all machine learning methods achieved mean AUC greater than 0.74. While the SVM algorithm with PIB features and PCA had the highest mean AUC, the RF algorithm may have generalized better, achieving forecasting greater than a chance predictor in all five canines studied. The ANN learning algorithm did not perform as well as the SVM and RF algorithms in any trial, suggesting it may not be as accurate or robust as the other methods. Good performance was obtained in some capacity for all feature sets studied, suggesting preictal changes may be present and identifiable in all three feature sets.

We also analyzed the effectiveness of the different features extracted on individual subjects. Out of the eighteen cross-validated trials of a subject, a feature (one of PIB, TMCO and SPCO) was used in six trials. The AUC metrics for these six trials were averaged to obtain the mean prediction performance of that feature for a particular subject. The bar plots in Figure 7.2

show the mean performance of each feature on different subjects. The confidence intervals of the mean AUC were obtained by calculating the standard errors (SE).

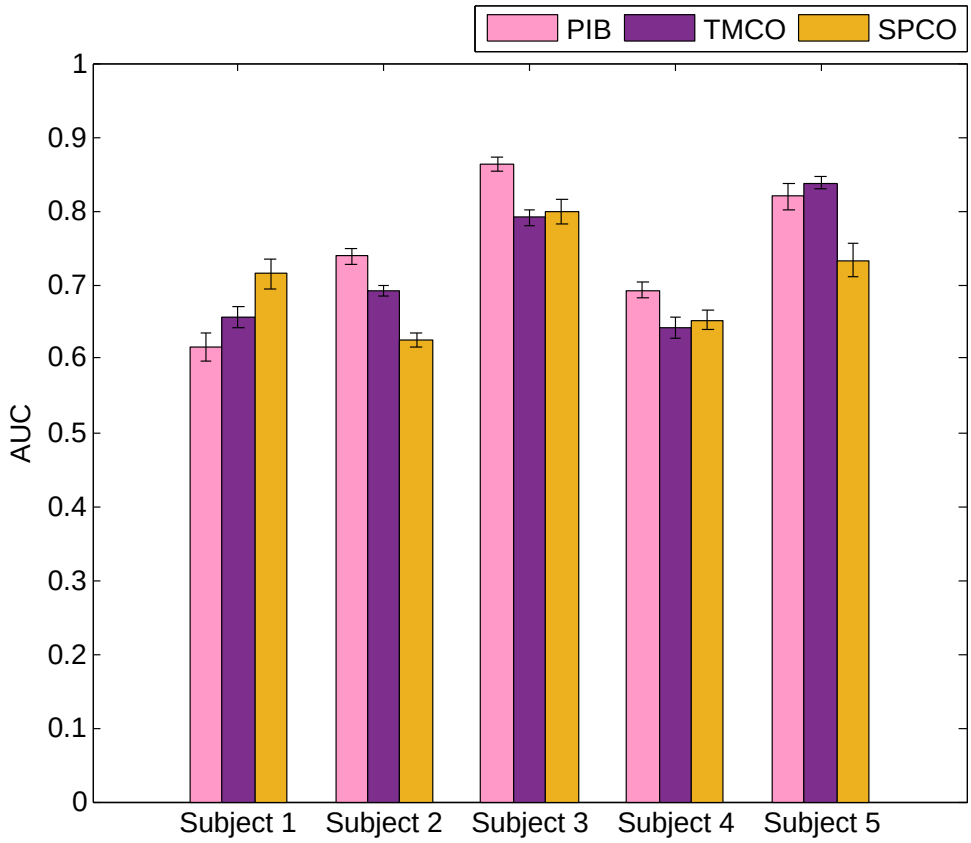

Figure 7.2: Mean prediction performances of the features on individual subjects

It is clearly observable from Figure 7.2 that the features show difference in their prediction capabilities depending on the subject. Table 7.4 lists the prediction capabilities of the features based on the subjects, with 1 being the highest prediction capability and 3 being the lowest. The ordering of features based on the mean prediction performance is done individually for each subject. We also provide a general classification of the prediction performance of a feature by classifying feature's performance into three categories, i.e. AUC > 0.8 being characterized as "Excellent", $0.8 >$ AUC > 0.7 being characterized as "Good" and $0.7 >$ AUC > 0.6 being characterized as "Average". Table 7.4 also lists the general classification of the feature's predictive

Table 7.4: Mean prediction performances of the features

Subject	PIB	TMCO	SPCO
Subject 1	3 (Average)	2 (Average)	1 (Good)
Subject 2	1 (Good)	2 (Average)	3 (Average)
Subject 3	1 (Excellent)	3 (Good)	2 (Good)
Subject 4	1 (Average)	3 (Average)	2 (Average)
Subject 5	2 (Excellent)	1 (Excellent)	3 (Good)

performance based on the mean AUC obtained.

Although the differences between the mean AUCs of different features are small (i.e. in the range of 0 - 0.1), these differences are significant for the following reasons. An AUC value can only be in the range 0 - 1. Further, a prediction mechanism should achieve AUC > 0.5 to be considered better than a chance (50-50) prediction algorithm. Therefore, a standard prediction algorithm will have AUC values between 0.5 and 1. Moreover, the closer the AUC value is to 1, the harder it becomes to further improve the AUC. Therefore, provided that all the mean AUC values are greater than 0.6 and these values can only be within 0.5 and 1, the small differences we observe here are indeed significant. Further, the small changes in the AUC values in fact correspond to larger changes in the true positive and false negative rates. Depending on the size of the testing dataset, these small improvements in AUC can considerably, improve the number of true positives and reduce the number of false positives.

This finding is important in the way that it opens up the possibility of personalizing the seizure prediction algorithm based on the feature that shows most predictive ability for that particular subject. For instance, for Subject 1 (based on row 1 of Table 7.4), the feature SPCO should be used for the seizure prediction algorithm. Although these results being the average results across all the trials with different machine learning and dimensionality reduction techniques is a reinforcement, the randomness involved in the machine learning algorithms requires further evidence for this claim. Since these results suggest that machine learning algorithms can identify subtle changes in iEEG-derived features preceding seizures, it may be possible to analyze and derive specifically what changes are occurring prior to a seizure for each subject, and this information could help us better understand the physiological changes underlying ictogenesis.

7.2 Preictal Changes

Identifying the changes in the iEEG data during the transition from the interictal baseline to a preictal, seizure-permissive state is complicated by the fact that the exact timing of this transition is unknown. Prior seizure forecasting studies have analyzed data between 30 and 90 minutes preceding seizure onset. In this context, the analysis in this work was begun 120 minutes before the seizure to ensure adequate coverage of the preictal period. Similarly, the duration of postictal effects following seizure termination is variable, and subtle effects may persist beyond visible changes in the iEEG. Two hours post seizure was also analyzed, as this extends beyond published postictal studies.

A data segment containing the two hour period prior to the seizure, the seizure and the two hour period after the seizure was extracted for each lead seizure of each subject. These data segments were divided into non-overlapping one minute clips. Each of these clips contains one minute recording of iEEG over sixteen channels. For each one-minute clip, the features PIB, TMCO and SPCO were extracted as explained in Section 3.2. For each one-minute clips 1109 PIB features, 120 TMCO features and 1820 SPCO features, the two dimensionality reduction techniques were applied and 10 PCs of PIB, 10 PCs of TMCO and 10 PCs of SPCO were extracted. Likewise, 10 PCs of each feature was extracted for all the one-minute clips in the data segments extracted around lead seizures. To quantify the 10 PCs as a single numerical value, squared L2 norm of the 10 PCs was calculated. So, each one-minute clip was assigned six numerical measures, which are the L2 norms of the 10 PCs of the three different features extracted using two different dimensionality reduction techniques. The means and variances of the six numerical measures of a one-minute clip were calculated considering all the lead seizure data segments extracted on the specific subject. Figures 7.3-7.7 show the changes that occurred in these features around lead seizures by means of the mean values and variances of those features.

In Figure 7.2 (Subject 1), we can see that except for the time of seizure, there are no noticeable differences in the means and variances of the features PIB ((i), (ii)) and TMCO ((iii), (iv)). However, there is a noticeable difference in the variances of SPCO in (v) and (vi). The variance is small in the preictal region (entire 2 hour period) and large in the postictal region. This

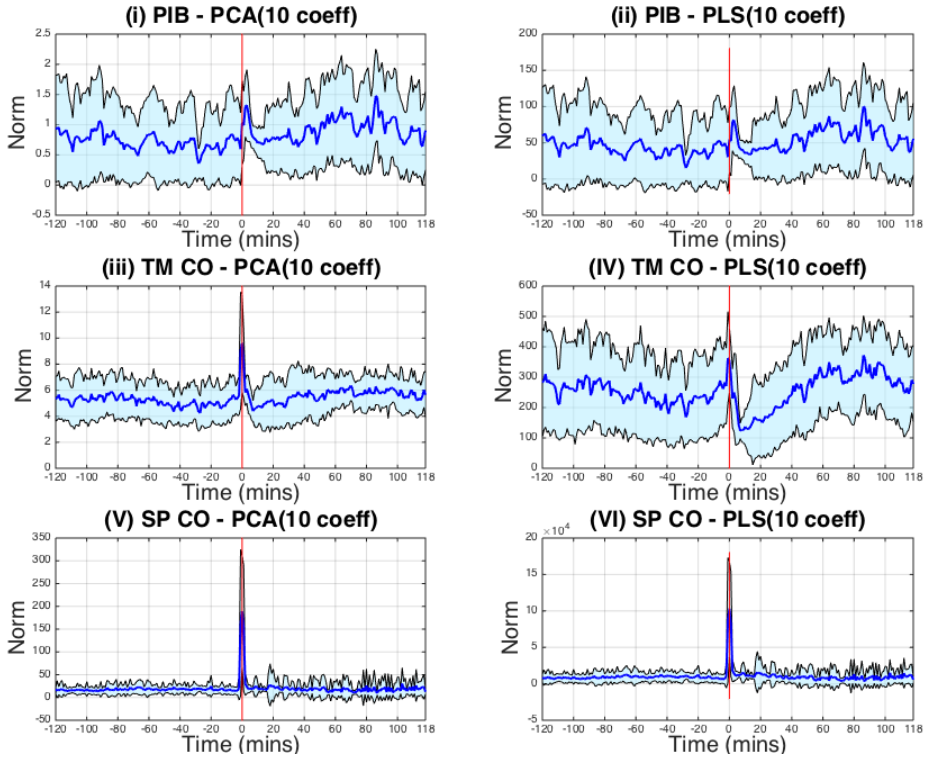

Figure 7.3: Preictal changes oberved in Subject 1

change in the variance of SPCO reinforces the observation from Figure 7.2 (Subject 1), explaining the "Good" performance by SPCO features. However, these plots do not reveal enough to make any conclusions about the difference in the performances of PIB and TMCO features.

Similarly, in Figure 7.2 (Subject 2), we can observe preictal changes in PIB and TMCO. The variances of PIB ((i), (ii)) show a decrease almost 50 (on average) minutes before the seizures and the variances of TMCO ((iii), (iv)) show an increase from 50 minutes (on average) before the seizures. However, there are no noticeable changes observed in SPCO preictally. These observations explain the difference in the performances of the features in Figure 7.2 (Subject 2). In Figure 7.2 (Subject 2), we observe that although only PIB features provided "Good" performance, both PIB and TMCO provided significantly better predictions than SPCO. However, the difference between the performances of PIB and TMCO in Figure 7.2 (Subject 2) cannot be explained using the preictal analysis in Figure 7.4.

In Figure 7.5, we can observe preictal changes in all three features about 10-15 minutes (on average) before the seizures. These are very significant

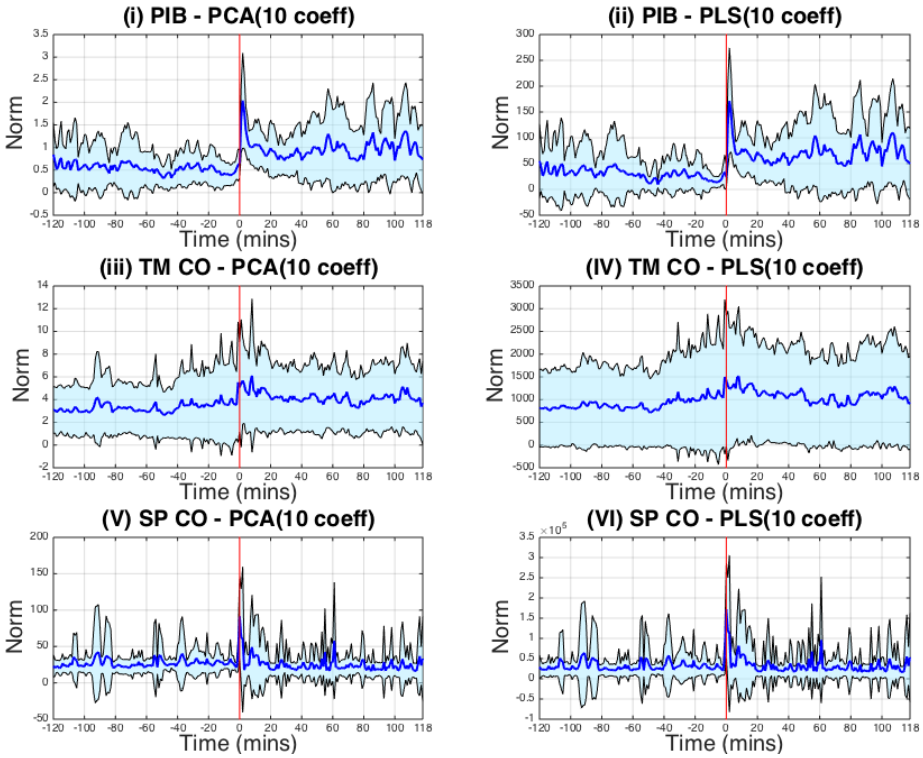

Figure 7.4: Preictal changes observed in Subject 2

changes because unlike other cases, we can see significant changes in the means of all the features 10-15 minutes before the seizures and settling to the respective interictal means about 40 minutes (roughly) after the seizures. This observation explains all three features providing performances better than "Average" in Figure 7.2 (Subject 3). However, the reason for PIB features providing the best performance among all the features is not clear from the preictal analysis in Figure 7.5.

In Figure 7.6, we can observe preictal changes in PIB ((i) and (ii)) and TMCO ((iv)). SPCO features do not show any preictal changes. The changes in PIB features are very significant because the variance becomes very small in (i) and almost zero in (ii). We can also observe a noticeable change in the variance of TMCO features in (iv). The significant changes of PIB features are resembled in Figure 7.6, where PIB features provided comparatively better performance than other features. However, the preictal change observed in TMCO features in (iv) is not resembled in Figure 7.2 (Subject 4).

In Figure 7.7, we can observe changes in TMCO ((iv)) and SPCO ((v) and (vi)). The change in (iv) is very prominent, i.e. the variance is much

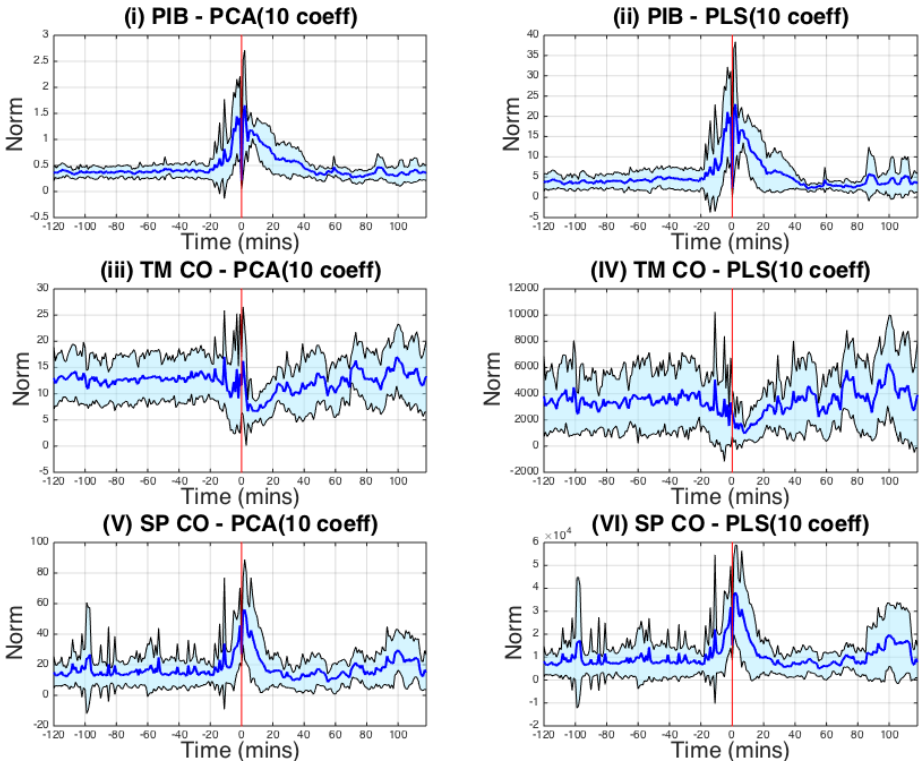

Figure 7.5: Preictal changes observed in Subject 3

larger preictally. Although the changes in (v) and (vi) are important in the way that they show a rise in the mean, they do not occur well before the seizures. This could be a limiting factor for the ability of SPCO features in predicting seizures. The changes in PIB features are very noisy, and it is not possible to derive any conclusions from (i) and (ii). If we look at Figure 7.2 (Subject 5), TMCO features, as expected provided the best performance. While SPCO provided a reasonable prediction performance, PIB features performed better, even though it was not clearly evident from the preictal analysis in Figure 7.7.

Further, the preictal changes observed persisted for different time durations in different subjects. However, the duration of a preictal changes on the different features (which showed preictal changes) of the same subject were very similar. In Figure 7.3 only SPCO features ((v), (vi)) showed preictal changes and the duration was 120 minutes (possibly longer than that). In Figure 7.4, both PIB ((i), (ii)) and TMCO ((iii), (iv)) showed preictal changes roughly around 50 minutes before the seizures. In Figure 7.5, all three features showed preictal changes approximately 10-15 minutes before

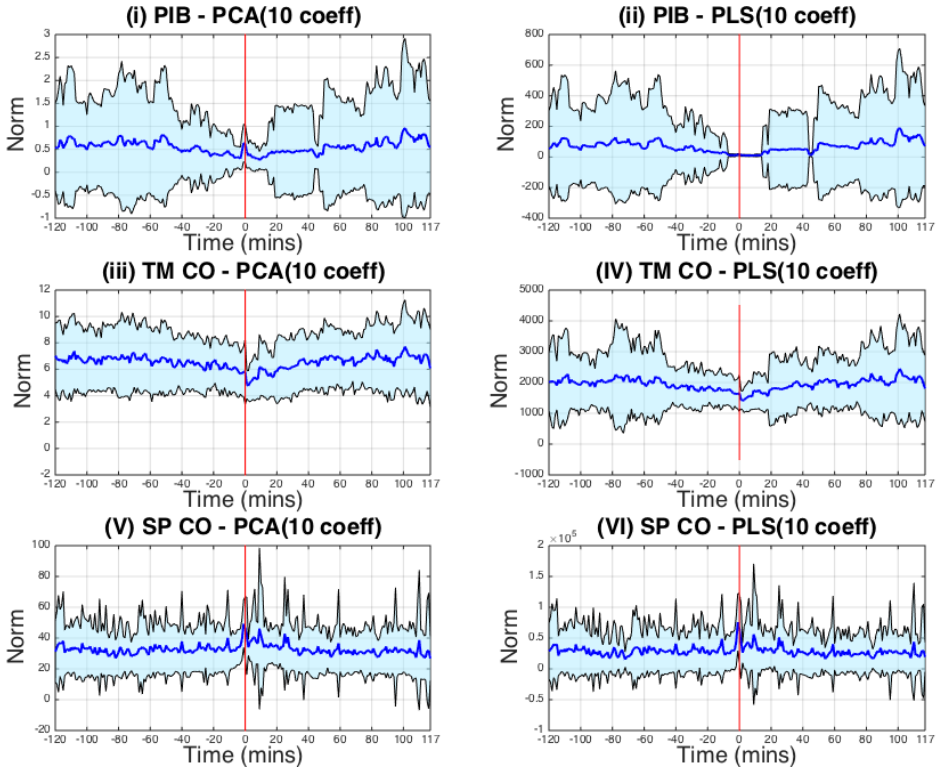

Figure 7.6: Preictal changes observed in Subject 4

the seizures. In Figure 7.6, PIB ((i), (ii)) and TMCO ((iv)) showed preictal changes around 40 minutes before the seizures. In Figure 7.7, only TMCO ((iv)) showed significant preictal changes which persisted for 2 hours (possibly longer) before the seizures. This observation suggests that durations of preictal changes are different for different individuals and preictal changes on the different features (which showed preictal changes) of the same subject are very similar.

These observations suggest that different features show different preictal changes and it is not wise to assume that there exists a common feature which shows similar preictal changes in all the subjects. Due to the dynamics in brain activity and the individual differences between subjects, it is highly unlikely that there exists a common bio-marker for epilepsy. By performing a preictal analysis as shown here, we can get a fair idea about the features which might be worth considering in developing a machine learning based seizure prediction algorithm. Also, it is evident from the analysis that the features which showed preictal changes were able to provide better predictions than

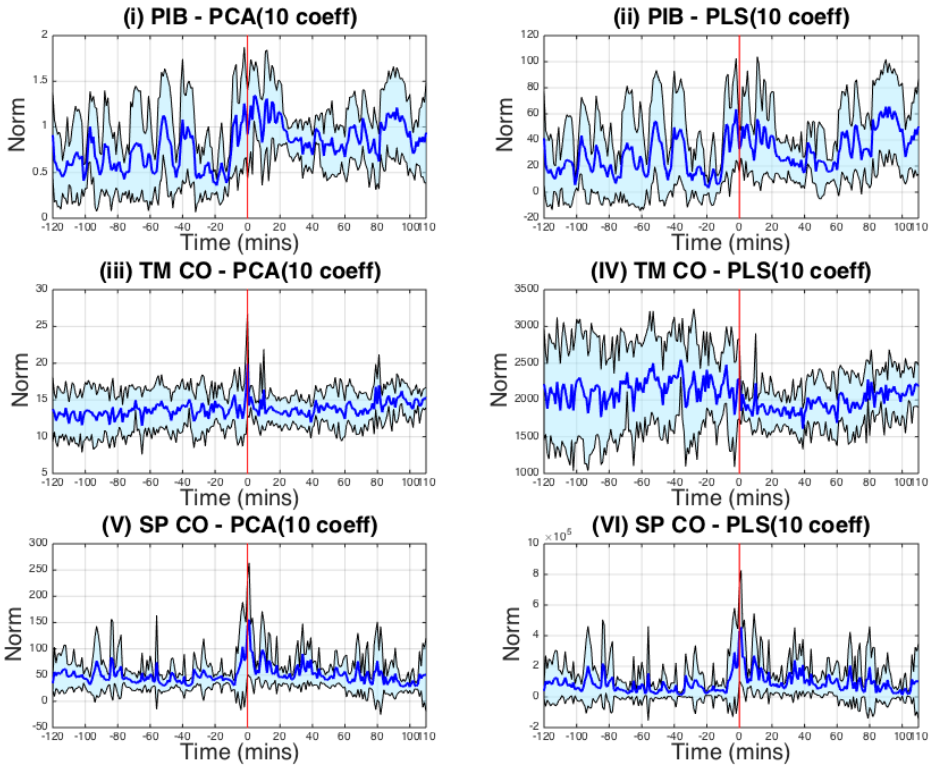

Figure 7.7: Preictal changes observed in Subject 5

other features for a particular subject. Therefore, this analysis could also be useful in individualizing a seizure prediction algorithm.

CHAPTER 8

CONCLUSION

In this thesis, we developed a machine learning based seizure prediction pipeline, which we used to evaluate different combinations of features, dimensionality reduction techniques and machine learning methods and their applicability in seizure prediction. The results demonstrate that all sets of features, dimensionality reduction, and machine learning techniques investigated showed some capability to forecast seizures, but SVM and RF machine learning classifiers performed consistently better than ANN. All feature sets tested produced forecasting results greater than chance in all five canines studied with some combination of dimensionality reduction and machine learning algorithms. These results also demonstrate changes in a number of iEEG features prior to seizures and support the concept of a distinct, measurable, preictal state that has an increased probability of seizure occurrence. Preictal changes in mean and variance PIB, TMCO, and SPCO metrics were observed in multiple canines and occurred as early as 40 minutes before seizures. These results may provide insight into the timing and duration of the underlying physiological changes that lead to seizures.

8.1 Future Directions

The ultimate objective of this thesis is to design a device with implanted electrodes which can predict seizures in a real-time manner. This work stands as a strong foundation to the design of the device as it demonstrates the suitability of a machine learning based approach and as it shows directions for how to choose the features with predictive capability which will provide better predictions. There are further research problems that need to be solved in order to successfully develop such a device. These include the following:

- Decision on how much time should be spent on training the machine

learning based framework.

- Development of a mathematical strategy to optimize the algorithm online while new data is arriving.
- Automation of the device to operate independently of human intervention.
- Understand where human input both clinically and technically, is required.
- Design of a minimally invasive hardware and taking measures to account for the convenience of the user.
- Efficient design of the software so as to minimally consume power and elongate battery life.
- Use of a fault tolerance strategy to deal with occasional failures that might occur in the hardware and software.

Provided that these problems are solved, a device could be realized which can predict seizures (with few false positives and few false negatives) providing warnings to epilepsy patients in a timely manner giving them an ample amount of time to prepare for the seizures.

REFERENCES

[1] S. Shorvon and P. Farmer, "Epilepsy in developing countries: A review of epidemiological, sociocultural, and treatment aspects," *Epilepsia*, vol. 29, no. s1, pp. S36–S54, 1988.

[2] C. J. Murray, A. D. Lopez et al., "Global comparative assessments in the health sector: Disease burden, expenditures and intervention packages," *World Health Organization, Geneva*, 1994.

[3] P. Kwan and M. J. Brodie, "Early identification of refractory epilepsy," *New England Journal of Medicine*, vol. 342, no. 5, pp. 314–319, 2000.

[4] A. Schulze-Bonhage and A. Kühn, "Unpredictability of seizures and the burden of epilepsy," *Seizure Prediction in Epilepsy: From Basic Mechanisms to Clinical Applications*, 2008.

[5] A. Schulze-Bonhage, F. Sales, K. Wagner, R. Teotonio, A. Carius, A. Schelle, and M. Ihle, "Views of patients with epilepsy on seizure prediction devices," *Epilepsy & Behavior*, vol. 18, no. 4, pp. 388–396, 2010.

[6] R. S. Fisher, "Epilepsy from the patient's perspective: Review of results of a community-based survey," *Epilepsy & Behavior*, vol. 1, no. 4, pp. S9–S14, 2000.

[7] R. S. Fisher, "Therapeutic devices for epilepsy," *Annals of Neurology*, vol. 71, no. 2, pp. 157–168, 2012.

[8] R. G. Andrzejak, D. Chicharro, C. E. Elger, and F. Mormann, "Seizure prediction: Any better than chance?" *Clinical Neurophysiology*, vol. 120, no. 8, pp. 1465–1478, 2009.

[9] J. J. Howbert, E. E. Patterson, S. M. Stead, B. Brinkmann, V. Vasoli, D. Crepeau, C. H. Vite, B. Sturges, V. Ruedebusch, J. Mavoori et al., "Forecasting seizures in dogs with naturally occurring epilepsy," *PloS One*, vol. 9, no. 1, p. e81920, 2014.

[10] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," *Clinical Neurophysiology*, vol. 116, no. 3, pp. 569–587, 2005.

[11] Y. Park, L. Luo, K. K. Parhi, and T. Netoff, "Seizure prediction with spectral power of EEG using cost-sensitive support vector machines," *Epilepsia*, vol. 52, no. 10, pp. 1761–1770, 2011.

[12] Wagenaar, Joost B and Brinkmann, Benjamin H and Ives, Zachary and Worrell, Gregory A and Litt, Brian, "iEEG Portal," https://www.ieeg.org/, 2013.

[13] J. B. Wagenaar, B. H. Brinkmann, Z. Ives, G. A. Worrell, and B. Litt, "A multimodal platform for cloud-based collaborative research," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*. IEEE, 2013, pp. 1386–1389.

[14] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[15] I. Leppik, *Contemporary Diagnosis and Management of the Patient with Epilepsy*. Handbooks in Health Care, 1997. [Online]. Available: https://books.google.com/books?id=WMxrAAAAMAAJ

[16] J. S. Ebersole and T. A. Pedley, *Current Practice of Clinical Electroencephalography*. Lippincott Williams & Wilkins, 2003.

[17] F. S. Tyner, J. R. Knott, and W. B. Mayer, *Fundamentals of EEG technology: Clinical Correlates*. Lippincott Williams & Wilkins, 1989, vol. 2.

[18] S. Viglione and G. Walsh, "Proceedings: Epileptic seizure prediction," *Electroencephalography and Clinical Neurophysiology*, vol. 39, no. 4, pp. 435–436, 1975.

[19] Z. Rogowski, I. Gath, and E. Bental, "On the prediction of epileptic seizures," *Biological Cybernetics*, vol. 42, no. 1, pp. 9–15, 1981.

[20] Y. Salant, I. Gath, and O. Henriksen, "Prediction of epileptic seizures from two-channel EEG," *Medical and Biological Engineering and Computing*, vol. 36, no. 5, pp. 549–556, 1998.

[21] A. Siegel, C. L. Grady, and A. F. Mirsky, "Prediction of spike-wave bursts in absence epilepsy by EEG power-spectrum signals," *Epilepsia*, vol. 23, no. 1, pp. 47–60, 1982.

[22] H. H. Lange, J. P. Lieb, J. Engel, and P. H. Crandall, "Temporo-spatial patterns of pre-ictal spike activity in human temporal lobe epilepsy," *Electroencephalography and Clinical Neurophysiology*, vol. 56, no. 6, pp. 543–555, 1983.

[23] J. Gotman and M. Marciani, "Electroencephalographic spiking activity, drug levels, and seizure occurence in epileptic patients," *Annals of Neurology*, vol. 17, no. 6, pp. 597–603, 1985.

[24] J. Gotman and D. Koffler, "Interictal spiking increases after seizures but does not after decrease in medication," *Electroencephalography and Clinical Neurophysiology*, vol. 72, no. 1, pp. 7–15, 1989.

[25] A. Katz, D. A. Marks, G. McCarthy, and S. S. Spencer, "Does interictal spiking change prior to seizures?" *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 2, pp. 153–156, 1991.

[26] L. D. Iasemidis, "Epileptic seizure prediction and control," *Biomedical Engineering, IEEE Transactions on*, vol. 50, no. 5, pp. 549–558, 2003.

[27] J. Martinerie, C. Adam, M. Le Van Quyen, M. Baulac, S. Clemenceau, B. Renault, and F. J. Varela, "Epileptic seizures can be anticipated by non-linear analysis," *Nature Medicine*, vol. 4, no. 10, pp. 1173–1176, 1998.

[28] M. Le Van Quyen, J. Martinerie, M. Baulac, and F. Varela, "Anticipating epileptic seizures in real time by a non-linear analysis of similarity between EEG recordings," *Neuroreport*, vol. 10, no. 10, pp. 2149–2155, 1999.

[29] M. Le Van Quyen, C. Adam, J. Martinerie, M. Baulac, S. Clemenceau, and F. Varela, "Spatio-temporal characterizations of non-linear changes in intracranial activities prior to human temporal lobe seizures," *European Journal of Neuroscience*, vol. 12, no. 6, pp. 2124–2134, 2000.

[30] M. Le Van Quyen, J. Martinerie, V. Navarro, P. Boon, M. D'Havé, C. Adam, B. Renault, F. Varela, and M. Baulac, "Anticipation of epileptic seizures from standard EEG recordings," *The Lancet*, vol. 357, no. 9251, pp. 183–188, 2001.

[31] V. Navarro, J. Martinerie, M. Le Van Quyen, S. Clemenceau, C. Adam, M. Baulac, and F. Varela, "Seizure anticipation in human neocortical partial epilepsy," *Brain*, vol. 125, no. 3, pp. 640–655, 2002.

[32] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients," *Physica D: Nonlinear Phenomena*, vol. 144, no. 3, pp. 358–369, 2000.

[33] M. Le Van Quyen, J. Martinerie, V. Navarro, M. Baulac, and F. J. Varela, "Characterizing neurodynamic changes before seizures," *Journal of Clinical Neurophysiology*, vol. 18, no. 3, pp. 191–208, 2001.

[34] M. L. V. Quyen, V. Navarro, J. Martinerie, M. Baulac, and F. J. Varela, "Toward a neurodynamical understanding of ictogenesis," *Epilepsia*, vol. 44, no. s12, pp. 30–43, 2003.

[35] M. Chávez, M. L. Van Quyen, V. Navarro, M. Baulac, and J. Martinerie, "Spatio-temporal dynamics prior to neocortical seizures: Amplitude versus phase couplings," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 5, pp. 571–583, 2003.

[36] K. Lehnertz and C. E. Elger, "Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain electrical activity," *Physical Review Letters*, vol. 80, no. 22, p. 5019, 1998.

[37] C. E. Elger and K. Lehnertz, "Seizure prediction by non-linear time series analysis of brain electrical activity," *European Journal of Neuroscience*, vol. 10, no. 2, pp. 786–789, 1998.

[38] L. D. Iasemidis, P. Pardalos, J. C. Sackellares, and D.-S. Shiau, "Quadratic binary programming and dynamical system approach to determine the predictability of epileptic seizures," *Journal of Combinatorial Optimization*, vol. 5, no. 1, pp. 9–26, 2001.

[39] B. Litt, R. Esteller, J. Echauz, M. D'Alessandro, R. Shor, T. Henry, P. Pennell, C. Epstein, R. Bakay, M. Dichter et al., "Epileptic seizures may begin hours in advance of clinical onset: A report of five patients," *Neuron*, vol. 30, no. 1, pp. 51–64, 2001.

[40] S. Gigola, F. Ortiz, C. Dattellis, W. Silva, and S. Kochen, "Prediction of epileptic seizures using accumulated energy in a multiresolution framework," *Journal of Neuroscience Methods*, vol. 138, no. 1, pp. 107–111, 2004.

[41] K. Schindler, H. Leung, C. E. Elger, and K. Lehnertz, "Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG," *Brain*, vol. 130, no. 1, pp. 65–77, 2007.

[42] F. Mormann, R. G. Andrzejak, T. Kreuz, C. Rieke, P. David, C. E. Elger, and K. Lehnertz, "Automated detection of a preseizure state based on a decrease in synchronization in intracranial electroencephalogram recordings from epilepsy patients," *Physical Review E*, vol. 67, no. 2, p. 021912, 2003.

[43] M. De Vos, A. Vergult, L. De Lathauwer, W. De Clercq, S. Van Huffel, P. Dupont, A. Palmini, and W. Van Paesschen, "Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone," *NeuroImage*, vol. 37, no. 3, pp. 844–854, 2007.

[44] M. Winterhalder, T. Maiwald, H. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, "The seizure prediction characteristic: A general framework to assess and compare seizure prediction methods," *Epilepsy & Behavior*, vol. 4, no. 3, pp. 318–325, 2003.

[45] R. Aschenbrenner-Scheibe, T. Maiwald, M. Winterhalder, H. Voss, J. Timmer, and A. Schulze-Bonhage, "How well can epileptic seizures be predicted? An evaluation of a nonlinear method," *Brain*, vol. 126, no. 12, pp. 2616–2626, 2003.

[46] M. A. F. Harrison, I. Osorio, M. G. Frei, S. Asuri, and Y.-C. Lai, "Correlation dimension and integral do not predict epileptic seizures," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 15, no. 3, p. 033106, 2005.

[47] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, "Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic," *Physica D: Nonlinear Phenomena*, vol. 194, no. 3, pp. 357–368, 2004.

[48] M. A. F. Harrison, M. G. Frei, and I. Osorio, "Accumulated energy revisited," *Clinical Neurophysiology*, vol. 116, no. 3, pp. 527–531, 2005.

[49] Y.-C. Lai, M. A. F. Harrison, M. G. Frei, and I. Osorio, "Controlled test for predictive power of Lyapunov exponents: Their inability to predict epileptic seizures," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 14, no. 3, pp. 630–642, 2004.

[50] Y.-C. Lai, M. A. F. Harrison, M. G. Frei, and I. Osorio, "Inability of Lyapunov exponents to predict epileptic seizures," *Physical Review Letters*, vol. 91, no. 6, p. 068102, 2003.

[51] L. D. Iasemidis, J. C. Sackellares, H. P. Zaveri, and W. J. Williams, "Phase space topography and the Lyapunov exponent of electrocorticograms in partial seizures," *Brain topography*, vol. 2, no. 3, pp. 187–201, 1990.

[52] P. E. McSharry, L. A. Smith, and L. Tarassenko, "Prediction of epileptic seizures: Are nonlinear methods relevant?" *Nature Medicine*, vol. 9, no. 3, pp. 241–242, 2003.

[53] K. Gadhoumi, J.-M. Lina, F. Mormann, and J. Gotman, "Seizure prediction for therapeutic devices: A review," *Journal of Neuroscience Methods*, 2015.

[54] C. A. Teixeira, B. Direito, M. Bandarabadi, M. Le Van Quyen, M. Valderrama, B. Schelter, A. Schulze-Bonhage, V. Navarro, F. Sales, and A. Dourado, "Epileptic seizure predictors based on computational intelligence techniques: A comparative study with 278 patients," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 3, pp. 324–336, 2014.

[55] N. Moghim and D. W. Corne, "Predicting epileptic seizures in advance," *PLoS One*, vol. 10, no. 3, Mar 2015.

[56] A. Eftekhar, W. Juffali, J. El-Imad, T. G. Constandinou, and C. Toumazou, "Ngram-derived pattern recognition for the detection and prediction of epileptic seizures," *PLoS One*, vol. 9, no. 6, 2014.

[57] A. Aarabi and B. He, "Seizure prediction in hippocampal and neocortical epilepsy using a model-based approach," *Clinical Neurophysiology*, vol. 125, no. 5, pp. 930–940, 2014.

[58] Y. Zheng, G. Wang, K. Li, G. Bao, and J. Wang, "Epileptic seizure prediction using phase synchronization based on bivariate empirical mode decomposition," *Clinical Neurophysiology*, vol. 125, no. 6, pp. 1104–1111, 2014.

[59] S. Li, W. Zhou, Q. Yuan, and Y. Liu, "Seizure prediction using spike rate of intracranial EEG," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, no. 6, pp. 880–886, 2013.

[60] K. Gadhoumi, J.-M. Lina, and J. Gotman, "Seizure prediction in patients with mesial temporal lobe epilepsy using EEG measures of state similarity," *Clinical Neurophysiology*, vol. 124, no. 9, pp. 1745–1754, 2013.

[61] M. J. Cook, T. J. O'Brien, S. F. Berkovic, M. Murphy, A. Morokoff, G. Fabinyi, W. D'Souza, R. Yerra, J. Archer, L. Litewka et al., "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: A first-in-man study," *The Lancet Neurology*, vol. 12, no. 6, pp. 563–571, 2013.

[62] J. R. Williamson, D. W. Bliss, D. W. Browne, and J. T. Narayanan, "Seizure prediction using EEG spatiotemporal correlation structure," *Epilepsy & Behavior*, vol. 25, no. 2, pp. 230–238, 2012.

[63] A. Aarabi and B. He, "A rule-based seizure prediction method for focal neocortical epilepsy," *Clinical Neurophysiology*, vol. 123, no. 6, pp. 1111–1122, 2012.

[64] L. Kuhlmann, D. Freestone, A. Lai, A. N. Burkitt, K. Fuller, D. B. Grayden, L. Seiderer, S. Vogrin, I. M. Mareels, and M. J. Cook, "Patient-specific bivariate-synchrony-based seizure prediction for short prediction horizons," *Epilepsy Research*, vol. 91, no. 2, pp. 214–231, 2010.

[65] P. Mirowski, D. Madhavan, Y. LeCun, and R. Kuzniecky, "Classification of patterns of EEG synchronization for seizure prediction," *Clinical Neurophysiology*, vol. 120, no. 11, pp. 1927–1940, 2009.

[66] J. C. Sackellares, D.-S. Shiau, J. C. Principe, M. C. Yang, L. K. Dance, W. Suharitdamrong, W. Chaovalitwongse, P. M. Pardalos, and L. D. Iasemidis, "Predictability analysis for an automated seizure prediction algorithm," *Journal of Clinical Neurophysiology*, vol. 23, no. 6, pp. 509–520, 2006.

[67] K. A. Davis, B. K. Sturges, C. H. Vite, V. Ruedebusch, G. Worrell, A. B. Gardner, K. Leyde, W. D. Sheffield, and B. Litt, "A novel implanted device to wirelessly record and analyze continuous intracranial canine EEG," *Epilepsy Research*, vol. 96, no. 1, pp. 116–122, 2011.

[68] L. D. Coles, E. E. Patterson, W. D. Sheffield, J. Mavoori, J. Higgins, B. Michael, K. Leyde, J. C. Cloyd, B. Litt, C. Vite et al., "Feasibility study of a caregiver seizure alert system in canine epilepsy," *Epilepsy Research*, vol. 106, no. 3, pp. 456–460, 2013.

[69] A. B. Gardner, A. M. Krieger, G. Vachtsevanos, and B. Litt, "One-class novelty detection for seizure analysis from intracranial EEG," *The Journal of Machine Learning Research*, vol. 7, pp. 1025–1044, 2006.

[70] G. A. Worrell, L. Parish, S. D. Cranstoun, R. Jonas, G. Baltuch, and B. Litt, "High-frequency oscillations and seizure generation in neocortical epilepsy," *Brain*, vol. 127, no. 7, pp. 1496–1506, 2004.

[71] T. C. Ferree and R. C. Hwa, "Power-law scaling in human eeg: relation to fourier power spectrum," *Neurocomputing*, vol. 52, pp. 755–761, 2003.

[72] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987.

[73] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[74] H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[75] Kaggle, "American epilepsy society seizure prediction challenge," http://www.kaggle.com/c/seizure-prediction, Aug 2014.

[76] C. Rodriguez, "Machine Learning, SUNY at Stony Brook," http://omega.albany.edu:8008/ml/, Aug 2008.

[77] N. de Freitas, "Machine Learning, University of British Columbia," http://www.cs.ubc.ca/ nando/540-2013/, Aug 2013.

[78] C. Gershenson, "Artificial neural networks for beginners," *arXiv preprint cs/0308031*, 2003.

[79] A. Ng, "Machine Learning, Stanford University," http://cs229.stanford.edu/info.html, Aug 2015.

[80] M. T. Hagan, H. B. Demuth, M. H. Beale et al., *Neural Network Design*. Pws Pub. Boston, 1996.

[81] D. E. Snyder, J. Echauz, D. B. Grimes, and B. Litt, "The statistics of a practical seizure warning system," *Journal of Neural Engineering*, vol. 5, no. 4, p. 392, 2008.