Exploring Machine Learning Techniques in Epileptic Seizure Detection and Prediction

Negin Moghim

Submitted for the degree of Doctor of Philosophy

Heriot-Watt University

School of Mathematical and Computer Sciences

May 2014

Abstract

Epilepsy is the most common neurological disorder, affecting between 0.6% and 0.8% of the global population. Among those affected by epilepsy whose primary method of seizure management is Anti Epileptic Drug therapy (AED), 30% go on to develop resistance to drugs which ultimately leads to poor seizure management. Currently, alternative therapeutic methods with successful outcome and wide applicability to various types of epilepsy are limited. During an epileptic seizure, the onset of which tends to be sudden and without prior warning, sufferers are highly vulnerable to injury, and methods that might accurately predict seizure episodes in advance are clearly of value, particularly to those who are resistant to other forms of therapy.

In this thesis, we draw from the body of work behind automatic seizure prediction obtained from digitised Electroencephalography (EEG) data and use a selection of machine learning and data mining algorithms and techniques in an attempt to explore potential directions of improvement for automatic prediction of epileptic seizures. We start by adopting a set of EEG features from previous work in the field (Costa et al. 2008) and exploring these via seizure classification and feature selection studies on a large dataset. Guided by the results of these feature selection studies, we then build on Costa et al's work by presenting an expanded feature-set for EEG studies in this area.

Next, we study the predictability of epileptic seizures several minutes (up to 25 minutes) in advance of the physiological onset. Furthermore, we look at the role of the various feature compositions on predicting epileptic seizures well in advance of their occurring. We focus on how predictability varies as a function of how far in advance we are trying to predict the seizure episode and whether the predictive patterns are translated across the entire dataset.

Finally, we study epileptic seizure detection from a multiple-patient perspective. This entails conducting a comprehensive analysis of machine learning models trained on multiple patients and then observing how generalisation is affected by the number of patients and the underlying learning algorithm. Moreover, we improve multiple-patient performance by applying two state of the art machine learning algorithms.

Dedication

To my father, Farhad, whose life-long struggle with epilepsy was the true inspiration behind this work.

Acknowledgments

First and foremost, I wish to thank my family and Pendar, for their endless support. Without their love and encouragement, I would not have the strength to see this work through.

I would also like to thank my supervisor Prof. David W. Corne for his excellent support and advice in the past few years. I am grateful to him for giving me the freedom and encouragement to follow my personal research interest.

I would like to thank Dr. Susan Duncan, who provided me with much needed insight into the medical practice of epilepsy management. She gave me excellent feedback on the applicability of this research and pointed out other epilepsy related issues which could benefit from machine learning and data mining techniques.

I also thank SICSA for financially supporting me for the duration of this PhD.

I am very grateful to EPILAB for sharing their code with me and I thank Freiburg EEG Database, for making their data available to the public.

I would also like to thank Dr. Albert Burger, Prof. Richard Baldock and the staff and students at the Human Genetics Unit of the MRC, who provided me with help and support at the start of my PhD.

I am very grateful to my examiners, Prof. Mike C. Chantler and Dr. Tony Bagnall for their helpful comments and feedback.

And last but not least, I wish to thank the friendly and helpful staff of MACS, who provided me with the necessary resources to undertake this study.

ACADEMIC REGISTRY
Research Thesis Submission

Name*:*	NEGIN MOGHIM		
School/PGI:	MATHEMATICL AND COMPUTER SCIENCES		
Version: *(i.e. First, Resubmission, Final)*	FINAL	Degree Sought (Award **and** Subject area)	PHD IN COMPUTER SCIENCE

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate*:*		Date:	14/05/2014

Submission

Submitted By *(name in capitals):*	NEGIN MOGHIM
Signature of Individual Submitting:	
Date Submitted:	14/05/2014

For Completion in the Student Service Centre (SSC)

Received in the SSC by *(name in capitals):*			
Method of Submission *(Handed in to SSC; posted through internal/external mail):*			
E-thesis Submitted (**mandatory for final theses**)			
Signature:		Date:	

Table Of Contents

Chapter 1
Introduction

This thesis explores the problem of the automatic detection and prediction of epileptic seizures from invasive Electroencephalography (EEG) data. It discusses the body of work behind this field of research and presents the state of the art. It also presents a data-driven road map for improving various aspects of this domain. In this chapter, the central contributions of this thesis are briefly presented. This chapter also motivates the principal theme of the thesis, as well as the work conducted throughout this study toward epileptic seizure prediction from invasive EEG records, as a potential direction for improved seizure management.

1.1 Motivation

In this section the motivation behind epileptic seizure prediction and the methods used in this thesis, is presented.

1.1.1 Why Epilepsy

Epilepsy is a neurological disorder, which affects 50 million people worldwide. The disorder can be managed in some patients using prescription drugs; The remaining 20-30% however, are likely to have a relapse after the initial remission, some of whom may develop drug resistant epilepsy (Mormann et al. 2007). Patients with uncontrolled epilepsy can be affected by accidents caused by unforeseen seizures and are at risk of Sudden Unexpected Death in Epilepsy (SUDEP), as well as a multitude of other unwanted side effects such as memory loss, depression and other psychological disorders (Reynolds et al. 1983).

Despite the design of new anti-epileptic drugs by the pharmaceutical industry, drug resistant epilepsy still lacks an ultimate solution (Mormann et al. 2007). Resective surgery, where the part of the brain that causes the seizures is removed (Elger & Schmidt 2008), can only be applied to a small fraction of drug-resistant patients, the outcome of which is highly unpredictable. Additionally, the cause of drug-resistance is unknown. According to (French 2007) a patient who has been prescribed more than one

type of Anti Epileptic Drug (AED) prior to seizure management is highly likely to develop drug-resistance epilepsy in the future. Resistance to AEDs in addition to the lack of effective seizure management treatments for this large population of patients demands for newer, more effective ways of seizure control therapies.

1.1.2 Why Seizure Prediction

With the wide use of digital EEG recording tools, this kind of data is becoming evermore accessible for electronic manipulation. While EEGs were used as a diagnosis and treatment specification tool for patients, access to the digitised form of this information has granted new fields of research from neonatal seizure detection to understanding how the seizure unfolds in the epileptic brain. EEGs have also widely been used in Brain Computer Interaction where a large number of sophisticated EEG handling techniques have been applied, creating an extensive non-seizure related body of work using EEG signal as the primary data type (Blankertz et al. 2004; Fazli et al. 2009; Fabiani et al. 2004).

The fundamental question researchers have addressed in the field of seizure prediction is to find characteristic features of EEG drawing from signal processing principles, which successfully correlate with the time of the seizure occurrence. This has given rise to a rich body of work, mainly single feature analysis studies, in recognizing and introducing features that can best capture the occurrence of seizures (Mormann et al. 2005).

Another main focus in the field of seizure prediction has been the tracking of seizure activity in the EEG signals leading up to the seizure onset in order to develop a better understanding of how seizures occur as well as studying the possibility of predicting seizures before their physiological onset (Le Van Quyen et al. 2001; Martinerie et al. 1998; Chávez et al. 2003). The various length of the prediction window has given rise to two distinctive prediction problems namely seizure detection and seizure prediction; seizure detection entails the prediction of seizure occurrence a few seconds or minutes prior to the seizure onset while seizure prediction entails the prediction of seizures several minutes/hours before the actual onset. Since there are no standard definitions of seizure detection in the literature, we have re-defined detection and prediction in the scope of this thesis: prediction is defined as the correct classification of a determined pre-ictal (pre-seizure) window which is longer than 30

seconds; detection refers to the correct classification of ictal (seizure) data or pre-ictal data when the length of the pre-ictal window is less than 30 seconds.

The prospect of developing new therapeutic strategies drawing from the advancements in epileptic seizure prediction from EEG records is immensely promising. The automatic prediction of seizure onsets could be used to warn the patient of the occurrence of seizures and therefore allowing them to take action to prevent risks associated with their seizures. Therapeutic plans could also move from long term strategies to fast acting and on demand strategies where the seizure could be prevented prior to full physiological manifestation (Theodore & Fisher 2004; Stein et al. 2000).

1.1.3 EEG Datasets

The public domain hosts several sources of digitised EEG data, but not all of these datasets are deemed useful for a seizure prediction study. Out of the publicly available datasets, a large proportion (~83%) were gathered from non-epileptic subjects, which are of no apparent use to this study. Those gathered from epileptic subjects fall in to two main categories: surface and invasive; surface EEG is placed on the scalp of a patient and captures brain signals from several pre-determined locations of the brain. This type of EEG is non-invasive, poses minimum risk to the subject and is mainly used for the diagnosis and treatment of epilepsy. Invasive EEG on the other hand has a higher signal to noise ratio as is mainly placed on a few focal points on the brain. Due to the invasive nature of this type of EEG, it is mainly used after several scalp EEGs are recorded from the subject, in order to find the exact foci of seizures (more on this in chapter 3). Due to the higher signal to noise ratio of the invasive EEG, this form of recording provides a far more accurate representation of the epileptic brain for the seizure prediction studies. In addition, the application of seizure prediction tools is envisaged to be of closed-loop deep brain form, for which the invasive EEG is a more suitable dataset. For this reason, out of the several online EEG datasets, we only consider the invasive EEG dataset of epileptic subjects.

1.1.4 Machine Learning as a method

Conventionally, statistical methods were used in epileptic seizure prediction studies (Mormann et al. 2007). The statistical methods provide a retrospective analysis where extracted features of EEG are compared for seizure and non-seizure states of the brain.

This method while widely used in this field of research, provides little applicable seizure detection strategies and is mainly used for the analysis of signal characteristics with respect to the seizure and non-seizure states of the brain. The machine learning approach, which is relatively new in this field of research, provides the ability to detect seizure state at any given timepoint and supports a more applicable model of seizure prediction.

Machine learning algorithms and techniques have been used in several fields of healthcare; originally used for diagnosis purposes, have now moved towards utilisation in disease prognosis and prediction (Kononenko 2001). Numerous studies have been carried out in order to improve our understanding of disease using machine learning, but the majority of these studies have not produced suitable outcome for real life application by clinicians.

The machine learning algorithms used in the field of seizure prediction comprise mainly of variations of Artificial Neural Networks (ANN). Some of the more recent algorithm have reportedly been tested solely on raw EEG data and do not take full use of the rich body of work behind feature extraction from EEG data. Some have only been tested on a small dataset comprising a single patient, while some others have not been correctly validated with respect to both Sensitivity and Specificity of detection. A small number of machine learning algorthims have been applied to multi-patient or advance prediction of seizures, and those reported have produced unsatisfatory results.

With the advancements in machine learning both in terms of more sophisticated algorithms as well as featuure selection techniques and fine-tuning of learning algorithms, and with correct statistical validation, better and more improved techniques of seizure prediction could evolve, further moving towards developing algorithms with real-life applicability.

1.1.5 Validation of Seizure Prediction Studies

The results produced for seizure prediction algorithms, should be highly accurate in order to be applicable in live systems. Several studies have reported high levels of Accuracy (low levels of error) on data. Although in some earlier studies the results were not validated in terms of Specificity or false positives (Le Van Quyen et al. 1999; Martinerie et al. 1998; Le Van Quyen et al. 2001), and some optimistic findings, which were obtained from selected small datasets, could not be reproduced for larger and more

diverse EEG files (M. Harrison et al. 2005a; Lai et al. 2004; De Clercq, Lemmerling, Van Huffel & Van Paesschen 2003a). In order for the reported outcome of a study to be valid, it should hold for a larger set of test data.

In health care systems including seizure prediction studies, in addition to high measures of Accuracy, Sensitivity and Specificity must also be high. Sensitivity and Specificity, define respectively, how many of the seizure states did the learner correctly classify as seizures and how many of the states classified as seizures were truly seizures. For real life applications, high Accuracy with low Sensitivity or Specificity is not desirable. The trade-off between Sensitivity and Specificity will be considered at the time of application depending on which is favored over the other, although in lab experiments, efforts should be made to maximise all three measures when possible. A large proportion of seizure prediction studies report results in terms of Accuracy, which is commonly high, and disregard the two measures of Sensitivity and Specificity. In order for a seizure prediction method to be truly validated, it must produce high values of Sensitivity and Specificity in addition to Accuracy.

1.2 Central Contributions

The principal research questions of the extensive body of work behind the field of epileptic seizure prediction from invasive Electroencephalography (EEG) records, revolves either around the evaluation of a single new feature which best characterises seizure data or a single machine learning algorithm that accurately detects seizure data. There is little evidence of successful studies where both these principles are considered. In this thesis, we aim to use a generally good machine learning algorithm that is suitable for seizure prediction with multiple established features in a combination of experimental settings on a large set of patient data in order to establish a road map to improved, individualised, seizure prediction.

The seizure prediction literature presents little evidence of seizure prediction in advance of the seizure onset, validated on a large population of patients. In this thesis we further explore the area of advance prediction of seizures in an aim to reveal whether statistically valid outcomes can be obtained across a diverse range of patients.

The advancements in machine learning algorithms have led to improved prediction results particularly among toy datasets, although very few of these novel

methods have been applied to the field of seizure prediction. In addition to exploring potential improvements in individualised seizure prediction, we aim to apply some of these powerful machine learning algorithms to the complex problem of multi-patient seizure prediction of unseen patients in order to evaluate the possible benefits of using these methods in this respect.

The first contribution is the implementation of extensive many-patient experiments on feature selection and machine learning in seizure detection and prediction, showing that robust and statistically validated performance can be achieved with appropriate feature selection strategies.

The second contribution is that by using feature selection methods and drawing from previous studies and empirical results, we can produce a new set of features that lead to a better performance than the previous set. The statistical significance of our findings is verified across the entire spectrum of patients from our dataset.

The third contribution shows that epileptic seizures can be predicted up to 25 minutes prior the physiological onset with high levels of Accuracy, Sensitivity and Specificity. The advance prediction of seizures can yield higher performance than the onset detection in special cases.

The fourth contribution in this thesis indicates that by using machine learning algorithms suitable for multi-source learning that are trained on the invasive EEG of multiple patients with epilepsy, the generalisation of the epileptic state for unseen patients can be improved with acceptable levels of Sensitivity and Specificity.

1.3 Overview of the Thesis

Chapter 2 provides a summary of machine learning and signal processing concepts and algorithms used in this thesis, with particular emphasis on supervised and semi-supervised machine learning algorithms and the main data features used in this study. The features presented in this chapter are derived from the work of Costa et al. (Costa et al. 2008). The chapter also provides a summary of the curse of dimensionality and ways of dealing with it, focusing on feature selection algorithms used in this thesis. There is a rich body of work associated with each topic presented in this chapter and it is out of the scope of this thesis to discuss every concept in great length. Therefore, this

chapter should mainly be regarded as a guided summary and full details of concepts should be sought from the relevant references.

Chapter 3 presents a summary on the main problems addressed in this thesis as well as relevant literature to provide sufficient background on the body of work, along with the state of the art. Similar to chapter 2, this chapter merely outlines concepts about epilepsy and seizure prediction that provide a sufficient basis to understanding the core questions of this thesis and for an in depth understanding of the particular topic, relevant references should be visited.

Chapter 4 presents the dataset used in this thesis and noteworthy characteristics of it. Some of the outlined characteristics are the number of patients, attributes of patients, data acquisition techniques and length and format of data. This chapter also includes general data preparation steps and the class distribution of features presented in chapter 2.

Chapter 5 is the first results chapter and provides insight into the preliminary study of seizure prediction, which determines the principal themes for the rest of the thesis. This chapter is part of a research paper published in the proceedings of the Third World Congress on Nature and Biologically Inspired Computing (Moghim & Corne 2011). While this work does not contain any of the contributions listed in section 1.2, all contributions of this thesis stem from the preliminary empirical findings of this chapter.

Chapter 6 presents the second set of seizure prediction experiments. It is largely based on the principles of feature selection. The chapter provides an exhaustive evaluation of features under several experimental conditions, such as various channel recording and dataset setting and further statistically verifies the findings on more patients from the dataset. The empirical results from parts of this chapter ultimately result in the derivation of a new and further improved feature-set which yields a relatively higher performance on held-out test-sets compared to our benchmark. This chapter largely supports the first and second acclaimed contributions of this thesis.

Chapter 7 presents the third set of seizure prediction experiments. This chapter is largely based on the advance prediction of seizures under several experimental conditions in an effort to explore high performance prediction of seizures prior to the physiological manifestation. The findings of this chapter were statistically validated on

all patients from the Freiburg EEG Database (Epilepsy.uni-freiburg.de 2007) where possible, and support the third contribution (see 1.2) of this research.

Chapter 8 presents automatic seizure detection in a multiple patient environment where classification models are trained on a set of patients whilst being evaluated on another set of unseen patients and, also on the unseen test-set of the same patients. Studies presented in chapters 5, 6 and 7 were solely based on individualised seizure prediction methods, while this chapter explores several aspects and potential improvement of multi-patient prediction. The chapter also highlights the role of diversity of the training-set on various forms of classification generalisation. Additionally, the chapter introduces other classification algorithms that are better suited for the task of multi-patient seizure prediction, with the aim to showcase the power of a well-selected machine leaning algorithm on improved multi-patient generalisation of seizure states. The work in this chapter supports the fourth contribution item of this thesis (see 1.2).

In Chapter 9 we revisit the contributions made at the start of the thesis under the light of evidence and information presented in the chapters 2 - 8. The main contributions are further broken down to elaborate auxiliary findings of this thesis. The chapter also gives an outlook on potential future work.

Chapter 2
Machine Learning and Signal Processing

In this chapter we present the machine learning, signal processing methods and evaluation techniques used throughout this document. This is, however, not an extensive review of such methods, but rather, a mere introduction to what they are. Throughout this thesis, we apply these methods in an attempt to aid our data-driven exploration of epileptic seizure predictability. Good recent books covering these methods include (Duda et al. 2012; Barber 2012; Bishop 2006; Hastie et al. 2009).

In the first section of this chapter we introduce terminology and mathematical notations that will be used throughout this document in an effort to avoid ambiguity. In section 2.2, we present the feature engineering process and in 2.3 we briefly review the intuition behind feature selection. In section 2.4 we review supervised learning methods used in this thesis and section 2.5 introduces some examples of semi-supervised learning methods. In section 2.6 we present data and model selection techniques and we conclude with a summary in section 2.7.

2.1 Machine Learning Terminology and notation

The term *Machine Learning* denotes algorithms that learn from data. The discipline intertwines with *pattern recognition*, the process of searching for patterns within data, historically used in physics.

The input to a machine learning algorithm is a *training-set*, which is a collection of *labeled* or *unlabeled* data. In the case of *supervised* learning, the training-set comprises data with labels. The machine learning *algorithm* uses a large set of data to find a model of mappings between training data and training labels. This dynamic model then predicts the label of the unseen data (or *test data*) and outputs the target vector; each unseen data point has a specific target vector.

Machine learning algorithms use a large collection of data points, namely training data to tune the *parameters* of a dynamic model. The algorithms used in machine learning are predominantly based on statistical methods; however, biologically inspired and graphical methods are also used in many cases. *Generalisation* in machine

learning is the ability of the constructed model to correctly categorise the unseen data. The training-set and test-set are sometimes *pre-processed* in order to simplify the pattern recognition or enhance the performance of the machine learning algorithm.

M : number of training examples

X : input variable/features

Y : output variable/target

(x, y) one training example, $(x(i),y(i))$ i^{th} training example,

$\left(x_1, x_2, \ldots, x_n\right)$ is the feature vector , where n is the number of the features.

H is a function that maps from $X \rightarrow Y$

The collection of *m* training example for a training-set with *n* number of features, is the feature matrix *X*.

2.2 Feature Engineering

The particular composition of features presented in this chapter was grouped together in a multi-feature study conducted by (Costa et al. 2008). The selected features are amongst some of the most powerful and well-studied features presented in the literature, which are commonly used in single-feature classification case studies (discussed further on in chapter 3). These features are all uni-variate, meaning that they are derived from a single channel.

The features are divided into three general categories: signal energy, wavelet transform and non-linear dynamics.

Signal Energy and Power

The EEG files are digitally recorded, where continuous data are discretised through digitisation in order to be digitally processed. Drawing from concepts of signal processing, we define the digital EEG as a sequence of complex numbers. Amongst the signal properties of discrete-time signals, signal energy and accumulated power are some of the common yet powerful features used in EEG feature engineering.

The energy of a discrete time signal (Prandoni & Vetterli 2008) is define as:

$$E_x = ||x||_2^2 = \sum_{n=-\infty}^{\infty} |x[n]|^2 \qquad (1)$$

The value of the signal is squared in order to produce a positive measure of the area under the signal curve, which could either be negative or positive. As presented in chapter 3, the changes in the transitions from pre-seizure to seizure states are tractable in signal energy values (Mormann et al. 2007). We define the power of the signal as:

$$P_x = \lim_{N\to\infty} \frac{1}{2N} \sum_{-N}^{N-1} \left| x[n]^2 \right| \tag{2}$$

This limit however is undetermined due to the periodic nature of infinite energy signals and is instead defined by the average energy over a time period:

$$P_x = \frac{1}{N} \sum_{n=0}^{N-1} \left| x[n]^2 \right| \tag{3}$$

Figure 2.1 displays the Signal Energy of patient 2 from the Freiburg EEG Database, captured from all 6 channels. From Figure 2.1, we observe that the signal energy produced by each channel varies particularly during the seizure.

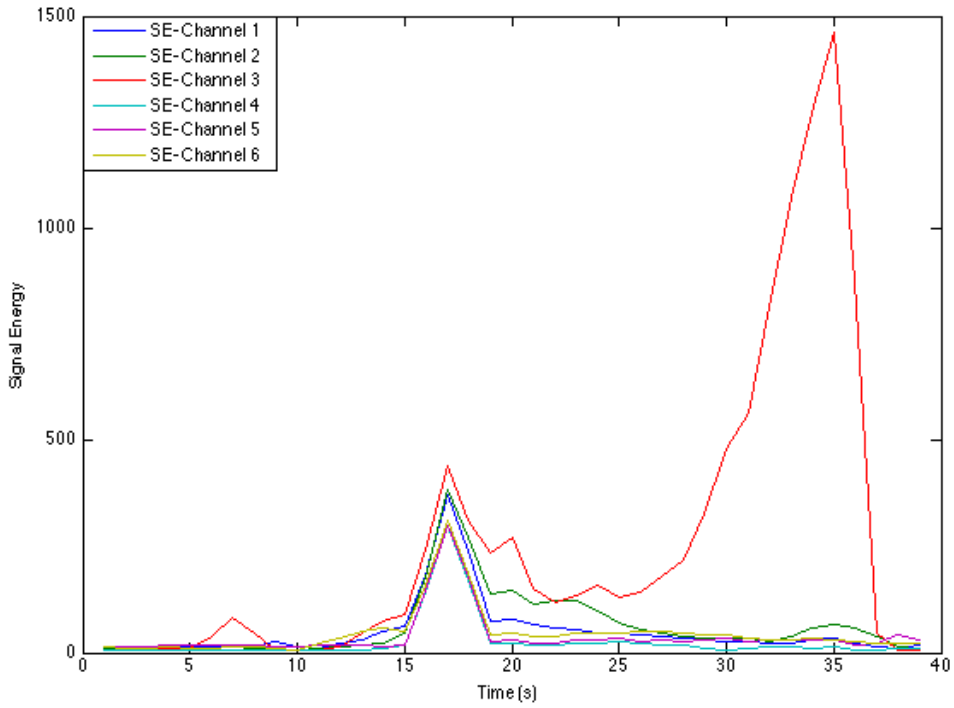


Figure 2.1 Signal Energy over 6 EEG channels for patient 2 of the Freiburg EEG Database – There is ictal activity from second 5 through 35.

The accumulated energy is another powerful means of finding abnormal behaviour in the brain. The accumulated energy is the sum of successive values of signal energy in a moving window analysis. Here, the signal power is integrated over a sequence of time windows:

$$AE(t) = \sum_{k=1}^{t} \sigma_k^2 \tag{4}$$

$AE(t)$ gives us the accumulated energy at time t, by the accumulation of the signal energy variance σ_k^2 in time window k, starting from 1 to t. Figure 2.2 displays the changes in accumulated energy of patient 2, calculated for each of the 6 EEG channels. The image displays similar trends of accumulated energy through different seizure-states for most channels. Channel 3 however displays a noticeably different trend of accumulated energy and signal energy (as seen in Figure 2.1).

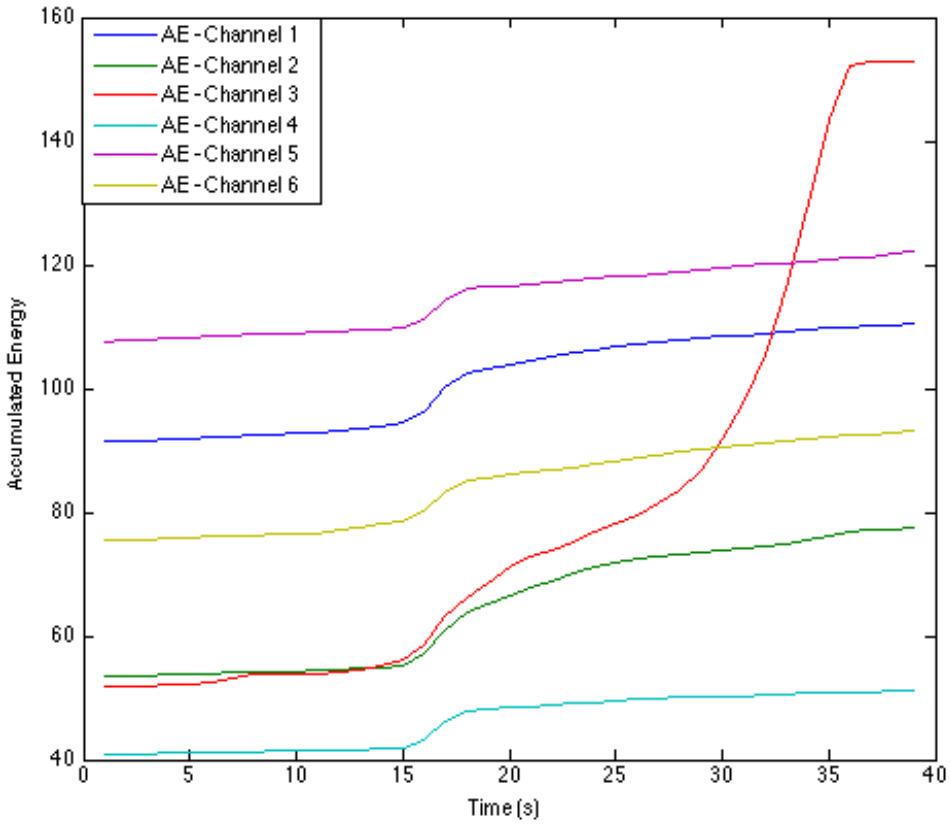


Figure 2.2 Accumulated Energy over 6 EEG channels for patient 2 – There is ictal activity from second 5 through 35.

Both of these features were calculated over an overlapping moving window analysis. The length of the window is 7 seconds with sampling rate of 256 Hz, and the overlap between each window is 50%.

Signal energy is further calculated over variable size windows to construct two additional features. We consider using two different lengths of sliding windows. The short-term energy and the long-term energy encoded respectively as STE and LTE. The length of the STE window is 9 seconds and the length of the LTE window is 180 seconds. This produced two features at each 5-second timepoint.

Wavelet Transform Accumulated Energy

Discrete Wavelet Transform is a space-time form of signal analysis, which decomposes the signal at different frequency bands and samples wavelets in a discrete way. It considers both spatial and temporal properties of the signal and can therefore capture characteristics that may have been missed by single-modal features (Gigola et al. 2004). The signal energy of each decomposition level is determined by:

$$E_j = \sum_{i=1}^{N_j} d_j^2(i) \tag{5}$$

Where j is the decomposition level and $d(i)$ corresponds to the decomposition co-efficient. The wavelet decomposition coefficients are used to determine the accumulated energy of each frequency band using a short term and long-term energy, moving-window analysis:

$$AE_j(k) = \sum_{i=a(k+1)+1}^{a(k+1)+b} d_j^2(i) + AE_j(k-1) \tag{6}$$

where k is the current time-segment, b is the length of the window and b-a is the overlap between adjacent windows.

Wavelet decomposition was carried out over the Daubechies mother wavelet with decomposition level 4 using EEGLAB, an open source EEG analysis and visualisation tool (Sccn.ucsd.edu 2011). Daubechies discrete wavelet transformation (Daubechies & Sweldens 1998) constructs orthogonal wavelets and through scaling, multi-resolution wavelets can be attained. The wavelet accumulated energy was

calculated over a Long-Term-Energy window (LTE) of 180 seconds and a Short-Term-Energy window (STE) of 9 seconds long.

Non-linear dynamics

Non-linear features have had a mixed review in the EEG signal-processing community. In some studies they have been suggested to be superior in performance in comparison to the linear features due to the aperiodic and unpredictable behaviour of seizures (Iasemidis et al. 1990; Le Van Quyen et al. 1999), while other studies suggest that linear attributes perform as well, if not better than non-linear dynamics (Mormann et al. 2005). Non-linear features are drawn from the theory of dynamical systems (Schuster & Just 2006; Kantz & Schreiber 2004; Ott 2002) in contrast to the direct derivation of linear methods from the time-series signal. Non-linear dynamical systems can represent chaos, a perceivably unpredictable behaviour that is fundamentally deterministic. Dynamical systems capture the behaviour of a system in different states in time through fixed deterministic rules, and the states at any given time are derived from a state space. The following are two dynamical system attributes used in this thesis:

Maximum Lyapunov Exponent

In deterministic systems, Lyapunov exponent (Rosenstein et al. 1993; Kantz 1994) is the rate of separation of infinitesimally, close trajectories in the phase space (i.e. the space where all possible states of the system are represented) and is defined as:

$$d_j(i) \approx C_j e^{L_{\max} i \Delta t} \tag{7}$$

Where $L_{\max}$ is the Lyapunov exponent and C is the measure of the initial separation.
The maximum (also referred to as largest or maximal) Lyapunov exponent is the largest Lyapunov exponent among the spectrum of Lyapunov exponents which encompasses various divergence rates for different orientations of the initial separation and is derived from the following approximation:

$$\ln d_j \approx \ln C_j + L_{\max}.i.\Delta t \tag{8}$$

and is calculated using a least square over

$$y(i) = \frac{1}{\Delta t}\left\langle \ln d_j(i) \right\rangle \tag{9}$$

where *y(i)* is a line averaged over *j*.

Correlation dimension

The correlation dimension (Grassberger & Schreiber 1991) is an estimation of the number of active degrees of freedom of random points within a state space. It is denoted with *v* and is calculated using the correlation integral (Grassberger & Procaccia 1983). The correlation integral (or sum) is the estimate local probability density of the points in a state space:

$$C(\varepsilon) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \Theta\left(\varepsilon - \left\|\vec{x_i} - \vec{x_j}\right\|\right) \tag{10}$$

where N is the number of states, $\|\bullet\|$ is a norm, Θ is the so-called Heaviside step function. The integral is calculated on the number of pairs of vectors in a radius ε.

The correlation dimension is defined as:

$$D_2 = \lim_{N\to\infty} \lim_{\varepsilon\to 0} d(\varepsilon) \tag{11}$$

$$\text{Subject to } d(\varepsilon) = \frac{d \ln C(\varepsilon)}{d \ln \varepsilon}$$

The correlation dimension is very quickly calculated and is robust to noise, making it one of the more suitable measures for dimensionality estimation.

Non-linear dynamics, namely Maximum Lyapunov Exponent and Correlation Dimension were calculated over a 7 second window size with 50% overlap between adjacent windows.

2.3 Feature Selection and Dimensionality Reduction

In machine learning studies, high-dimensional datasets may result in what is known as the curse of dimensionality (Bishop 2006). The existence of redundant or irrelevant data in the dataset, can lead to higher classification error while yielding low training error. This is also known as over-fitting. The redundant features provide no further information and irrelevant features are irrelevant to the learning task at hand. Methods exist which reduce the number of features in a dataset and have proven to be useful particularly in datasets, where the number of learning examples is low compared to the number of features. There are two general ways for reducing dimensionality: feature selection, which assembles a smaller subset of optimal features, and feature extraction, which projects the high-dimensional feature space to a low-dimensional feature space. We only consider feature selection methods in line with our research, due to the inherent ambiguity prevalent in feature extraction outcomes such as Principal Component Analysis, which make them unsuitable for this study.

2.4 Feature selection

Feature selection methods extract a subset of features from the full feature-set. This can improve the interpretability of the learning model, and can decrease computation time of learning and also avoid over-fitting.

Feature selection methods can produce feature-rankings as an auxiliary outcome of improving prediction. In some cases, these rankings are not only used in prediction, but are also used for better interpretation of complex features, examples of which are seen in microarray analysis, when rankings are used to discover drug leads. Ranking criteria vary among feature selection methods, with each having a particular feature-reduction goal. Some feature selection methods emphasise on improving computation time while others are more suited for interpretability and visualisation purposes.

Feature selection methods generally fall in two categories (Guyon & Elisseeff 2006):

2.4.1 Wrapper Methods

The Wrapper method is a black box feature selection method, which uses the learning algorithm as an evaluation tool. It uses the learning performance of a leaning algorithm to assess the suitability of the feature subset. This makes it highly dependent on the learner in use. Wrapper methods may perform well with a particular learning algorithm, while they may perform poorly with another. They are usually computationally expensive, as they require several runs of the learning algorithm prior to finding the optimal feature-set that yields the highest performance, although with suitable search strategies, the computation cost can be somewhat reduced.

2.4.2 Filter Method

Filter methods are a general type of pre-processing feature selection methods, often resulting in a ranking table. They are pre-processing methods, as they are used prior to classification and are therefore independent from the learning algorithm. The filter method ranks features according to a given criteria, some of the notable ones are minimum Redundancy-Maximum Relevance (mRMR) (Peng, Long & Ding 2005b) which, as the name suggests, finds features with the highest relevance and lowest redundancy in a two-step process, and ReliefF (Moore & White 2007) which ranks those features with the higher discrimination power with regards to the target label. In studies where features and their composition is of high importance, filter methods are mostly used, but are also, at times, used as a pre-processing step to other dimensionality-reduction techniques, such as wrapper methods or feature extraction. This form of feature selection algorithm is also preferred over some wrapper methods due to the low computational time and high scalability power.

For the purposes of this study, we have selected a filter feature selection method, which is widely used in the Information Theory and Bio-informatics community, particularly because of their ability to perform well on very large feature-sets. Although wrapper methods seem to be a better choice for directly improving the classification Accuracy, there are a number of reasons for choosing filter methods over the alternative wrapper methods for the purposes of our study (Guyon & Elisseeff 2006; Saeys et al. 2007); Some of these are explained below:

1- **Computation time:** Wrapper methods are wrapped around the learning model, repeatedly testing feature subsets in the cross-validation set. The implementation of wrapper methods may result in a decreased validation error rate; however, it demands a great deal of computation time for learning the model. Diversely, the filter feature selection methods are separate from the classification entity, reducing the overall computation time by allowing one-off calculation of the rankings prior to learning, followed by model learning on the now reduced feature-set. This is considerably faster than the repeated evaluation of features at each stage of cross-validation, when using wrappers. We will see that this is particularly important as the size of the feature-set and dataset increases in certain experiments, where the learning process becomes lengthier in time. Therefore, deploying a wrapper feature selection method, in addition to the overheads of slow learning process will negatively impact the computational time, without the additional benefits of a filter method.

2- **Evaluation consistency:** In order to be able to plausibly compare and contrast experiments in this chapter and chapter 7, we need to use homogenous methods and settings across all studies; feature selection is one of such methods. Although it would be computationally justifiable to use wrapper methods on smaller datasets/feature-sets and use filter methods once the dimensionality increases, this however poses inconsistency in our methods which makes evaluation of the results of these separate experiments erroneous. Therefore, in view of evaluation consistency, we have used consistent feature selection methods for all the experiments.

3- **Interpretability**: Wrappers are black box methods, providing little information about the selected features, and their corresponding impact on the classifier. They also impose further difficulty in terms of consistent and smooth feature selection in the face of varying cross-validation and training-sets. If the wrapper starts at 14 features and re-orders these features in accordance to the classification error of the validation-set, and then carries another round of feature selection, the ordering of the features could be different in each round of cross-validation, revealing little information about the intrinsic characteristic of the features and their direct impact on the validation-error rate. With filter methods, however, feature selection is carried out in one step, prior to

classification, where a solid and informative table of feature rankings is created. This use of ranking tables, along with full control over ranking updates are of the most important criteria in our choice of feature selection algorithm, which justify the use of filter methods for our series of experiments.

4- **Generalisation**: The wrapper methods are highly coupled with the learning algorithm, meaning they may perform differently under alternate learning models, whereas the filter methods are learner independent, allowing for a better generalisation across various forms of learning. This is particularly of importance for the purposes of our seizure detection study as i) the algorithm in use is not the focus of the research; we have followed a 'dirty' approach to learning, by selecting and further fine-tuning a well-known classifier, which has reportedly been applied to similar problems, according to our review of the literature ii) we aim to analyse the intrinsic properties of the features, independent of the classifier in use. The goal here is not to solely improve prediction, more so to improve prediction on a simple classifier. By manipulating the feature-set according to their inherent characteristic and relevance to the class labels (seizure states), in the effort to identify optimal features, regardless of the underlying learning algorithm.

ReliefF

Relief is an attribute estimator filter algorithm used for feature selection. In the earlier version of Relief (Moore & White 2007), the conditional dependency of features is estimated using a nearest neighbor algorithm, where the quality of the features are assessed by the discriminant power of them, and of values that are near each other. In this method, for each randomly selected data point, the nearest instance from the same class and the same data point from a different class are selected. The features that highly discriminate between the two different classes are given a higher 'quality estimation'. This re-evaluation of features is carried out for m number of random instances, on an n length dataset. This algorithm performs particularly well on large datasets, with as little as 4 minutes on average to rank the features, with computation complexity of $O(m \times n \times a)$ where a is the number of features. The algorithm has widely been used for feature selection on large datasets, as well as in guiding the

induction phases of regression decision trees and various other settings, due to their intuitive interpretability. ReliefF is a variation of the original Relief algorithm, which can deal with multiclass learning (Kononenko 1994).

2.5 Supervised Machine Learning

In supervised learning, examples of correctly labeled test data is used to predict the label of unseen data. One type of supervised learning is classification, in which the *classifier* is trained based on samples of collected data (known as training data), where a suitable model which closely represents the known training data is constructed and algorithmic parameters are estimated. The trained model is then optimised through validation techniques in order to produce an accurate prediction on unseen data. The primary outcome of this process is finding a model that generalises the data based on a particular training-set, and using the constructed model to make predictions on the target value of unseen data. We first present Artificial Neural Networks (ANN) and will follow by introducing the Support Vector Machine (SVM).

Artificial Neural Networks

Artificial Neural Networks (ANN) are perceptron based learning systems, originally attempted to model the information processing in the brain, which can deal with linearly inseparable and high dimentional data (Kotsiantis et al. 2007). Multi layer ANNs are created from three classes of connected neurons (Figure 2.3): input, hidden and output; where the neurons in the network are connected by weighted edges, input units receive training information, output units store the result of the classification, and hidden units are the neurons in between these two units. Feed-forward is the simplest form of ANNs, where signals only travel in one way.

One of the main decisions when designing neural networks is to determine the number of hidden layers, as low number of neurons could result in poor generalisation and too many neurons could result in the model over-fitting the data. Another important feature is the choice of the weights of each input connection; the value of weights are initially set to random values and change at each step of the training, as the result of the output is compared with the desired output.

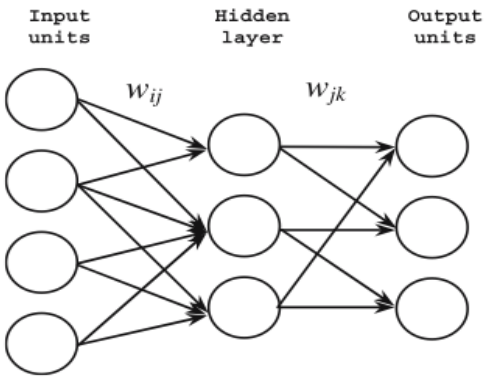


Figure 2.3 Feed Forward Artificial Neural Network (Kotsiantis et al. 2007)

Back propagation is a well-known learning algorithm used for training ANNs where input neurons send signals all the way to the output neurons, passing activation values and weights through the hidden layers. Evolved Neural Networks (EANN) are the standard neural network trained by an evolutionary algorithm.

Feed-forward neural networks most commonly use back propagation, as a result, the training algorithms are usually very slow, but estimating optimal initial weights as opposed to picking random weights can speed up the performance. Genetic algorithms work particularly well for training optimal ANN weights and architecture. Bayesian methods are also reported as being useful for training ANNs. Pruning, where useless nodes are removed, and constructive algorithms where extra nodes are added are also techniques used for improving ANN training algorithms.

The main drawback of ANNs is their black box approach towards problem solving, which does not provide reason or meaningful insight into how the learning problem was solved. Symbolic rules may be extracted from trained ANNs but this is nowhere near the comprehensiveness provided by other alternatives, such as decision trees. Among other problems, are long training times and tendency to converge to the local optima due to poor feature selection.

Support Vector Machine

A support vector Machine is a classification model in which an optimum separating hyperplane divides the instance-label pairs (xi, yi) (Hsu et al. 2003; Chang & Lin 2011; Cortes & Vapnik 1995). The SVM specifies a margin between the separating hyperplane and the data points at either side of the plane (the support vectors). The linear combination of the support vectors on either side of the margin represents the

feature-mapping model, ignoring other features. This makes the SVM suitable for learning tasks with a high number of features.

A support vector machine solves the following optimisation problem:

Given a training-set of instance-label pairs (xi, yi), $i = 1, \ldots, l$

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \text{Subject to} \quad & y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi \geq 0, i = 1, \cdots, l, \end{aligned} \tag{12}$$

Where C is the regularisation parameter, ξ_i is a non-negative slack variable used with soft margin SVMs to measure the degree of misclassification, and $\phi(x_i)$ is known as the kernel function. When data are linearly inseparable, the SVM simplifies this optimisation by mapping the training feature-vector to a higher dimensional space, through a kernel function φ,

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) \tag{13}$$

The kernel function decreases the computational power required for calculating the mappings of data to a higher dimensional space, as it enables making fast classifications without the need to essentially map each data point to a higher dimension. The Kernel used in this study is the Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0. \tag{14}$$

Figure 2.4 visualises the SVM-RBF classification. In order to find the best kernel function parameter, SVM can be used along Cross-Validation.

Multiclass SVM is used for multi-label classification problems. Multiclass SVM is constructed from one SVM for each class label. When presented with new data, the classifier makes a prediction with each SVM and chooses the class label associated with the SVM that placed the prediction furthest into the positive region.

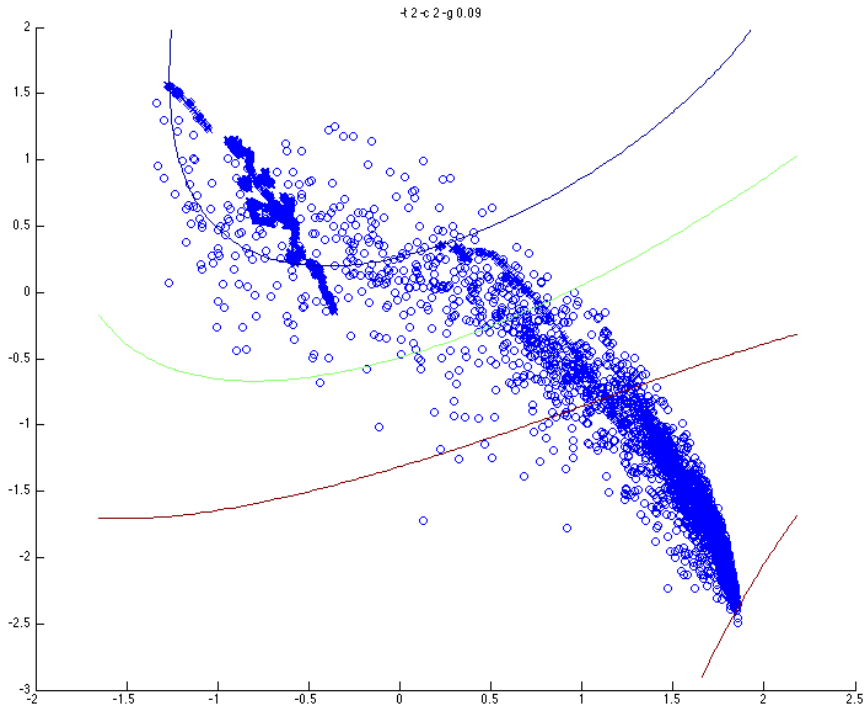


Figure 2.4 An example of binary SVM with RBF Kernel

2.6 Semi-supervised Learning

In machine learning, a learning task can have a combination of labeled and unlabeled data. The semi-supervised learners can be used to learn from both types of data. Two of the notable types of semi-supervised learning are Multi-Task Learning and Deep Learning.

2.6.1 Multi-Task Learning

Multi-Task learning (MTL) (Ando & Zhang 2005; Caruana 1997) is a form of semi-supervised learning, also a subset of inductive transfer learning (Baxter 2000), which enables learning of similar problems at the same time, by using the communally shared between the learning tasks. The diagram in Figure 2.5 depicts the difference between the architecture of Single-Task Learning and Multi-Task Learning. Single-Task learning trains a model on the training-set of a single task and applies it to the test-set, resulting in several learning models for similar tasks.

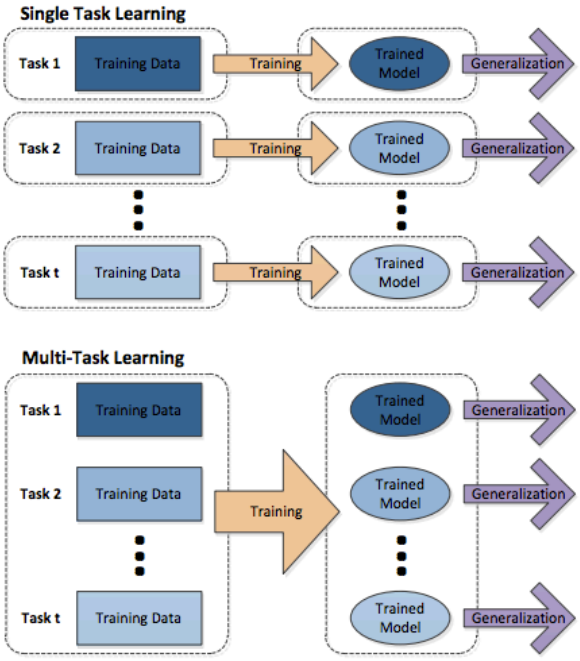


Figure 2.5 Single-Task Learning vs. Multi-Task Learning (Zhou et al. 2012)

MTL uses the common knowledge shared between tasks through a shared representation to generalise over the various *related* tasks. The assumption behind this form of learning is that what's learned from each task can aid the learning of other tasks. Multi-Task Learning, although a semi-supervised learning algorithm, can also be used in solely supervised learning tasks, instances of which are applied in multi-modal learning studies such as multi-source learning for multi-modality neuro-Imaging data (Yuan et al. 2012) where data may be extracted in various modalities.

Multi-Task Learning using Alternating Structural Optimisation

Multi-Task Learning using Alternating Structural Optimisation (ASO) (Ando & Zhang 2005) uses alternating structure optimisation to discover a shared feature mapping between the tasks. ASO is based on the assumption that a common predictive structure exists among related tasks where parallel learning can find the shared low dimensional predictive structure. This principle is displayed in Figure 2.6, where Θ is the shared low dimensional feature map and U_m is the weight vector learnt from task m and V_m is the weight vector learnt from the shared low dimensional feature space of task m.

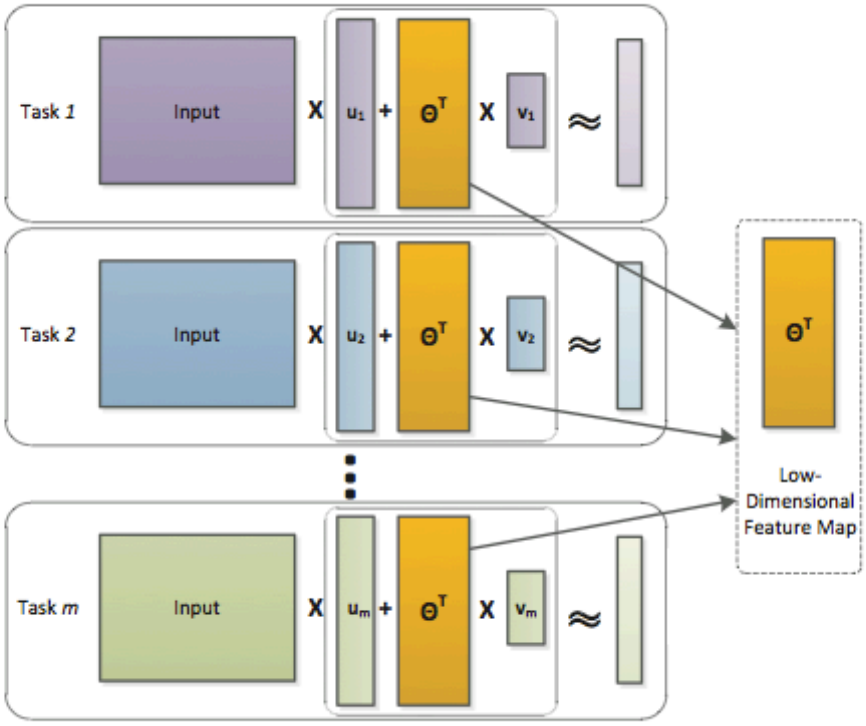


Figure 2.6 Multi-Task Learning using Alternating Structural Optimisation (Zhou et al. 2012)

The model of each task is divided into two main components, which are task-specific feature mappings and task-shared feature mappings. The learning goal in the ASO Multi-Task Learning algorithm is: -

$$\min_{\{v_t, w_t\}, \Theta} \sum_{t=1}^{T} \left(\frac{1}{n_t} L(\mathrm{w}_t) + \alpha \left\| w_t \right\|^2 \right) \tag{15}$$

$$\text{Subject to } \Theta\Theta^T = I, \ \ w_t = u_t + \Theta^T v_t,$$

Where n_t is the number of samples of the t-th task, L is the loss function, α is a predefined regularisation parameter and w_t is the weight vector of the high-dimensional model. The predictor of task t is determined by

$$f_t(x) = w_t^T x = u_t^T x + v_t^T \Theta x \tag{16}$$

This formulation, however, is non-convex. A relaxed, convex alternative, which is also scalable to large datasets is, the convex Alternating Structural Optimisation (cASO) (Chen et al. 2009) and is determined by:

$$\min_{\{w_t\},M} \sum_{t=1}^{T}\left(\frac{1}{n_t}\sum_{i=1}^{n_t} L(\mathrm{w}_t)\right) + \alpha\eta(1+\eta)tr(W^T(\eta I + M)^{-1}W) \tag{17}$$

$$\text{Subject to } tr(M) = h,\ M \preceq I,\ M \in S_+^d$$

Note that M is $\Theta^T\Theta$, and $\eta = \beta/\alpha > 0$, where α and β are pre-determined regularisation parameters. The performance of cASO has been suggested to be as good as ASO on benchmark data, but with reduced computation costs.

2.6.2 Deep Learning

Deep learning refers to a category of semi-supervised machine learning algorithms in which layered models of input data are learned (Bengio 2009). This form of learning is closely related to the development of the human brain in cognitive neuroscience and distributed representation. The assumption behind the learning algorithm, is that different factors of data, which may be known or unknown at time of training, result in several representations of the data, from which generalisation about similar factors in the unseen data is possible. This representation is layered, where each layer represents different levels of data abstraction. The layers in the model form different representations of the input, with the more abstract higher levels drawing from the lower levels.

Deep learning is commonly applied to Neural Networks and can be used in unsupervised and semi-supervised learning. It has been widely applied to studies in speech recognition and signal processing, object recognition, transfer learning and Natural Language Processing.

Deep Belief Networks

Deep Belief Networks (DBN) (Arel et al. 2010; Bengio et al. 2013) are a form of deep learning constructed of multi-layered stacks of Restricted Boltzman Machines (RBM).

Restricted Boltzman Machines are simplified neural networks, which are restricted to have a single hidden unit and a single visible unit. They are suitable as building blocks of more complex models, due to their simple architecture. Deep Belief Networks differ from traditional neural networks, in that they are generative models, which can capture the joint distribution over observed data, and labels from which new data can be generated. Neural networks on the other hand, are discriminative models and are limited to produce the probability distribution of class-labels given observations. More so, DBNs solve a number of problems associated with neural networks such as i) slow learning ii) converging to local optima due to poor parameter selection.

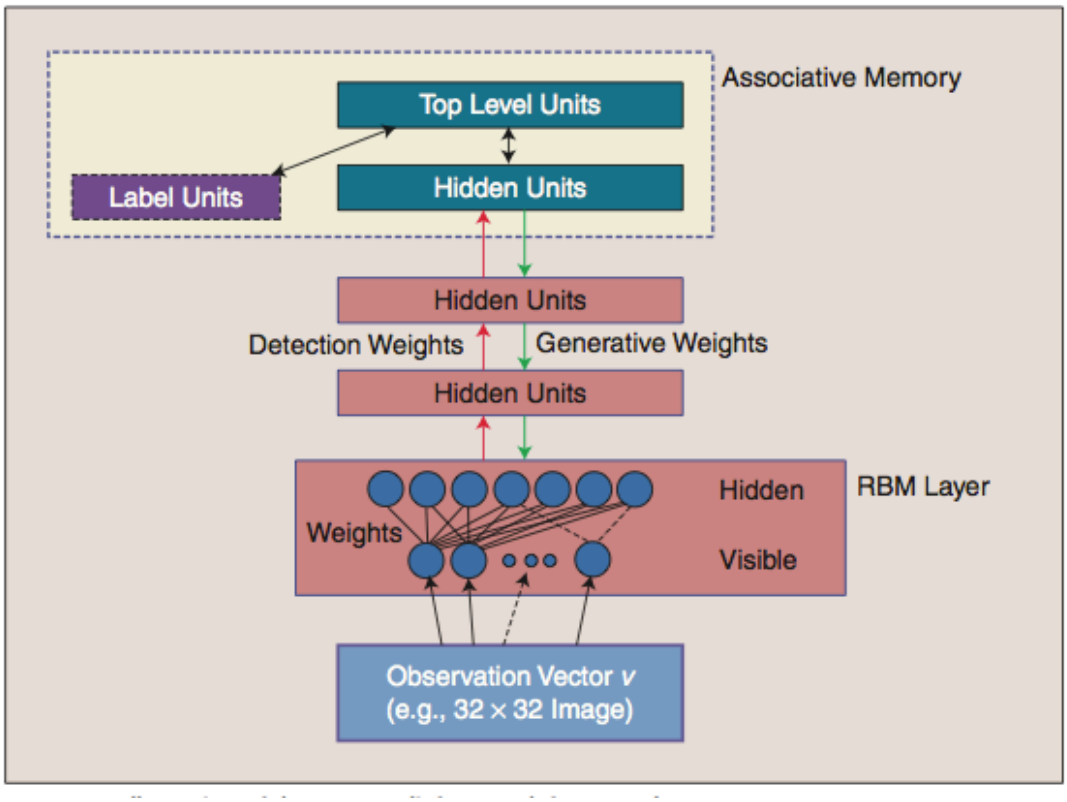


Figure 2.7 The Architecture of Deep Belief Networks (Arel et al. 2010)

The generative DBNs are constructed through a pre-training step where unsupervised, greedy learning, is conducted on a layer-by-layer basis. Visible units pass a vector *v* down to the hidden units, after which, the hidden units reconstruct the visible input according to the vector *v*. The traversal between hidden units and visible units is called Gibbs sampling. The weights are updated based on correlation between hidden layer activations and visible input.

DBN can be used in classification after fine-tuning the outcome of the pre-training step. In order to achieve this, labeled data are used in back propagation where a

top layer is added in order to clarify classification boundaries. This top layer is connected to the pre-training top 2 layers, which contains associative weights. The lower level generations are linked to the top layer and new sets of bottom-up weights are learned.

2.7 Model Selection and Evaluation

In this section, we present methods used in the selection and evaluation of the models employed throughout this thesis. Some of these steps may differ in some applications.

2.7.1 Training Set vs. Test Set Error

Training-set error is the number of records where the predicted value is different from the actual value. The test-set error on the other hand, is when the data at hand, are divided into two parts, training data and test data. We use training data to train the model and calculate the training-set error, and we use the test data to test the model, as if it were the future real data from which the test-set error rate can be calculated.

The aim is to reduce the training-set error rate as much as possible in order for a model to be less error prone when used on future real data. However, in this process we should be aware of noise in the data and over-fitting it. A sign of over-fitting could be a higher error rate of the test data.

2.7.2 Leave-one-out cross-validation (LOOCV)

In this method, one of the records is temporarily removed from the dataset, and the model is trained using the remainder of the data. The error is then found with respect to the left-out data point. This procedure is repeated for each data point. At the end of the loop, the mean error rate is calculated. This method does not waste data but is computationally expensive.

2.7.3 K-fold cross-validation

Cross validation is an approach used for preventing the model from over-fitting the data. Cross validation works by estimating the Accuracy of the model learnt from some training data, against future unseen data. In this method, the dataset is randomly broken

down to k partitions. For each partition, the model is trained on the data point that is not in the partition and is tested against those in the partition. In essence, there will be k models, from all of which the mean error is calculated.

CV can be used in model selection; using k-fold CV. the best model among a number of candidates is selected. That model will be used and trained with all of the data. A CV can also be used for choosing the kernel parameter in kernel regression or locally weighted regression and the Bayesian prior in the Bayesian regression. These involve real-valued parameters. In a classification problem, CV can be used to calculate the total number of misclassifications on a test-set instead of sum-squared errors on the test-set. CV is also used for feature selection, where the features that are most useful to the learning algorithm are picked, using stochastic search (simulated annealing or genetic algorithms), hill climbing, backward elimination or forward selection.

2.7.4 Evaluation Measures

There are several evaluation measures in the field of machine learning that are used for evaluating both the training outcome and the test outcome. Some of the most notable methods are mean squared error and Accuracy. The choice of evaluation measure, mainly depends on the common measure used in the relevant line of research for ease of cross-study result comparison. The following are the evaluation criteria that will be used throughout this report:

Accuracy

Accuracy is the most common evaluation measure in machine learning research. It compares the predicted output y against the target output $\hat{y}$ and calculates the percentage of those output labels predicted accurately.

$$Accuracy = \frac{TruePositives + TrueNegative}{Positives + Negatives} \tag{18}$$

Specificity (true negative rate)

In skewed datasets as well as particular lines of research such as seizure prediction, Accuracy is not solely representative of the performance of the model. In a life-critical

application such as seizure prediction, it is also important to evaluate the positive predicted values also known as the true negative rate or Specificity.

$$Specificity = \frac{TrueNegative}{FalsePositives + TrueNegatives} \tag{19}$$

Specificity is the ratio of the true negatives (the correctly predicted negative target value) against the sum of all negative target values. In the case of seizure detection, Specificity is the measure of how many of the states that were predicted as non-seizure were correctly identified.

Sensitivity (Recall)

The true positive rate, also known as Sensitivity or recall, is the measure of true positives to the true positives and false negatives.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{20}$$

In the context of seizure prediction, this measure reveals the percentage of the target seizure states, which were accurately predicted. The preference between higher values for Sensitivity or Specificity, is domain dependent. Table 2.1 presents the confusion matrix of the seizure prediction scenario. A true positive is when a prediction was made and this prediction was followed by a seizure; the true negative on the other hand is when neither a prediction of a seizure has been made, nor a seizure has occurred.

	Seizure occurred	Seizure did not occur
Alarm raised	a TP	b FP
Alarm was not raised	FN c	TN d

Table 2.1 The confusion matrix for seizure prediction.

S1-Score

When Accuracy is not solely sufficient for evaluating the prediction outcome, precision and recall are often used as main or auxiliary measures. The trade-off between Sensitivity and Specificity is mainly domain specific, and is subject to application requirements.

Sensitivity and Specificity are also of importance when dealing with skewed data. In a highly skewed dataset with 99% negative examples and 1% positive examples, a random prediction can have an Accuracy of 99%. In such datasets, the Accuracy is not a representative measure of the prediction, and performance should be verified using Sensitivity and Specificity, which take the number of positive and negative instances into account. In order to facilitate loss functions to only include one measure, instead of both Sensitivity and Specificity, and for ease of evaluation we use the following measure, which incorporates both measures of Sensitivity and Specificity:

$$S_1 = 2 \times \frac{Sensitivity \times Specificity}{Sensitivity + Specificity} \tag{21}$$

The S1-Score is the harmonic mean of Sensitivity and Specificity, and equally captures both measures.

2.8 Summary

This chapter presented an introduction to some of the machine learning and signal processing concepts and algorithms used throughout this thesis. Some of the topics were presented at a very high level of abstraction. There are two reasons behind this: i) All presented concepts have rich bodies of research attached to them, which mean they are a deep research field on their own. Providing full details of the topics is out of the scope of this thesis. ii) Implementation details of some methods were omitted and will be presented in future chapters.

Chapter 3

Epilepsy, Seizure Detection and Prediction

This chapter presents a brief introduction to Epilepsy and seizure prediction in addition to research relevant to the subject of this thesis. In section 3.1 an introduction to epilepsy, its categorisation, diagnosis and treatment is presented. This section is particularly important as it provides the neuro-scientific details crucial for understanding the thesis and the questions it poses. In section 3.2 EEG and its various roles with respect to epilepsy are outlined. Section 3.3 introduces seizure prediction and presents relevant research in the field ranging from the early stages of research to the state of the art. This section also provides an overview of common techniques used in seizure prediction studies as well as common challenges associated with research in this field.

3.1 Epilepsy

Epilepsy is the second most common neurological disorder, affecting 0.6-0.8% of the population of the world (Mormann et al. 2007). In this chronic, neurological disorder, abnormal activity of the brain causes seizures (Perucca et al. 1998). Seizures are defined as "sudden, brief attacks of altered consciousness" (Elger & Schmidt 2008). Various types of epilepsy are categorised by the types of seizure, the causes of seizure, the age at which the seizures begin (also known as age of onset), the patterns of EEG during and between seizures, severity and frequency of seizures, the part of the brain involved, whether the disease is inherited, other disorders, prospects of recovery or worsening. The International League Against Epilepsy (ILAE) has come up with a standardised classification and terminology for epileptic seizures and syndromes (Berg et al. 2010).

3.1.1 Categorisation of Epileptic Syndromes

There are three main types of epilepsy syndrome. Cryptogenic, where epilepsy is due to an unidentified focal abnormality; Symptomatic, where the epilepsy is due to a known, structural abnormality, such as mesial temporal sclerosis, cortical dysplasia, arterio-

venous malformation, stroke, or cerebral palsy; Idiopathic, which is due to genetic factors, such as childhood absence epilepsy and juvenile myoclonic epilepsy (Kwan & Brodie 2000). Generally speaking, the cause of epilepsy is unknown for about half of the patients suffering.

Another common categorisation of the epilepsy syndrome is defined by the location and extent of spread of epilepsy. The two common classes are generalised (Cascino & Sirven 2011) and focal (Elger & Schmidt 2008). In generalised epilepsy, seizures occur due to the general lowering of a seizure threshold. They start in the entire cortex, and are usually due to genetic factors. In partial (also known as focal or localised) epilepsy, focal and localised changes in the function of the brain causes seizures that vary in speed and extent of spread and onset location. The distinction between the two is important, as Anti Epileptic Drugs (AED) that are suitable for one syndrome, may cause poor seizure management and drug-resistance in the future if applied to other syndromes.

Focal seizures fall into two categories of simple and complex, and can cause seizures such as smacking of lips, visual hallucination, complex automatic behaviour etc., depending on what part of the brain is the focal area. Where the focal changes spread throughout the brain rapidly, the seizure is categorised as being a generalised tonic-clonic seizure, and categorised as *secondarily generalised,* while the onset is primarily focal. A full reference of further classification of epilepsies can be found in (Elger & Schmidt 2008).

3.1.2 Diagnoses

The diagnosis of epilepsy is made after the patient has already had a number of seizures (Elger & Schmidt 2008). Depending on the number, duration and type of seizures they have experienced, they may be diagnosed with epilepsy. After hypothesizing about the type of seizure, type of epilepsy and type of epilepsy syndrome, tools such as Electroencephalography (EEG), Magnetic Resonance Imaging (MRI), and blood tests (to determine levels of glucose, sodium, and magnesium) are carried out. MRIs are used to find structural causes of the seizure, and EEG is used to assess the diagnosis hypothesis. EEG is used in the inter-ictal (between seizures) state to confirm the syndrome as well as localisation in the case of focal epilepsy.

In an EEG, the neuronal discharges are recorded through electrodes placed on the scalp (surface EEG), the shape and motion of which (the morphology), could be used to confirm whether epilepsy exists or not and classify the epilepsy syndrome according to established spike patterns. Figure 3.1 shows a series of discharges for different types of spikes and waves. The EEG is usually taken several times in order to confirm the epileptic state of the patient, as the inter-ictal recording can bear normal neuronal activity 30% of the times. MRIs on the other hand, are used to detect the existence of a structure, which could potentially lead to symptomatic epilepsy, such as a brain tumor, brain injury, etc.

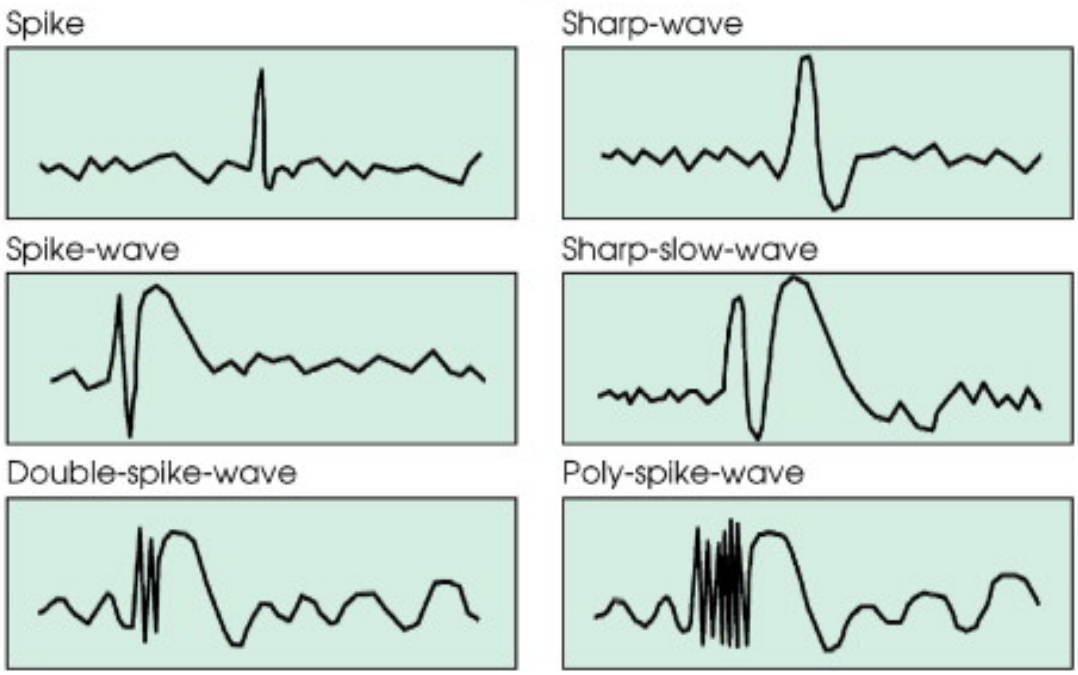


Figure 3.1 Various Forms of Inter-ictal discharges observed in EEG recordings (Elger & Schmidt 2008)

3.1.3 Treatment

The most common method of treatment prescribed is Anti-Epileptic Drugs (AED), which are highly effective in most patients (65%) while ineffective for 35%. The most suitable AED for a patient is prescribed on clinical grounds rather than the pharmaceutical knowledge of the chemical mechanism. This is due to a lack of understanding of the seizure mechanism itself (Elger & Schmidt 2008).

The choice or combination of AEDs depends on a variety of factors such as the type of the syndrome, safety, tolerability, pharmacokinetics (drug-drug interaction) and other factors. These drugs are usually advised for patients who have debilitating seizures such as generalised tonic clonic. AEDs are generally used to prevent future seizures from happening, as well as reducing the severity of the seizures while maintaining the normal functionality of the brain. They do not however, cure epilepsy and are merely, a tool for long-term seizure management. Another disadvantage of AEDs are the associated side effects, some of which are; weight problems, central nervous system toxicity, and depression.

AEDs only prevent seizures from occurring and do not affect the underlying epilepsy syndrome and the cause of it. They treat the seizure but not the epilepsy. One new trend in epilepsy treatment is Anti-epileptogenesis, which are strategies that are hoped to prevent the generation of seizures themselves. So far, animal models of chronic epilepsy have been developed and the synaptic plasticity of underlying causes of susceptibility to epilepsy are being studied in an effort to understand epilepsy itself, and which may someday, be able to prevent or reverse it.

There also exist a number of non-pharmacological therapeutic methods for preventing seizures. These consist of measures a patient can take in order to avoid seizure precipitation, such as sticking to a scheduled and fixed sleep-plan. Interruptions to sleep/wake cycles could cause seizures in some patients. In patients whose seizures are triggered by television, watching at a distance and using small screens can avoid this. There are many more examples such as this, requiring life-style changes in order to avoid the likelihood of seizures.

Another method, popular in drug-resistant patients, is resective surgery. In this surgery the part of the brain that causes the seizures is removed, resulting in fewer seizures post-surgery. Those patients on whom the surgery can be performed and who are deemed suitable are carefully selected. These are usually patients who have focal epilepsy and are resistant to drug therapy. 25-30% of patients are cured by this method and a further 25-30% become seizure free, or relatively seizure free. In order to correctly localise the region of the brain, which is to be operated on, MRI, Single Photon Emission Tomogram (SPECT) and invasive EEG are used.

Neuro-stimulation is another class of therapeutic methods for treating epilepsy. These are Vergas Nerve Stimulation (VNS) (Elger & Schmidt 2008) and brain

stimulation (Theodore & Fisher 2004). In VNS, a pacemaker is implanted near the Vergas nerve. The patient can then activate the pacemaker when they sense the onset of a seizure occurring. This makes VNS suitable for patients who experience an aura immediately before seizures. It is usually an alternative to surgery, has minimal complications and reduces partial seizures by one-third.

Brain stimulation has only been investigated within the last 50 years as an alternative to surgery. It has yet to evolve more before it is suitable for real-life application, but a number of research teams are working on pacemakers that reset the state of the brain as the seizure is automatically detected. Neuropace (Vachtsevanos 2003) was recently granted a Food and Drug Administration (FDA) approval for their deep brain stimulation pacemaker in early 2013. The field is still very new and requires extensive research to improve patient-specific tuning and more sensitive detections.

3.1.4 Drug-resistant Epilepsy

Antiepileptic Drugs are used as long-term therapeutic solutions to treat epilepsy, under which, the epilepsy is not cured, but rather controlled. 65% of patients suffering from epilepsy, who have bean treated by AEDs, fall into long-term seizure remission. However the remaining 35% have a relapse and become resistant to anti-epileptic drugs, therefore, continue to have seizures (Schiller 2009).

Drug resistance (also known as refractoriness) among epilepsy patients can result in many undesirable and even life-threatening consequences for the patient (French 2007). Drug resistance can result in prolonged and unforeseen seizures. Lack of control in seizures can lead to a higher risk of body injuries and can increase the likelihood of the occurrence of Sudden Unexpected Death in Epilepsy (SUDEP) (Elger & Schmidt 2008). Besides physical risks, patients with uncontrolled seizure are also prone to neuropsychological, psychiatric and social impairments, which are believed to reduce employment and can affect marriage rates and decrease the overall quality of life. The increasing of drug dosage and change of AEDs, which is commonly practiced for refractory epilepsy, can also result in memory-loss and other unwanted side effects.

Risk Factors

The study (Baumgartner et al. 2005) has reported substantial evidence that shows one third of patients with newly diagnosed epilepsy will in the long term develop drug-

resistant epilepsy. This has even been proven in the AED trials where 10% of patients who are resistant to drug, remain drug resistant throughout the trial; Having tried out several treatments, their epilepsy will improve in the short term, but there seems to be no resolution for long term seizure control. Another study (Kwan & Brodie 2000) argues that patients, who have failed their first few AED treatments, are likely to develop drug-resistant epilepsy in the long term.

(French 2007) has reviewed predictive markers of refractory epilepsy. They point out the role of genetic factors in predetermining the rate of drug absorption in an individual, its metabolism and blood brain barrier permeability. Genetic factors can affect how a drug is broken down and what metabolic effects are formed and how effective is the delivery of the drug to the brain. The Blood Brain Barrier (BBB) is a complex structure, protecting the brain against chemicals in the blood. The cells on BBB have several pumps, which allow glucose to pass into the brain and prevent other substances from entering the brain. One of these pumps is P-glycoprotein, which is believed to be the main cause of decreased absorption of the AEDs by the brain, and consequently results in making the drugs less effective (Epilepsyresearch.org.uk 2013).

Other predictors for drug resistance are early age of seizure onset, abnormal neurological exam, partial seizures at diagnosis, mixed seizure types associated with developmental delay, abnormal EEG activity, failure to gain control of seizure early in the therapy, multiple seizures prior to treatment. Multiple seizures after the treatment and certain genetic syndromes are also potential indicators of drug resistant epilepsy (Berg et al. 2001; Cockerell et al. 1997).

3.2 EEG

Electroencephalography (EEG) is used for recording the electrical activity in the brain, using multiple electrodes placed on the scalp. The EEG has the ability to capture waves of neuronal activities across the brain which are the result of a cascade of neuron movements, and can therefore pick up on abnormal activity in the brain (Niedermeyer & da Silva 2005).

3.2.1 Diagnostic

The EEG is a very powerful diagnostic tool for many neurological disorders, epilepsy in particular. It can be used to distinguish between epileptic and non-epileptic seizures. EEG can provide information about the location of the brain where the abnormality is created, and also can be used for identifying the type of epilepsy syndrome (Niedermeyer & da Silva 2005).
The EEG is used for covering 3 main diagnostic grounds (Noachtar & Rémi 2009):

Epilepsy or Not

The EEG can determine whether a patient has epilepsy or not. This is done via analyzing the inter-ictal (between seizures) epileptiform discharges (IEDs). An IED is the none-seizure state of the brain. IED is hardly observed among normal patients without epilepsy (2%) and is highly accurate for the majority of epileptic patients (98%). The IEDs are not always present in the first EEG; therefore, through repeated or long-term EEG recordings, the diagnostic Sensitivity is enhanced. One important thing to note about IEDs is that they are difficult to identify even by the experts due to a lack of objective definition. Therefore, the diagnosis is subject to under-interpretation and over-interpretation. The key is to find well-known epileptic patterns in the brain signal using the following established IED forms: Spikes, Sharp waves, Benign epileptiform discharges in childhood, Spike–wave complexes, Slow spike–wave complexes 3-Hz spike–wave complexes, Polyspikes, Hypsarrhythmia and Seizure pattern, examples of which are illustrated in Figure 3.1.

Identifying the epileptic zone

The EEG patterns are linked to a number of established, epilepsy syndromes, which can help with choosing the appropriate treatment and prognosis. The main categories of the syndromes are generalised and focal epilepsy. The different forms of the two are distinguished using both ictal (seizure) and inter-ictal (non-seizure) EEG recordings. A correct diagnosis is crucial to suitable long-term management of epilepsy. For instance, the generalised spike wave complexes are very responsive to certain AEDs and have a good prognosis, given that the correct treatment is used. In the case of using the wrong AED, although the seizure may be managed in short-term, but in long-term there lies the risk of the patient developing drug-resistant epilepsy.

Treatment

The EEG can be used in aid of assessing the effects of therapeutic treatment on the patient. It can also identify side effects of certain AEDs. The occurrence of certain EEG events is considered a risk factor for one patient, while indicating a positive impact of the AEDs on others. The EEG can also be used with patients who have previously undergone surgery, in order to predict the long-term outcome of the surgery as a prognostic tool.

3.2.2 Invasive EEG Recording

In most diagnostic and treatment-monitoring settings, the non-invasive EEG is used in the form of a scalp EEG. There are however, settings in which an invasive EEG serves to be more informative (Noachtar & Rémi 2009). One of these settings is pre-surgery evaluation for accurate localisation of the seizure-focus particularly in mesial temporal epilepsy. This information is used along with other imaging information to better evaluate the patient prior to surgery. This form of intervention is used only when non-invasive methods result in poor localisation or when the epileptic zone is too close to the eloquent cortex.

The scalp EEG is susceptible to low resolution of recordings due to muscle activity or ballistic movement, creating a poor signal-to-noise ratio. This means that the surface EEG may miss out on underlying epileptic patterns otherwise captured by the invasive EEG as depicted in Figure 3.2. Invasive EEG provides information at a higher resolution, although it does have a high risk of complications due to the invasive nature.

The invasive EEG electrodes are placed in one of four standard ways, the choice of which varies based on the purpose of the EEG:

Depth electrode: The electrodes are implanted into the brain using MRIs to accurately choose the location of the placement with minimum damage to the important nerves of the brain. These are widely used in temporal lobe where mesial frontal and parietal areas can be investigated with high Sensitivity.

Subdural strip or grid electrode: These electrodes are subdurally planted on the cortex. These are particularly useful for inferring cortical mapping due to the large connected area they cover. This method is particularly used in Neuro-informatics research for encoding and decoding neuronal firing (Dayan & Abbott 2005).

Epidural electrode: In this method, mushroom-head electrodes are placed on the dura via holes and are amongst the less invasive methods.

Forman Ovale electrode: This method is used as a lower risk alternative to depth electrodes with lower levels of complication. The placement of electrode can record mesial temporal lobe areas. They can also be used in other types of epilepsy syndromes for recording how the seizure spreads to the temporal lobe.

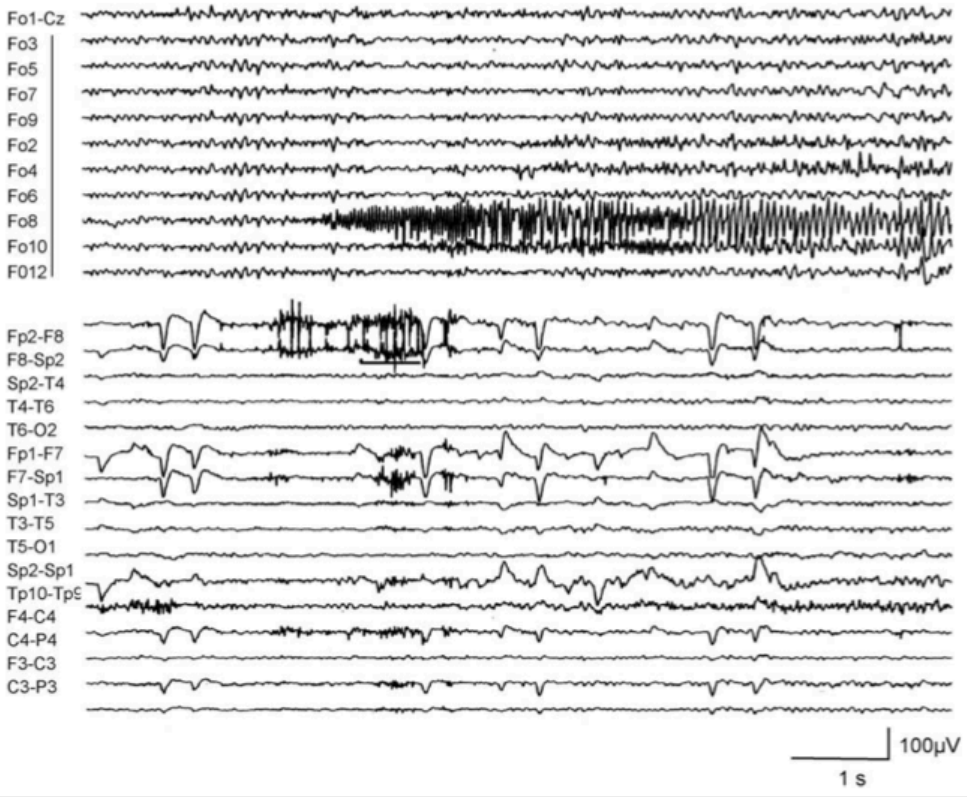


Figure 3.2 - EEG Recording with Invasive Foremen Ovale (top) and Scalp electrode (bottom) (Noachtar & Rémi 2009)

The invasive EEG methods are, as mentioned before, of higher risk to the patient (1-4% patients). Despite offering a more accurate resolution of the seizure, they are sensitive to picking up signals in a small radius of the area they are placed in. If the seizure zone is even millimeters away from the electrode, it is not picked upon, therefore providing little information on the accurate epileptic zone. This form of EEG provides enough information to support the interpretation of how the seizure is spread in the recorded area. Therefore, it is best to use surface EEG for multiple and longer recordings along with other imaging techniques and only resort to invasive methods if there exists an

underlying hypothesis about the epileptic zone and to accurately identify the location for surgical treatment.

3.3 Seizure Detection and Prediction

In recent years, there has been growing research interest on seizure detection and prediction from EEG recordings. The unpredictable nature of seizures imposes high risk on the health of epilepsy patients. Being able to predict seizures and couple this information with state of the art technology, will allow patients to take action prior to the occurrence of the seizure, minimizing the risk caused by seizures (Niedermeyer & da Silva 2005).

The main question researchers have been addressing is whether characteristic features can be extracted from an EEG, which have a correlation with the occurrence and time of the occurrence of seizures. In that case, treatments could move from therapeutic and long-term preventive plans to on-demand strategies (i.e. immediately before the seizure occurs). (Stein et al. 2000) have envisioned this using fast-acting anticonvulsant substance while (Theodore & Fisher 2004) have proposed deep-brain stimulation technology in order to reset the brain as soon as seizure activity is detected, to avoid the occurrence of seizures.

There is also the question of how a seizure occurs: Is it a result of a sudden transition or a gradual change in the dynamics of the EEG. The latter can be predicted through dynamics and is more likely the case for focal epilepsies, whereas the former is impossible to predict through dynamics and is more likely to be the case in general epilepsy (Da Silva et al. 2003).

Prior to the occurrence of seizures, a number of clinical symptoms have been proven to exist. These symptoms include an increase in oxygen availability, cerebral blood flow, and blood-oxygen-level-dependent signal and changes in heart rate (Baumgartner et al. 1998; Adelson et al. 1999; Federico et al. 2005; Kerem & Geva 2005). In addition to these changes, it is believed that the neuronal networks of the brain are involved in a process where an increasing number of critical interactions among the neurons in the focal region unfold over time. This concept has allowed researchers to study EEGs in an alternative way, in order to find such processes and identify the pre-ictal (pre-seizure) state.

Table 3.1 presents a summary of research in the field of seizure prediction and detection. The first attempts at seizure prediction were carried out by (Viglione & Walsh 1975) in order to find seizure precursors using linear approaches for absence seizure EEGs. Others (Rogowski et al. 1981; Salant et al. 1998) were able to find changes 6 seconds before seizure onset, using an autoregressive model of the neuronal activity. Another group (Siegel et al. 1982) found changes among 1-minute epochs prior to the seizure, and conducted further analysis on the spike occurrence rates in the EEG, indicating decreased focal spike-rate along with an increased rate of bilateral spikes before the seizure. This was followed by (Le Van Quyen et al. 1999) who compared pre-ictal dynamic variations to those of inter-ictal EEG and discovered a dynamical similarity index which seemed to decrease before seizures. Another groundbreaking discovery was made by (Iasemidis et al. 1990) using Lyapunov exponent and an open window analysis, revealing chaotic behaviour in invasive EEG and a decrease in this behaviour before the seizure.

Authors	Year	Features	Patients	Mean Prediction time (min)	Sen.	Spec
Liu et al	1992	Autocorrelation	13	-	42.9	90.2
Lehnertz and Elger	1998	Correlation dimension	16	12	94	0
Martinerie et al.	1998	Correlation density	11	3	89	-
Le Van Quyen et al.	1999	Similarity index	13	6	83	-
Le Van Quyen et al.	2000	Similarity index	9	4	94	-
Mormann et al	2000	Phase synchronization	2	-	100	100
Cerf et al.	2000	Lerner density	7	-	100	100
Iasemidis et al.	2001	Dynamical entrainment	5	49	91	-
Litt et al.	2001	Accumulated energy	5	19	90	88
Le Van Quyen et al.	2001	Phase synchronization	8	Several min	77	-
Lehnertz et al.	2001	Correlation dimension	59	19	47	100
Jerger et al.	2001	7 different measures	4	2	100	-

Authors	Year	Features	Patients	Mean Prediction time (min)	Sen.	Spec
Navarro et al.	2002	Similarity index	11	8	83	69
Celka & Colditz	2002	Signal complexity	13	-	66.1	56
Mormann et al.	2003	Synchronization/correlation	10	-	86	100
Khan and Gotman	2003	Frequency spectrum	13	-	62.5	64
Mormann et al.	2003	Phase synchronization	18	4-221	81	100
Niedehauser et al.	2003	Sign periodogram transf.	5	5-80 seconds	94	92
Chavez et al.	2003	Phase synchronization	2	>>30	-	-
D'Alessandro et al.	2003	Feature selection	4	3	63	72
Iasemidis et al.	2003	Dynamical entrainment	5	100	83	83
Winterhalder et al.	2003	Similarity index	21	-	42	85
Aschenbrenner et al.	2003	Correlation dimension	21	-	34	90
Shoeb et al.	2004	Multiple wavelet decomposition features	36	6	-	85
Maiwald et al.	2004	Accumulated energy	21	-	30	85
Gigola et al.	2004	Accumulated energy	4	-	92	100
Esteller et al.	2005	Accumulated energy	4	85	71	89
Harrison et al.	2005	Accumulated energy	5	-	0	-
Iasemidis et al.	2005	Dynamical entrainment	2	78	82	85
Jouny et al.	2005	Complexity/synchrony	2	-	0	-
Le Van Quyen et al.	2005	Phase synchronization	5	187	69	-
Mormann et al.	2005	30 different measures	5	-	-	-
Kalitzin et al.	2005	Phase clustering	3	-	-	-

Authors	Year	Features	Patients	Mean Prediction time (min)	Sen.	Spec
Navarro et al.	2005	Similarity index	13	>13	64	-
Chaovalitwongs et al.	2005	Dynamical entrainment	10	72	69	85
Harrison et al.	2005	Correlation dimension	20	-	0	-
Schelter et al.	2006	Phase synchronization	4	-	70	85
Costa et al.	2008	14 Features	2	5	98.5	99.5
Mirowski et al.	2009	Several bivariate features	21	5 s	71	100
Santaniello et al.	2011	High dimensional channel dependency features	4	9.6 s	100	84
Park et al	2011	Multiple Spectral Band Power features	18	-	97.5	73
Williamson	2012	High dimensional feature-set of channel correlations	19	828	95	85

Table 3.1 Summary of research in seizure prediction and detection. Adapted from (Mormann et al. 2007) with modifications.

(Litt et al. 2001) conducted a controlled experiment on continuous multi-day EEG recordings of a population of 5 patients evaluated for epilepsy surgery. The statistical study revealed that, quantitative signal changes were detected 7 hours, 2 hours and 50 minutes prior to the seizure onset, with an increase in accumulated energy 50 minutes prior to the seizure onset, suggesting that the cascade of electrophysiological events which have evolved from several hours before the seizure onset, can be identified as a reliable and timely indication of seizures. The optimistic findings of this study were, however, not reproducible in later studies (M. A. F. Harrison et al. 2005b). More so, the approach used in this study was statistical, bearing no confirmation for successful prospective implementation. Some other studies used an algorithmic approach on similar multi-day EEG recordings. (Iasemidis et al. 2005) reported 68% Sensitivity and 0.15 false postives rate (which is the same as 85% Specificity), for 78 minutes in advance, on a dataset of only 2 patients. The small population of patients used in the dataset lead to inconclusive results. The method used is an algorithmic real-time

statistical method which continuously calculates the single feature in use (the short-term maximum Lyapunov exponent) and monitors a T-index curves of this measure, and produces an alarm, if and when a the measure exceeds a threshold; it only considers a single feature. The same research group (Chaovalitwongse et al. 2005) reported a Sensitivity of 68% and the same false positive rate (specificity of 85%), for an average prediction window of 72 minutes in advance of seizures, when the same algorithm was tested on a population of 10 patients; the low sensitivity reported by (Chaovalitwongse et al. 2005) had a large standard deviation of 24.42%, indicating that the method performed poorly on 20% of the patients The results produced by these studies when tested on a larger dataset, were considerably low and are unsuitable for real-life implementation.

The diversity in the length of the prediction windows in the reported literature has lead to a further classification of the studies into seizure detection and seizure prediction. Seizure detection denotes the automatic recognition of seizures shortly before or after the actual onset, commonly in a short prediction window of a few seconds long. Seizure prediction represents the automatic recognition of seizures well in advance of the actual onset where the prediction window can be several minutes long (Mormann et al. 2007). The distinction between the two is important as i) the target application of each scenario, and hence, the potential treatment strategy appropriate for each method is likely to be different ii) the results reported for detection are generally higher than those reported for prediction, as for many cases, seizure prediction is used for exploratory purposes and to prove the existence of predictive markers well in advance of the seizure onset.

One other important distinction is between the underlying prediction methods of a study. There are two general approaches in seizure-prediction studies: statistical and algorithmic. In the statistical approach, distinct characteristics of the EEG are evaluated in a retrospective manner, for their capability to discriminate between known ictal and non-ictal states of the brain. These methods are mainly used for exploratory analysis of the seizure state and provides information such as peaks and drops of a measure (Mormann et al. 2005; Mormann et al. 2006). This renders it impractical for real-time application, where the ultimate goal is to label new unlabeled EEG data, as seizure or non-seizure. The algorithmic approach, labels every timepoint in the dataset as either ictal or non-ictal. In this thesis, we regard findings from the statistical studies as

background information about predictability of seizures and discriminatory power of features, and pay more attention to algorithmic studies for method and results comparison.

The following section presents some important studies in the field of automatic seizure detection:

3.3.1 Prominent Seizure Prediction and Detection Case Studies

Detection

Seizure detection in newborns is of significant importance, as there is not enough expertise in the newborn ward for detecting seizures and could therefore benefit from automated seizure-classification tools. (Faul et al. 2005) have evaluated the works of three published automated algorithms (Khan & Gotman 2003; Liu et al. 1992; Celka & Colditz 2002), for automatically detecting neonatal seizures.

The (Khan & Gotman 2003) algorithm is from a frequency point of view, in that, it uses Fast Fourier Transform (FTT) for finding rhythmic discharges, multiple spikes and very slow discharges. The problem with this method is that there is an overlap of frequency spectrum characteristics of ictal and non-ictal EEG.

(Liu et al. 1992) Also implemented a frequency based seizure detection algorithm, particularly searching for periodic, rhythmic data using autocorrelation, which is the cross-correlation of a signal with a delayed version of itself, used for finding repeating patterns in a signal. The approach is prone to producing high Specificity but low Sensitivity on neonatal seizure data, due to the simplicity of the autocorrelation function, which fails to detect rapidly changing frequency, amplitude and shapes.

(Celka & Colditz 2002) examines the complexity of EEG data to detect seizure signals, based on the knowledge that, ictal EEG is different in complexity from non-ictal EEG. This approach led to high Sensitivity and Specificity rates in the neonatal seizure detection implementation.

The studies were carried out on one-minute EEG segments of 13 neonates. The Specificity produced by (Khan & Gotman 2003; Liu et al. 1992; Celka & Colditz 2002) are respectively 64.0%, 90.2% and 56.0% and the Sensitivity was respectively 62.5%, 42.9% and 66.1%. In their review of these studies, (Faul et al. 2005) concluded that based on the Specificity and Sensitivity rates of these methods, none of them are fit to

be used for live clinical application, and despite providing useful information, they are not suitable for being used without other analysis and classifiers.

Prediction

(Costa et al. 2008) whose work is the starting point of this thesis, have tested various neural networks for classifying EEG records into one of the four classes: ictal, pre-ictal, inter-ictal, post-ictal. Ictal corresponds to the seizure activity, pre-ictal to the few seconds before the occurrence of the seizure, post-ictal corresponds to the EEG recordings immediately following the seizures and inter-ictal to the period between post-ictal and pre-ictal signals.

They used 14 features for classifying the EEG signals, which are based on signal energy attributes, wavelet transforms and non-linear system dynamics. They carried out their study on EEG recordings of two patients from the Freiburg EEG Database (Epilepsy.uni-freiburg.de 2007). The study compared the performance of Artificial Neural Networks trained in three different situations:

1) Trained on 70% of a single patient's recordings and tested on the remaining 30% of the recordings.
2) Trained on one patient's recordings and tested on another patient's recordings.
3) Trained on both patients' recordings and tested on either patient.

The following neural networks were used in this experiment:

Radial Basis Function - which means the number of neurons in the 1st layer is equal to the number of instances in the input.

Feed Forward Back-propagation and Layer-Recurrent Networks - composed by an arbitrary number of layers with a feedback loop around each layer, except for the output layer, where the feedback loop provides a single delay to the network. Networks were configured using 2 layers, the hidden layer composed of 10 neurons and the output layer by 4 linear neurons.

Elman and Distributed Time Delay - used with one hidden layer, composed of 10 neurons followed by a linear output layer, with a back propagation function, in conjunction with Distributed Time Delay networks which are dynamic Artificial Neural Networks where the output of the various layers depends on past output of these layers

Feed Forward Input Time-Delay Back-Propagation - using inputs from the training data and also a pre-defined time-delay from the data, meaning they can deal with temporal and spatial data.

This study reported experimental results in terms of test-set Accuracy, Specificity (the capacity of correctly identifying negative cases) and Sensitivity (the capacity of identifying positive cases), which we saw in section 2.3.6. The results reveal that 'single' experiment has the best mean performance (Sensitivity 98.5%, Specificity 99.5% and Accuracy 98.5%) followed by the 'multiple' experiments (Sensitivity 90.5%, Specificity 99% and Accuracy 97.5%), the least accurate being the 'different' case (Sensitivity 14.5%, Specificity 66.5% and Accuracy 54.5%). The results may suggest that individual-based EEG prediction is more effective than collective EEG training. However, this experiment was only run on 2 single patients, therefore the results are highly inconclusive. The study did not indicate sufficient experiment runs to make the findings statistically valid. Little specification was reported on the selected 14 features, as to which EEG channel they were extracted from.

(Shoeb et al. 2004) have developed a patient specific support-vector machine classifier, using wavelet decomposition, in order to detect patient-specific seizure onset. The detector passes 2-second epochs from each of 21 bipolar EEG channels through a feature extractor to compute features characterizing the morphology of each channel's waveform. The features extracted from each channel are grouped in one large feature vector in order to find correlation between those channels. The feature vector is then used with a support vector machine in order to be trained on ictal and non-ictal records. Seizure onset is detected when 3 consecutive 2-second epochs are classified as members of seizure class which helps avoid false detections.

The classifier was used on 36 subjects, recording 2 to 5 bipolar EEG recordings sampled at 256 Hz, each set of recording lasting for 35 minutes for 30 subjects, 2 hours for 4 subjects and 12 hours for 2 subjects. The first set of experiments used leave-one-out cross-validation testing for each subject and the classifier was trained on the recordings of all but one subject, for each subject, which produced a mean latency of 10 seconds. The inconsistent seizure of one patient resulted on poor performance of the algorithm and was prone to false positives when seizure like artifacts was longer than 6 seconds. True negatives were prone to inconsistency in spatial distribution of seizure activity.

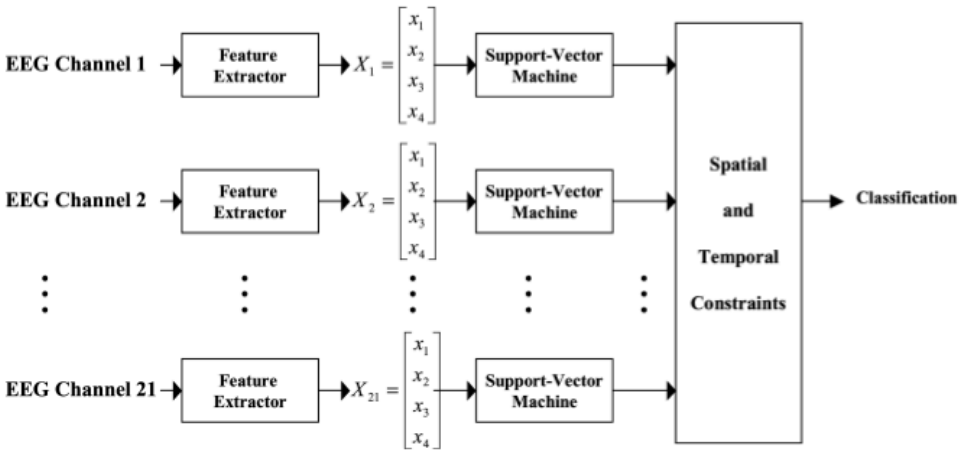


Figure 3.3 – Alternate SVM Architecture of (Shoeb et al. 2004)

The study then presents an alternative architecture, (Figure 3.3), where 2-second epochs from each of the 21 channels are passed through a wavelet transform extractor to compute features. The four features extracted are assembled into a distinct feature vector and assigned to seizure and non-ictal classes independently (no correlation taken into account), where seizure onset is detected, after the classification of all support vector machines, taking into account patient-specific localisation and temporal constraints by explicitly imposing patient-specific constraints at classification level. In comparison to the other architecture, this has a smaller mean prediction window and a larger number of true detections, which may be due to the small feature vectors, used for the support vector machines.

(Mirowski et al. 2009) This seizure detection study covers several bi-variate features with an emphasis on non-linear dynamics and synchronisation properties, namely, cross correlation (measure of linear dependence between two signals), non-linear interdependence (measures the distance between the trajectories of two EEG channels), dynamical entrainment (using the measure of EEG chaos and Lyapunov exponent) and three wavelet synchrony features (the difference between the frequency phase of two channels), alongside a number of machine learning algorithms, under several experimental settings. High dimensional features are calculated over several channels.

The features are pre-calculated across all EEG channels of a single patient. In an M channel setting, a single measure would create an $M \times (M-1)$ dimensional feature-set for each timepoint, i.e. all possible combination of EEG channels are used.

The extracted features are aggregated in several ways and used in all three classification methods, each yielding a total of 16 experiments per patient. The machine learning algorithms used are SVMs, Convolution Neural Networks and Logistic regression for a binary classification.

The results revealed that convolution networks on the wavelet coherence feature-set produced the highest classification measures, (75% Sensitivity averaged over all patients). They also were able to find at least one combination of features, feature aggregates and classification methods that could successfully classify all seizure data for each patient. After the preliminary step of extracting features from the dataset, the machine learning algorithms were trained and tested on inter-ictal and pre-ictal EEG, while excluding ictal and post-ictal data from the study.

The results also suggest that features extracted over a longer sliding window (5 minutes) proved to produce higher performance than those extracted over a shorter window (1 minute). This is one of the few studies where the prediction of the seizure was compared against the epilepsy syndrome of the patient and the localisation of the epileptogenic focus. In their findings, they observed no correlations between patient condition and the number of successful features and classifiers.

(Santaniello et al. 2011) A group at Baltimore University has applied a Hidden Markov Model to the multi-channel EEG data of 4 epileptic patients in an attempt to develop an optimal control-based Quickest-Detection (QD) strategy as well as higher Specificity rates. The QD strategy was used to predict the ictal state on a patient-specific basis via minimizing a cost function of detection delay and false positive probability.

They use a graphical representation of EEG channels in order to create a feature-set, which captures the co-dependences of any 2 channels. The channels are regarded as nodes and are connected to other nodes depending on their co dependence in a given frequency band. For any 2 channels, the connectivity matrix, which is calculated from the connectivity of the nodes in the graph are calculated over a 5 second moving window.

Singular value decomposition (SVD) of each matrix specifies the rank of the matrix, which is an indication of the inherent complexity of the matrix and therefore, the brain. The higher rank indicates higher complexity of the brain and higher probability of seizure state. The SVD is calculated for the connectivity matrix A:

$$A = USV^* = \begin{bmatrix} u_1 & u_2 & \dots & u_m \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix} \quad (1)$$

Where u_m are the eigenvectors of matrix AA^*, v_n are the eigenvectors of matrix A^*A and σ_r are the non-zero values of A for *r* ranks of *A*.

This method is based on the assumption that the SVD of the patient at each second is generated by the Hidden Markov Model (HMM), and that each state *k* is historically dependent on previous state transitions (this is usually not the case in the default HMM). The study yields 100% Sensitivity for all 4 patients on test and validation samples. The detection time-delay averaged over all patients is 9.6 seconds and standard deviation of 10.56. False positive values vary for each of the 4 patients, but averages to 1.39 per hour. The penalty for detection delay can be tweaked to values anywhere in the range [0, 1], in order to reduce the delay while maintaining high Sensitivity values and low false positive rates. There is however no evidence of a general rule for the penalty detection delay which yields the highest performance measures across all 4 patients.

The difference between this work and other research, apart from the use of time series analysis in the form of Hidden Markov Models, is the incorporation of the performance requirement in terms of a cost function, which is to be minimised. The modeling of the state transitions from ictal to non-ictal in the form of HMMs has allowed for an evolving and dynamic detector, which changes course and evolves based on performance measure of its current state.

The use of a minimizing cost function indicates a frequentist approach to building the HMM rather than alternative Bayesian approach. The dataset used in this study consisted of more recordings per patient (~42 hour and ~10 seizures per patient), which could in turn be a significant reason to the improved performance.

3.3.2 Seizure Prediction: More on Feature Engineering

Research in the field of seizure prediction, comprises two underlying trends for improving epileptic seizure detection algorithms, namely, improving

prediction/classification models and enhancing EEG features. The latter has a wider body of work attached to it, with studies drawing from signal processing and mathematical disciplines to engineer improved features. The former, however, has more recently been the playground of machine learning experts, instigating an exceeding number of studies on applying various algorithms to the seizure classification problem. The feature engineering literature is dominated by statistical approaches; the algorithmic studies, when used, tend to use a simple well-documented machine learning algorithm with new sets of features, amongst which, Artificial Neural Network has been the most popular. The machine learning studies on the other hand, usually use a simple feature such as the raw signal power, which does not require pre-engineering. In the literature presented thus far, the principal theme has revolved around the machine learning algorithm used in seizure prediction and detection experiments. Some of the most commonly used features in seizure prediction studies and their characteristics (Table 3.2) as well as outstanding feature engineering research papers are highlighted below.

Feature	Description	Linearity	Mode
Statistical moments	mean, variance, kurtosis and skewness of the signal time-series	Linear	Univariate
Spectral band power	relative power contained in 5 pre-determined spectral bands	Linear	Univariate
Spectral edge frequency	minimum frequency in which 50% of <40 Hz spectral power is contained	Linear	Univariate
Accumulated energy	The sum of the signal power of a sequence of windows	Linear	Univariate
Hjorth parameters	Three parameters of activity, mobility and complexity of an EEG which can be calculated from the statistical moments.	Linear	Univariate
Autoregressive modelling	Composed of three linear models of the signal; white noise, moving average and autoregressive process.	Linear	Univariate
Correlation dimension	The number of degrees of freedom of the probability density of signal in state space.	Non-Linear	Univariate
Correlation density	A correlation sum calculated by time-series delay and spatial embedding of EEG.	Non-Linear	Univariate
Correlation entropy	A measure which describes the level of uncertainty about the future state of the dynamical system.	Non-Linear	Univariate

Feature	Description	Linearity	Mode
Marginal predictability	A measure of signal predictability based on the correlation sum.	Non-Linear	Univariate
Dynamical similarity index	The similarity between a fixed window of reference and a running window of EEG.	Non-Linear	Univariate
State space dissimilarity measures	Measures the dissimilarity between two EEG time-series based on correlation sum.	Non-Linear	Univariate
Maximum Lyapunov exponent	Measures the divergence between trajectories in state space.	Non-Linear	Univariate
Local flow	A measure which identifies whether a dynamical state is stochastic or deterministic.	Non-Linear	Univariate
Algorithmic complexity	Based on symbolic dynamics, where a time-series of symbols are created and the complexity is measured via the size of the symbol sets.	Non-Linear	Univariate
Loss of recurrence	A quantification of degree of non-stationary in the EEG.	Non-Linear	Univariate
Maximum linear cross-correlation	Quantifies the similarity between two time-series.	Linear	Bivariate
Linear coherence	Measure the linear synchronisation between two signal time-series in a given frequency band.	Linear	Bivariate
Non-linear interdependence	A quantification of similarity between two signal time-series based on the re-construction of a state space.	Non-Linear	Bivariate
Dynamical entrainment	The statistical difference between the Maximum Lyapunov exponent of a sequence of windows of two time-series.	Non-Linear	Bivariate
Measures for phase synchronization	Composed of 3 measures of mean phase coherence, index based on conditional probability and index based on Shannon entropy.	Non-Linear	Bivariate

Table 3.2 The most common features used in seizure prediction literature.

(Williamson et al. 2012) engineered a new feature which comprised spatial and temporal information of all channel recordings. This was tested in an individual-patient mode using a support vector machine. They also used Principal Component Analysis (PCA) to minimise the feature-set dimension. The experiment was conducted on the patients from the Freiburg EEG Database using a 15-minute moving window analysis. The features are based on correlations across EEG channels, extracted from the space-delay covariance matrix of the EEG over four levels of time delays, resulting in a high dimensional feature-set. They use a linear SVM with a Radial Basis Function (RBF) kernel, with Sensitivity measured over three predictive threshold of t = -0.1, 0, 0.1 yielding Sensitivity of 0.95, 0.88, 0.86 respectively in total. Specificity on the other

hand was ignored. One of their important findings was a correlation between the subset of features selected using PCA and the improved Sensitivity, indicating the importance of the feature-reduction step in increasing the performance in terms of Sensitivity. They also discovered that the smallest time-delay was most contributive towards the better feature-set.

(Mormann et al. 2005) presents one of the more comprehensive, seizure detection feature studies, which exhaustively examines predictability of epileptic seizures using several features and experimental schemes. In this study, they constructed 30 features from the EEG recordings, containing bi-variate and multi-variate features as well as linear and non-linear ones. These features comprise signal variance, signal skewness, signal kurtosis, the relative power of the five common spectral bands, spectral edge frequency, Hjorth complexity, Hjorth mobility (Hjorth 1970), correlation dimension (Elger & Lehnertz 1998), maximum Lyapunov exponent (Iasemidis et al. 1990), local flow (Kaplan & Glass 1992), algorithmic complexity (Bai-Lin 1989), loss of recurrence (Rieke et al. 2002), surrogate correlation dimension, surrogate maximum Lyapunov exponent, surrogate local flow, surrogate algorithmic complexity, mean phase attributes (Mormann et al. 2000) and non-linear interdependence (Arnhold et al. 1999), mathematical definitions and details of which can be found in (Mormann et al. 2005). The features were evaluated individually, using a statistical approach, where amplitude of the inter-ictal and pre-ictal distributions of single features were analysed, checking for discriminatory power using the Receiver Operating-Characteristics (ROC) curve, which plots Sensitivity against Specificity. A moving window technique was used with a fixed number of data points, as opposed to a fixed window length, due to differences in sampling rates of the dataset.

Seizure prediction was evaluated under four experimental schemes, two of which are considered for this literature review: 1) All values from all EEG channels 2) EEG values from the best performing channel. The best channel was determined after the results of each channel were separately evaluated. The study suggests that linear methods perform as good as, if not better than non-linear features. Results also reveal that bi-variate features are sensitive to dynamical changes up to 1 hour before the seizure, while uni-variate measures appear to discriminate between seizure and non-seizure in a small time-window. Among the different schemes, 'all seizure-all channels' did not produce predictive outcome for neither of the features. For the case of 'all

seizures-one channel separately', with the notion of the best channel in mind, the bi-variate measures performed better than uni-variate methods.

The study claims that achieving a prospective seizure prediction with 100% Sensitivity and Specificity is unrealistic and combining bi-variate and uni-variate features to achieve a probability of seizure occurrence may be a more plausible scenario. They suggest that probabilistic seizure-anticipation for predictor implants will perform better than random but this may not be enough for clinical application.

They also found results to be consistent among the various tested patients but argue that the inter-ictal results which ultimately are the predictors in this study, may not be transferable to real-life applications, as the signals produced in the study were under particular pre-surgical circumstances. The study demonstrates a lack of consistent optimum locality for a feature and also concluded that bi-variate measure are not capable of predicting seizures in a shorter prediction window, therefore may not be solely suitable for predictive applications due to the long prediction window, but rather, are suitable for being used as an indication to the likelihood of a seizure happening in the long timescale along with uni-variate features.

This important paper statistically proves the existence of pre-ictal activity and claims seizures can be predicted in a large time window (240 minutes), given the correct combination of uni-variate and bi-variate features. The main shortcoming of this paper is in its statistical approach to seizure prediction, which as mentioned earlier, can merely lead to statistically validation of existence of predictive markers rather than predicting the state of unseen and new data points.

EPILAB (Teixeira et al. 2011) has developed a Matlab (Mathworks.co.uk 1994) platform which allows EEG seizure prediction using several features, and supports high dimensional epileptic seizure classification. They engineered uni-variate, bi-variate, linear and non-linear features from a dataset of EEG recordings, a full list of which is enlisted in (Teixeira et al. 2011). They suggest that for multiple channel real-time feature extraction, uni-variate linear methods are suitable and easily computed followed by bi-variate features over a limited number of channels and finally, uni-variate non-linear features are slow to compute compared to uni-variate linear features and cannot be computed in real-time for all channels at once.

In a recent study, (Park et al. 2011) used SVMs on 18 out of the 21 patients on the Freiburg EEG database. They produced high patient-specific detection results

(97.5% Sensitivity and 73% Specificity) on linear features of the Spectral Band Power, in what they refer to as a bipolar pre-processing setting, where unwanted artifacts are removed and ultimately results in a better spatial resolution. They used a binary SVM in which the cost-function takes the imbalance between the ictal and inter-ictal instances into account, during parameter selection. The level of Sensitivity is significantly high but is undermined by the lower level of Specificity. Other studies such as Costa et al. (Costa et al. 2008) have reported higher levels of Sensitivity and Specificity for the same patient-specific seizure detection experiments. However, the results do reveal the predictive power of Spectral Band Power measures.

3.3.3 Common Techniques in Seizure Prediction

Some of the common techniques and best practice used in characterizing EEG records are listed as follows (Mormann et al. 2007):

Moving window analysis

The performance of the algorithm can be assessed using what is called a moving window analysis. The moving window analysis requires that EEG recording are broken down into 10 - 40 second time frames, for each of which, several characteristics and features are measured. Moving window analysis can be uni-variate (i.e. characterizing a single-channel), bi-variate (finding relations between two channels) or multi-variate (finding relations between more than two channels).

Statistical versus algorithmic approaches

EEG recording time profiles are analyzed in either of two approaches: the statistical approach, which is retrospective of the distribution of some characteristics of the inter-ictal state against those of the predicted pre-ictal state, which is potentially useful for comparing the state of certain characteristics under different conditions; algorithmic approaches on the other hand, produce an output which is the function of the information given at every point of a time profile.

With prediction algorithms, it is good practice to define a prediction horizon, which defines a period of time after an alarm within which a seizure is expected. The definition of a prediction horizon allows for taking count of false positives (when an alarm is not followed by a seizure within this time frame) and true positives (when an

alarm is followed by a seizure within the horizon).

Sensitivity and Specificity

Sensitivity and Specificity (see section 2.3.5) of an algorithm can be defined using the predefined prediction horizon: Sensitivity is quantified as the number of seizures with at least one alarm in the preceding prediction horizon divided by the total number of seizures, and Specificity is the portion of time from inter-ictal period during which the patient is not in the state of falsely awaiting a seizure. Depending on the clinical requirements, an algorithm can be tuned to have higher Sensitivity but lower Specificity or vice versa.

3.3.4 Seizure Prediction and Related Issues and challenges

In this section we present a number of the reported issues within the seizure prediction line of research:

Poor use of Machine learning

When machine learning algorithms made their debut in the field of EEG seizure prediction, they were merely used as feature selection tools to narrow down the pre-engineered features and channels that yielded the highest in-sample performance. The community had yet to use machine learning for the actual task of classification. With papers such as (Shoeb et al. 2004; Costa et al. 2008; Santaniello et al. 2011) more sophisticated machine learning methods were used as classifiers for seizure prediction tasks, particularly Artificial Neural Networks. In the majority of the machine learning enhanced studies, a single EEG feature is used for the classification task in order to showcase the power of the particular machine learning algorithm in use. There is little work that combines the powerful features developed in early, seizure prediction research with state of the art machine learning algorithms.

Out-of- sample algorithms

The main fraction of research in EEG seizure prediction involves using simple tuning of well-know binary classification of in-sample data (Mormann et al. 2007). This entails that the classifier has not been verified on unseen data and is likely to perform poorly on out-of-sample predictions. In order for the findings of a seizure detection study be

extended to other datasets, the performance of the algorithm should only be published for the test data, which is essentially required to be unseen 'out-of-sample' data (Mormann et al. 2007).

Confounding variables

One other point worth consideration is the confounding variables during the inter-ictal state, which may influence the characterizing features used for the prediction algorithm. Failing to identify and understand such variables could potentially affect the Specificity and Sensitivity of the algorithm. Therefore, building the algorithm should involve features from both the ictal and inter-ictal stage and work towards a better understanding of confounding variables which may be a result of slow-wave sleep, emotional and cognitive states (Mormann et al. 2007).

Mechanisms of Ictogenesis

Another issue that has been somewhat overlooked in most studies, is the process and mechanism of seizure generation itself (Ictogenesis). Some studies have found the mechanism behind certain types of epilepsy (Kalitzin et al. 2002) and some suggest that there may be different seizure generation mechanisms for different brain structures and pathologies, implying that seizure initiation could vary from person to person. Therefore, using EEG prediction algorithms in understanding these mechanisms, and also using these mechanisms to develop better prediction algorithms could possibly result in better research outcome. Some studies have suggested modeling EEG signals in order to have an insight into the dynamical process of seizure-generation through time (Wendling et al. 2003; Suffczynski et al. 2006).

Seizure prevention

One potentially groundbreaking area of research that is often disregarded, is designing intervention systems, which in addition to warning the patient about a seizure, will also prevent this from happening. (Stein et al. 2000) have looked into the local application of short-acting powerful drugs. In another study, electrical stimulations have been suggested, with major focus on deep brain stimulation intervention (Theodore & Fisher 2004) which in a nutshell, uses electrodes to alter the state of the brain from the ictal state. These forms of intervention could benefit from seizure prediction, but also, from

early seizure detection. Seizure prediction, predicts the time at which the seizure could occur, well in advance of EEG ictal state, whereas early seizure-detection, focuses on detecting the seizure onset before the clinical symptoms occur with little time for intervention. The research in this area is very recent and further studies should be carried out in order to investigate potential real-life applicability of such concepts.

Surface EEG vs. Invasive EEG

When it comes to using EEG recordings for experiments, much care has to be taken as to which type of recording is being used. Intracranial recordings are used in the majority of seizure-prediction studies and provide much better signal to noise ratio and less artifact, with the added benefit of being recorded directly from the seizure-generating area of the brain, whereas the surface EEG can provide an overall image of all areas of the brain, which is useful for understanding the effects of the environment on seizure generation. However, if closed-loop interventions were to be used, patients would have to wear the EEG cap the entire time in order to monitor their surface recordings. Therefore there are doubts regarding the usefulness of scalp EEGs for intervention purposes (Morrell 2006).

Data requirements

In order to have a reasonable separation between inter-ictal and pre-ictal stages, it is advised to use EEG recordings, which not only have large number of seizures but also have sufficient time interval between the seizures (Mormann et al. 2007).

3.4 Summary

In this chapter, a rich and comprehensive background on epilepsy and an introduction to EEG and its role in various areas of epilepsy diagnosis and treatment were presented. This information is crucial in understanding the problem of seizure prediction and therefore a great segment of this chapter was dedicated to the description of these two topics.

The development of seizure prediction research in the early days up until the state of the art was also presented along with relevant research literature. Finally, common techniques and challenges in the field of seizure prediction were identified.

Chapter 4

The Freiburg EEG Database and EEG Feature Extraction

The first section of this chapter presents and summarises the Freiburg EEG database which is the data source used throughout this thesis. The second section of this chapter discusses the data preparation steps carried out prior to the implementation phase of the experiments in this thesis.

4.1 The Freiburg EEG Database

The Freiburg EEG Database is one of the most cited resources used in prediction detection experiments. It is also one of the few publicly available invasive EEG (see section 3.2.2) datasets. The database contains 24 hour-long continuous pre-surgical invasive EEG recordings of 21 patients suffering from epilepsy. The patients are from a wide range of varying age, sex, seizure type and seizure locality, but they all suffer from focal medically intractable epilepsy and were admitted for pre-surgical evaluation (see section 3.1.3) at the Epilepsy Centre of the University Hospital of Freiburg, Germany (Epilepsy.uni-freiburg.de 2007)

The epileptic foci of each patient vary from a range of neocortical brain structure (11 patients), hippocampus (8 patients) or both (2 patients). A summary of patient characteristics is listed in Table 4.1.

Pat.	Sex	Age	Seizure type	H/ NC	Origin	Electrodes	# Seiz.	Ictal Dur.
1	F	15	SP CP	NC	Frontal	g,s	4	86400 s
2	M	38	SP CP GTC	H	Temporal	d	3	86400 s
3	M	14	SP CP	NC	Frontal	g,s	5	86400 s
4	F	26	SP CP GTC	H	Temporal	d,g,s	5	86400 s

Pat.	Sex	Age	Seizure type	H/ NC	Origin	Electrodes	# Seiz.	Ictal Dur.
5	F	16	SP CP GTC	NC	Frontal	g,s	5	86400 s
6	F	31	CP GTC	H	Temporal/ Occipital	d,g,s	3	86400 s
7	F	42	SP CP GTC	H	Temporal	d	3	88597 s
8	F	32	SP CP	NC	Frontal	g,s	2	86979 s
9	M	44	CP GTC	NC	Temporal/ Occipital	g,s	5	86163 s
10	m	47	SP CP GTC	H	Temporal	d	5	88047 s
11	F	10	SP CP GTC	NC	Parietal	g,s	4	86570 s
12	F	42	SP CP GTC	H	Temporal	d,g,s	4	89326 s
13	F	22	SP CP GTC	H	Temporal/ Occipital	d,s	2	86400 s
14	F	41	CP GTC	H, NC	Frontal/Te mporal	d,s	4	85894 s
15	M	31	SP CP GTC	H, NC	Temporal	d,s	4	86400 s
16	F	50	SP CP GTC	H	Temporal	d,s	5	86400 s
17	M	28	SP CP GTC	NC	Temporal	s	5	86634 s
18	F	25	SP CP	NC	Frontal	s	5	89569 s
19	F	28	SP CP GTC	NC	Frontal	s	4	87780 s
20	M	33	SP CP GTC	NC	Temporal/ Parietal	d,g,s	5	92219 s
21	M	13	SP CP	NC	Temporal	g,s	5	86177 s

Table 4.1 Characteristics of Patients of The Freiburg Invasive EEG Database (Epilepsy.uni-freiburg.de 2007).

The invasive EEG records are captured with three types of grid (g), strip (s) and depth (d) electrodes (see section 3.2.2). These data were recorded at 256 Hz sampling rate, using the Neurofile NT digital EEG over 128 channels. From these electrodes, 6 channels of data are extracted by visual analysis of EEG experts, 3 of which are in the epilepsy focal area and the remainder are from extra-focal area of the brain, the distinction of which was previously made in chapter 3 (see section 3.2.1). The channels are labelled from 1 – 6, where channels 1 – 3 correspond to focal recordings and 4 – 6 comprise extra-focal recordings. The locations of these channels vary for each patient.

The seizure foci vary for each patient but are typically one of the frontal (originating in the frontal lobe of the brain), temporal (starting in the temporal lobe of the brain), frontal/temporal (originating from both frontal and temporal lobes), temporal/parietal (localised in both temporal and parietal lobes) or temporal/occipital (originating from both temporal and occipital lobes) in origin. The seizure type also varies amongst patients and is one of generalised tonic clonic (GTC), complex partial (CP) and simple partial (SP), which were previously described in chapter 3 (see section 3.1.1).

There are two types of signal files per patient: Ictal and Inter-ictal. The ictal files contain at least 8616 seconds of EEG signals per patient. There are 3 ictal files on average per patient; each contains signals of a single seizure. The ictal file typically holds ictal signals (signals corresponding to seizure as seen in section 3.1.2) as well as pre-ictal (signals immediately preceding seizures), post-ictal (signals immediately following seizure activity) and inter-ictal (in between seizures) activities.

The patient is monitored for over 24 hours during which time several seizures are triggered and recorded. Windows of 8616 seconds are extracted and annotated as ictal files and the remaining bulk is classified as inter-ictal.

The data files are in ASCII format and contain signal voltage of the corresponding EEG segment. Each of the 6 channels and each of the ictal/inter-ictal segments are categorised in a separate ASCII file. The dataset comes with information on electrode specifications of 6 channels and seizure onset and offset markers for all patients.

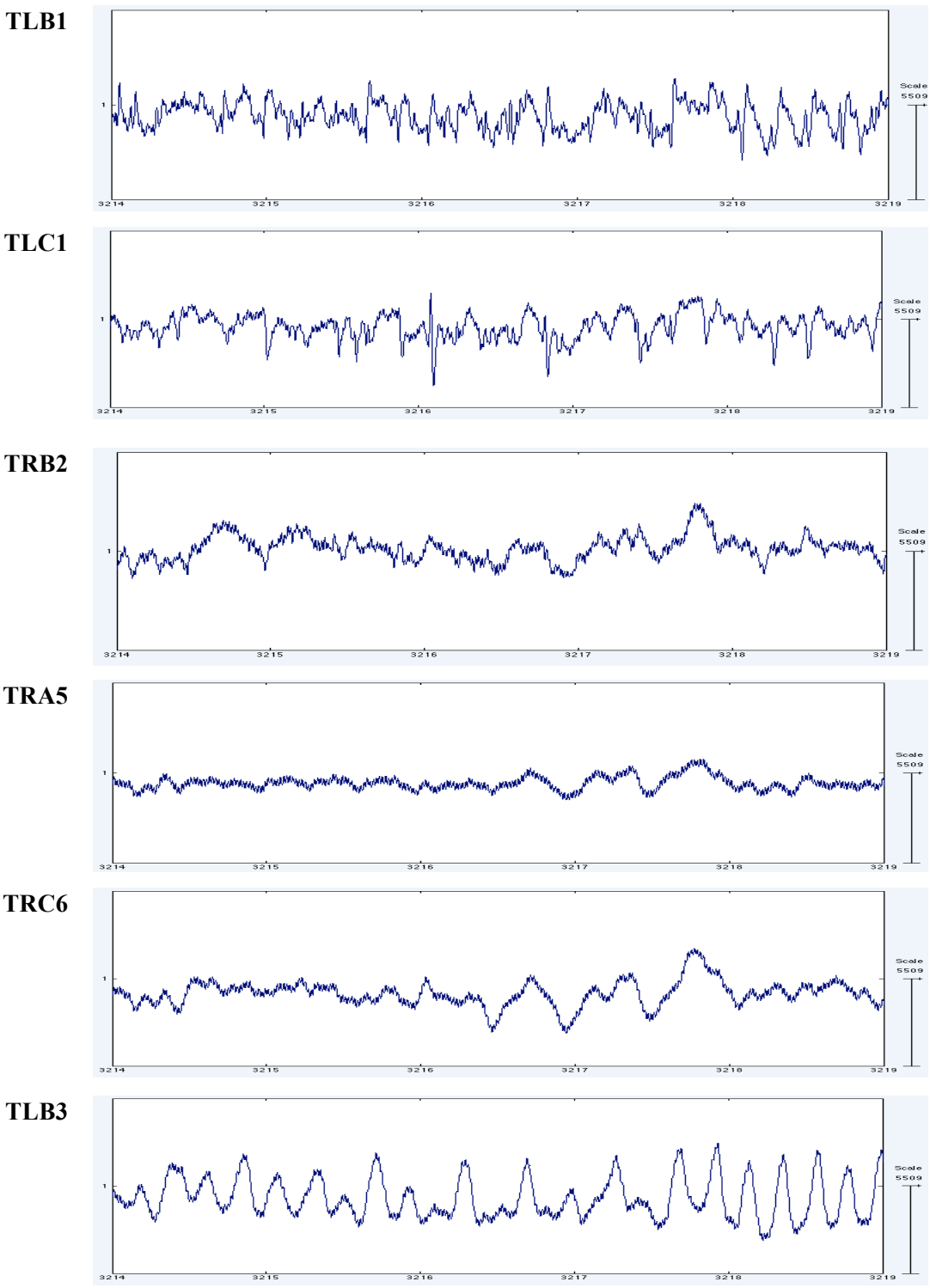


Figure 4.1 Invasive EEG Recording of Patient 2 from the Freiburg EEG Database – The image corresponds to pre-seizure and seizure data. Each row displays 1 of the 6 channel recordings. The name of the relevant EEG channel is listed to the right of each signal.

4.2 Data Preparation

In this section, we present the steps taken for preparing the raw data files from the Freiburg EEG database in order to be used in experiments described in future chapters. We start by describing how we deal with missing data and outliers for each patient. We then continue the section with details of the structure of the method of data extraction from the raw dataset which will be used to generate the original feature-set for 21 patients and will further be used as a framework for future feature engineering in upcoming chapters.

4.2.1 Data Sampling

In the Freiburg EEG Database, there are two types of ictal and inter-ictal files captured across 6 different channels. We originally select ictal files (which contain both ictal and non-ictal signal data) over 1 focal channel, which is typically channel 1; channel 1 refers to different locations of the brain for each patient. The data for each patient was recorded at 256Hz as was not down sampled for this study. The inter-ictal files only include non-seizure data and hence may not contain traces of seizure activity. More so, using the full 24-hour dataset for all 21 patients will lead to increased computational cost.

The runtime complexity of an SVM is $O(\max(n, d) \min(n, d)^2)$ (Chapelle 2007). The 24-hour recording comprises 17280 instances with traces of seizure data, while an ictal file on average comprises 2160 instances. The run-time complexity of training an SVM on the 24-hour recordings is 8^3 times that of the ictal files; the 24-hour files are therefore omitted from the original dataset. Instead, we only include the ictal files which comprise both seizure and non-seizure data leading up to and following the seizure, summing up to 1 hour for most seizures. The sampling of consecutive 1-hour segments of all states of ictal activity allows for faster processing time for our experiments without compromising the representation of seizure-state continuity.

4.2.2 Missing Data and Outliers

The common practice (Hand and Mannila et al. 2001) for handling outliers in a dataset is to represent them in a way, which makes them usable for the learning model, causing

the least structural damage to the dataset while taking extreme care when replacing outliers with numerical values in order to avoid introducing bias.

Artifacts in EEG recordings are forms of outliers and are considered as disturbances in a measured brain-signal, not originating from the brain. The different sources of artifacts are classified to external and internal categories. External artifacts result often from unsatisfactory technology such as exceeding measurement range of signals and disconnection of the electrode box. Internal artifacts arise from body activities that are either due to movements or bioelectrical potentials. The potential between electrodes changes as a result, from effects such as eye movement or muscular activity, causing an artifact (Mormann et al. 2005).

The common outlier correction approach does not apply to EEG artifacts. When dealing with EEG data, the common practice of dealing with artifacts (Mormann et al. 2005) is by visually detecting and clipping out of the outliers. We manually removed the artifacts for recordings of all 21 patients of the Freiburg EEG Database, according to pre-determined artifact specification accompanying each patient profile, using EEGLAB Matlab software package (Sccn.ucsd.edu 2011).

4.2.3 Ictal file extraction

The EEG data for each patient in the Freiburg database is organised into separate ASCII files, each file containing 1 hour long recordings of a single channel. As mentioned in 4.2.2 only recordings of channel 1 are analysed for each patient. Each ASCII file contains a combination of ictal and non-ictal data. All ASCII files of each patient in their raw format were transformed to the Costa et al. feature-set (see chapter 2), resulting in a $n \times 14$ dimensional Patient-File per patient. Figure 4.2 presents the 14 calculated features of a 1hour long ictal segment of patient 2 against the actual EEG recording.

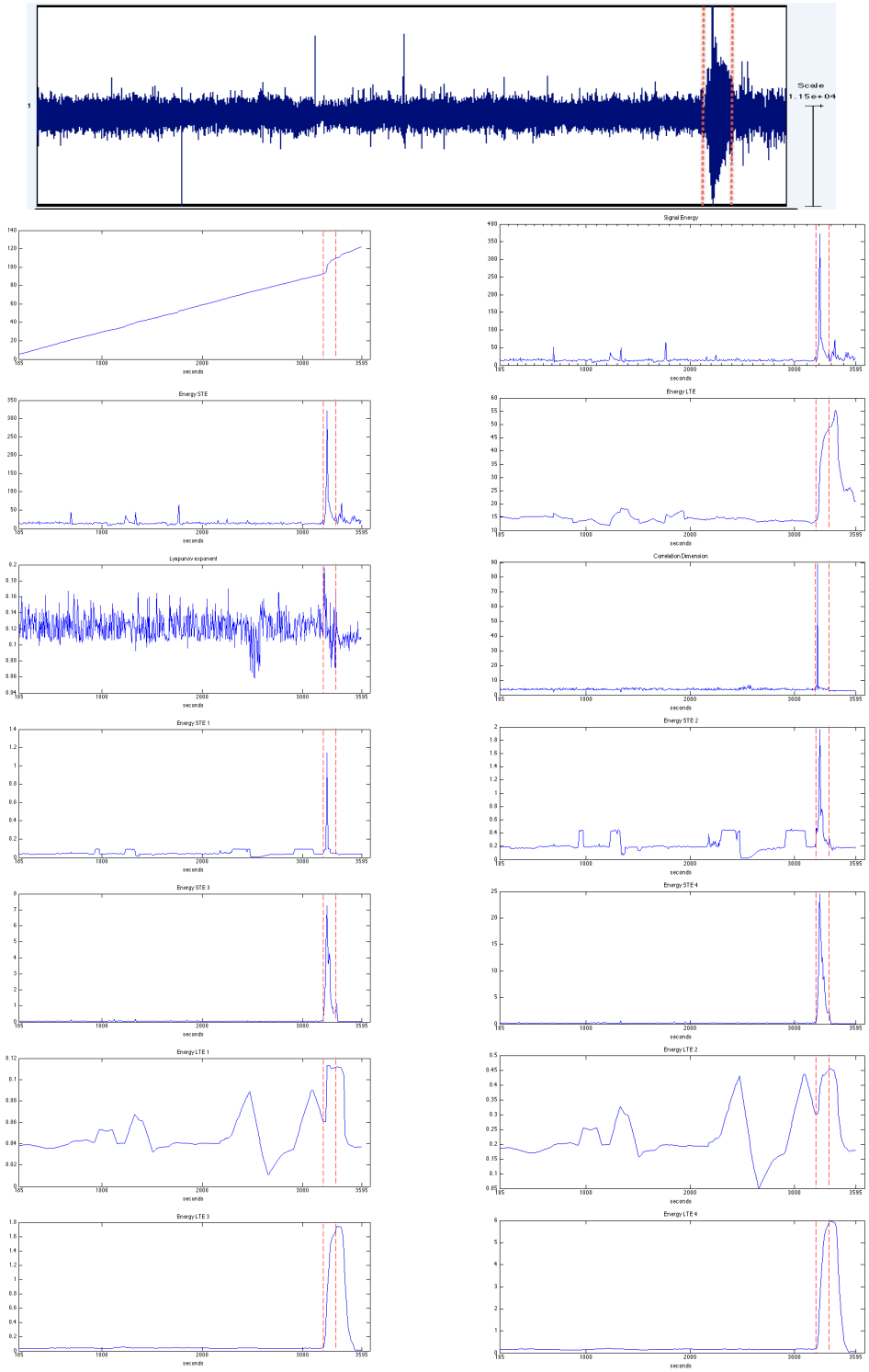


Figure 4.2 The raw EEG signal and the 14 features of a single ictal file of patient 2, derived from the first focal channel. The section in enclosed in between two red dotted lines indicates the seizure event.

Data from all seizures of a patient and the respective inter-ictal sequences were grouped together in a single file. Each ictal event and its surrounding inter-ictal data are 1 hour

long and the composition of these sets are based on the original composition of the respective ASCII files. The grouping of the ASCII files into a single Patient-File is performed after the feature engineering step, entailing that the moving window process described in section 2.2 is reset for each 1-hour segment. Figure 4.3 displays the accumulated energy of Patient-File 2. The single patient file contains three 1hour segments of ictal data. The accumulation of signal energy is reset at the start of each 1hour segment and is calculated independently of other ictal files of the same patient. Table 4.2 presents a summary of the dataset of all Patient-Files and Table 4.3 summarises the statistics of the 14 calculated features of Patient-File 2.

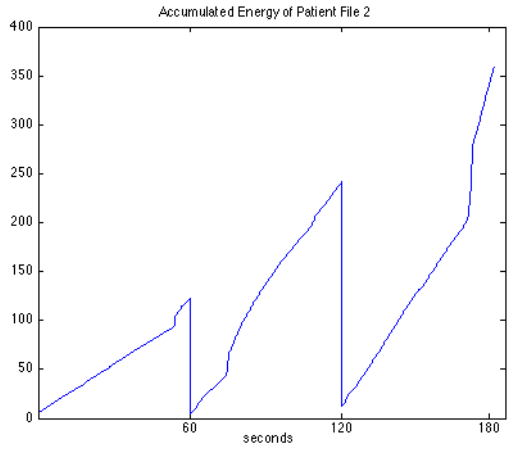


Figure 4.3 The accumulated energy of Patient-File 2 which contains three 1hour segments of ictal data.

Patient	#instances	#seizures	Patient	#instances	#seizures
1	2732	4	12	2732	4
2	2049	3	13	1366	2
3	2831	5	14	2732	4
4	3415	5	15	2732	4
5	3415	5	16	3190	5
6	2049	3	17	3415	5
7	1998	3	18	3394	5
8	1060	2	19	2732	4
9	3415	5	20	3415	5
10	3148	5	21	3415	5
11	2049	4			

Table 4.2 The data specification of the constructed Patient-Files of all 21 patients from the Freiburg EEG Database.

The EEG datasets are accompanied with markers of seizure start and end times, suggesting the natural formation of two rather obvious classes: ictal and non-ictal. In Figure 4.4 The distribution of each of the 14 features for Patient-File 2, which contains three 1-hour segments, is displayed.

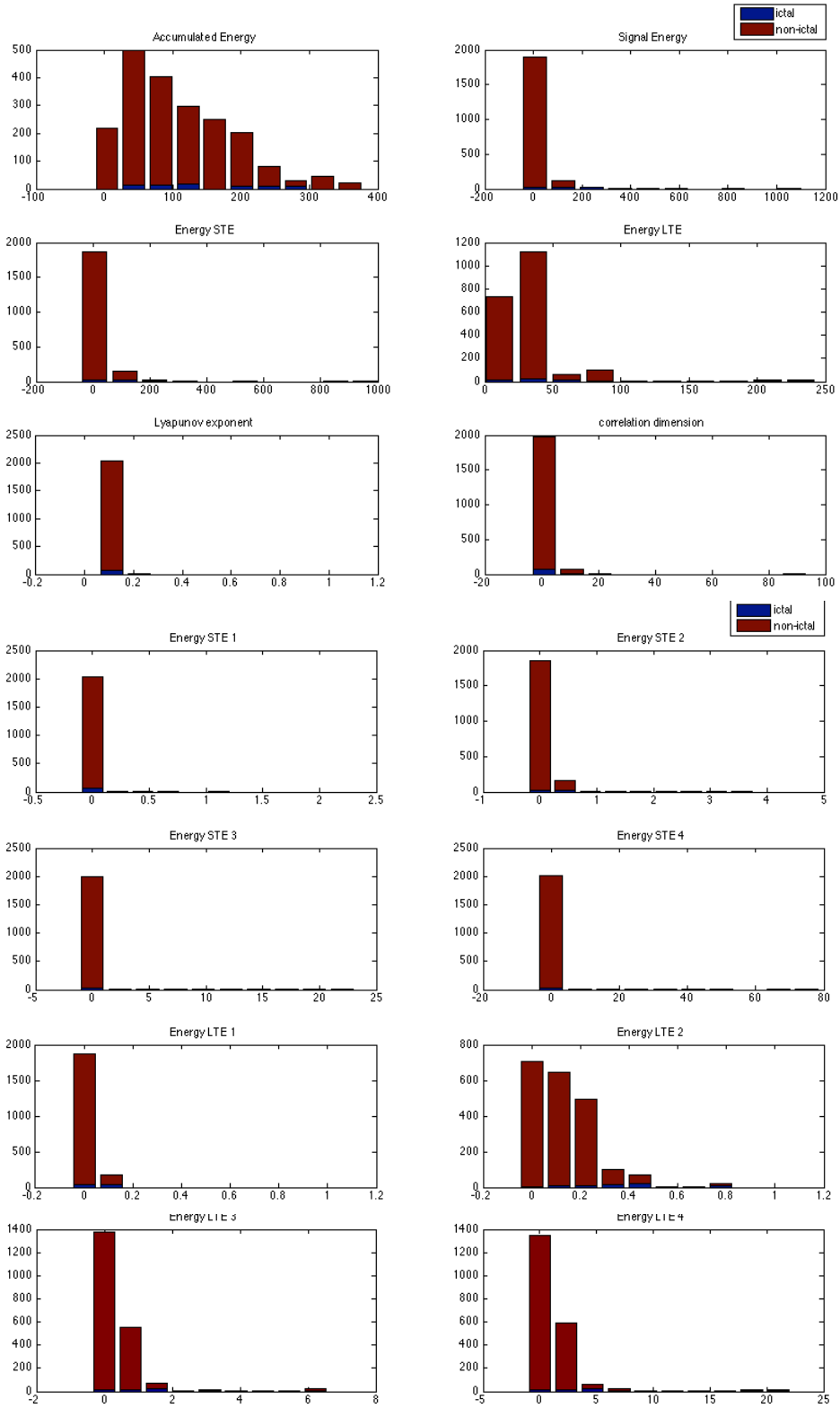


Figure 4.4 The ictal/non-ictal class distribution of all 14 features of Patient-File 2.

Feature	Min.	Max.	Mean	Stdev
accumulated energy	4.848	358.148	108.820	78.702
Signal energy	6.947	1052.623	34.183	47.527
energy STE	6.974	956.240	34.188	46.243
energy LTE	11.917	231.922	33.466	29.406
Lyapunov exponents	0.058	0.196	0.105	0.019
correlation dimension	1.981	88.976	4.386	2.098
energy STE 1	0.001	1.145	0.024	0.041
energy STE 2	0.007	3.375	0.131	0.194
energy STE 3	0.004	21.489	0.364	1.317
energy STE 4	0.021	74.201	1.376	4.323
energy LTE level 1	0.003	0.113	0.023	0.023
energy LTE level 2	0.013	0.751	0.132	0.123
energy LTE level 3	0.009	6.058	0.367	0.757
energy LTE level 4	0.049	20.279	1.383	2.534

Table 4.3 The statistical properties of all 14 engineered features of patient 2.

4.2.4 Data Labels

Following the feature extraction phase, each data point was labeled as one of the following states:

Ictal – This labels the seizure activity in the brain and is marked precisely by EEG experts. It is of varying length but is typically around ~3 minutes long.

Pre-ictal – Pre-ictal is marked as 5 minutes immediately prior to the seizure onset and is believed to hold predictive markers of seizure activity (De Clercq, Lemmerling, Van Huffel & Van Paesschen 2003b; Martinerie et al. 1998).

Post-Ictal – Post-Ictal is marked as brain activity following the seizure offset for duration of 5 minutes. Abnormal excitement in the signals may be observed in this state, particularly as patients are recovering from the seizures.

Inter-Ictal – non-seizure data preceding the pre-ictal state and proceeding the post-ictal state are marked as inter-ictal, where studies have traced early predictors of future seizure activity (Le Van Quyen et al. 2001; Quyen et al. 2003; Chávez et al. 2003). These labels were manually set according to the seizure onset and end markers provided in the Freiburg EEG Database notes. We use 4 states rather than 2 states as practiced in (Costa et al. 2008), in order to distinguish the activity among the four states and to ensure the data are more evenly balanced and to avoid dominance of the inter-ictal states over ictal activity. More so, in order to achieve prediction rather than detection, the definition of a pre-ictal window is vital (see chapter 3).

Although some seizure prediction studies that have defined a pre-ictal window use a two-class definition, but this practice rather is misleading since there is no indication in these studies as to what is considered pre-ictal and what is considered ictal; also, these studies fail to specify the utilisation of data instances, which are neither ictal, nor pre-ictal. Therefore, by defining 4 states, we utilise all the data instances in our learning algorithm. Figure 4.5 left displays the class distribution for binary classes of ictal and non-ictal for patient 2 and Figure 4.5 right displays the class distribution under the new multi-class scenario.

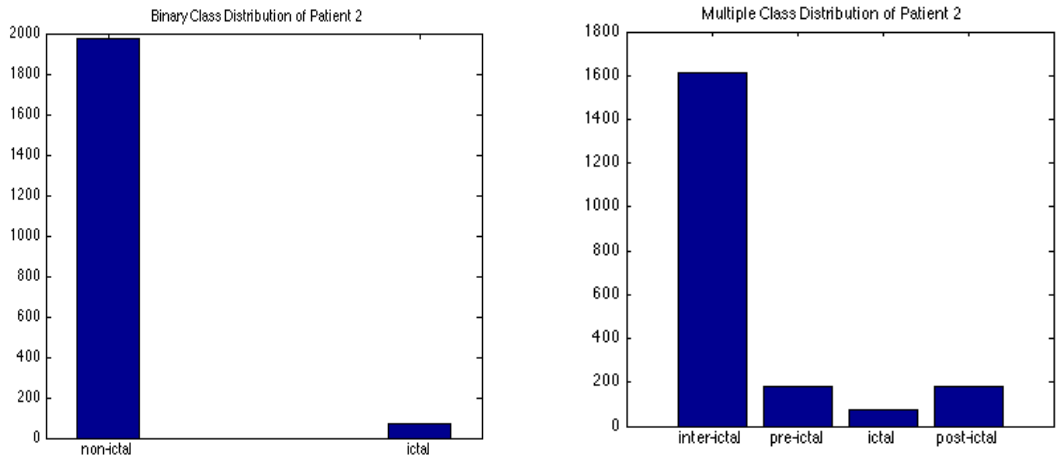


Figure 4.5 The class distribution of Patient-File 2: Left) the binary class distribution right) the multi-class distribution

4.3 Summary

This chapter presented the core data source used throughout this thesis, as well as data preparation steps undertaken in advance of algorithmic and machine learning implementation of experiments in future chapters. The described steps also form a basic framework according to which, other variations of data preparation and feature engineering presented in future chapters will be carried out.

Chapter 5
Preliminary Studies

Epilepsy is the most common neurological disorder, affecting between 0.6% and 0.8% of the global population. During an epileptic seizure, the onset of which tends to be sudden and without prior warning, sufferers are highly vulnerable to injury, and methods that might accurately predict seizure episodes in advance are clearly of value. Building on recent work by Costa et al. (Costa et al. 2008), we compare and contrast the Sensitivity, Specificity and Accuracy of a selection of algorithms that attempt to predict the onset of epileptic seizures on the basis of 14 features extracted from EEG monitoring data. We focus on how predictability varies as a function of how far in advance we are trying to predict the seizure episode, and also consider feature selection issues.

In section 5.1 of this chapter we further motivate the necessity of Automatic Seizure Prediction in general. In section 5.2 we summaries background and recent work presented in chapter 3, focusing on seizure prediction from EEG data, and finish with a summary of the results from (Costa et al. 2008). Section 5.3 presents the machine learning methods used in the experiments of this chapter, giving further detail on the dataset we use, the extracted features, and the machine learning and feature selection algorithms that we test. Section 5.4 presents our feature selection experiments and section 5.5 describes our advance-prediction experiments. In section 5.6 we discuss the findings of these experiments and in section 5.7 we summaries our findings and conclude.

This chapter was published as a paper in the proceedings of the Third World Congress on Nature and Biologically Inspired Computing (Moghim & Corne 2011).

5.1 Motivation

Epilepsy is the second most common neurological disorder affecting 0.6-0.8% of the population of the world (Mormann et al. 2006). In this chronic neurological, disorder abnormal activity of the brain causes seizures (Perucca et al. 1998).

Generally speaking, the cause of epilepsy is unknown for about half of all patients (Kwan & Brodie 2000). Epilepsy treatment requires long-term management and, being one of the most common chronic neurological diseases, gains a lot of attention from disease researchers (Aldenkamp et al. 2003). In the current work, our main focus is on the brain seizures that are endured by epileptic patients. Given the increasing availability of electroencephalograph (EEG) recording equipment, a growing body of work is investigating the EEG traces in association with epileptic seizures. Among other aspects of interest, one possibility raised by this technology is the advance prediction of seizures, which may provide ways to warn sufferers (and their carers) long enough in advance of a seizure for preventative measures to be taken, thus avoiding injury.

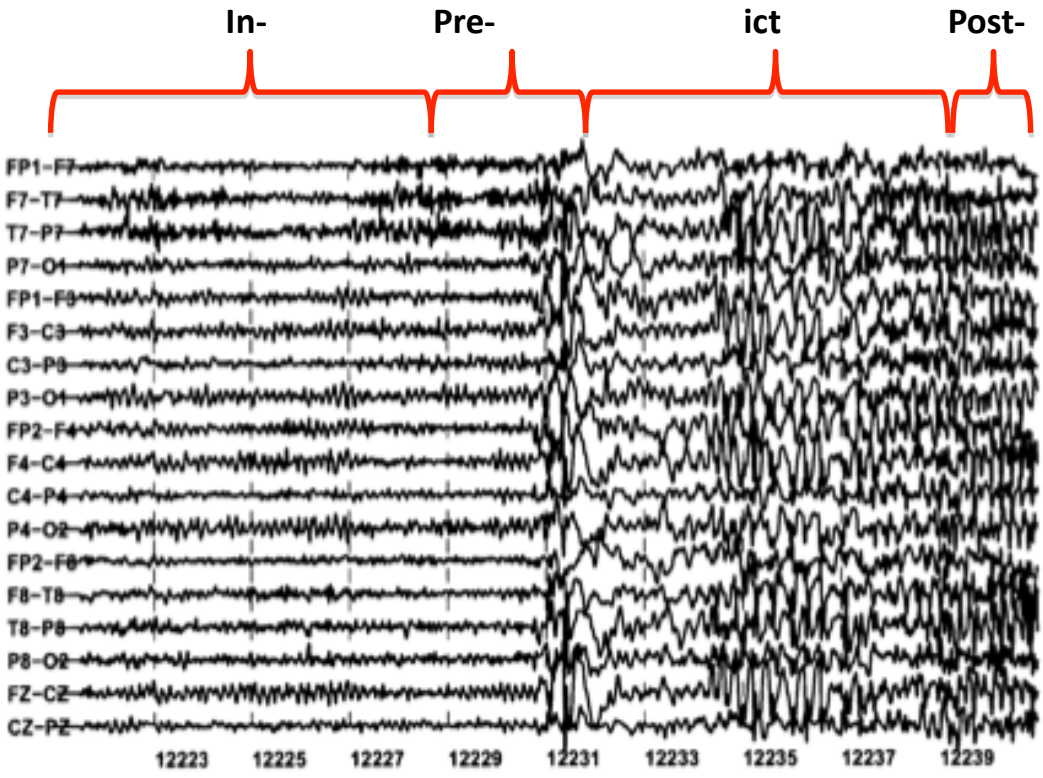


Figure 5.1 Illustration of the four classes of interest where *ictal* indicates the epileptic seizure. The plot shows EEG signal over time for an epilepsy patient from each of 18 electrodes.

We contribute to the enterprise of seizure prediction by building on recent work of Costa et al. (Costa et al. 2008). In Costa et al's study, 14 features were extracted from recorded EEG signals, and the EEG recordings were thus preprocessed into a time series of vectors. Each such vector summarised signal characteristics over a 5-second period, and was classified (with the aid of clinical expertise) into one of four classes,

specifically *inter-ictal* (normal brain activity), *ictal* (epileptic seizure), and *pre-ictal* (around 5 minutes of activity immediately preceding an *ictal* phase), and *post-ictal* (around 5 minutes of activity immediately following an *ictal* phase). Figure 5.1 illustrates these classifications against a plot from 18 channels of raw EEG data.

Classification of *pre-ictal* vectors corresponds to advance prediction of a seizure, and this is our main measure of interest. In (Costa et al. 2008), a variety of neural network based approaches were tested, and evaluated according to their Sensitivity and Specificity in regard to predictions of *pre-ictal* vectors in an unseen test-set. Costa et al. found that RBF Network and Recurrent Network were generally the most robust in performance, and we summarise their main results in section 5.2. Here, we compare Costaet al' s results with a Multi-Class Support Vector Machine (MCSVM), and Evolved Neural Networks, (EANN) applied to the same variations on the prediction problem that were studied in (Costa et al. 2008). Moreover, we investigate the performance of the SVM and EANN when attempting to predict the seizure further in advance than was addressed in (Costa et al. 2008). We also investigate the relative contributions of Costa et al's 14 features via further experiments in which two separate feature selection methods are explored.

5.2 Background and Related Work

In their recent work, Costa et al. compared various neural network approaches for classifying EEG recordings into one of the four classes, *ictal*, *pre-ictal*, *inter-ictal*, and *post-ictal*, as described in section 5.1, with *pre-ictal*, the target class, corresponding to a short period of time immediately before the occurrence of the seizure. They extracted 14 features for classifying the EEG signals, based on signal energy attributes, wavelet transforms and non-linear system dynamics. We have presented these features in section 4.2.3 and will briefly discuss these features in section 5.3. The study in (Costa et al. 2008) was carried out on EEG recordings of two patients taken from the Freiburg EEG Database (Epilepsy.uni-freiburg.de 2007) .They study three types of experiment:

- *Single*: Train a neural network on 70% of a single patient's recordings, and tested on the remaining 30% of the recordings from the same patient.
- *Different*: Train a neural network on one patient's recordings and test on the other patient's recordings.

• *Multiple*: Train a neural network on 70% of both patients' recordings, and test on the remaining 30%.

They tested six neural network methods: Radial-basis function, standard feed-forward, two kinds of recurrent networks, and two kinds of time-delay networks. Results were presented in terms of *Specificity*, *Sensitivity*, and *Accuracy*, defined as follows:

Specificity: the percentage of non-*pre-ictal* cases in the test-set that were correctly classified.

Sensitivity: the percentage of *pre-ictal* cases in the test-set that were correctly classified.

Accuracy: the overall percentage of samples in the test-set that were correctly classified.

We reproduce a selection of Costa et al's results in Table 5.1, for the radial basis function, and the standard recurrent network. These are generally representative of Costa et al's results, with the RBF network being among the worst performing, and the Recurrent network among the best, particularly with reference to the 'Single' case.

The results from (Costa et al. 2008), illustrated in Table 5.1, show that better performance is possible for 'single' (i.e. patient-specific) experiments. Concerning both 'Different' and 'Multiple' scenarios, they show some seeds of promise (using the specialised set of extracted features) for building *pre-ictal* predictors for unseen patients based on training data from multiple patients. However in this paper we focus on the single-patient case.

	Single	**Different**	**Multiple**
	RBF Network		
Specificity	96%	89%	100%
Sensitivity	98%	2%	97%
Accuracy	93%	63%	100%
	Recurrent Network		
Specificity	99%	32%	98%
Sensitivity	97%	94%	76%
Accuracy	97%	28%	95%

Table 5.1 Selected results from Costa et al. (Costa et al. 2008). Note, Costa et al. present only results from a single run to two significant digits.

Overall, seizure prediction studies are beginning to show excellent sensitivities and specificities for classifying *pre- ictal* brain activity, where *pre-ictal* tends to correspond to a few seconds (or at most a small number of minutes) in advance of a seizure. A key

element in achieving this is the use of features extracted from the EEG data that relate to signal energy attributes and nonlinear dynamics metrics meanwhile, patient-specific *pre-ictal* predictors seem much more readily achievable than predictors that can distinguish *pre-ictal* activity in patients whose data were not involved in training.

Here, we first examine additional algorithms on the Costa et al. data, and then we look further into the features used in (Costa et al. 2008) by running some feature selection experiments. We also begin to look at whether effective Sensitivity and Specificity values might still be attainable if we re-engineer the data towards distinguishing *pre-ictal* states that are more than a given time in advance of the ictal phase.

5.3 Experiments: SVM, EANN And Feature Selection Methods

In this section, we build on the work of Costa et al. to evaluate the effect of using feature selection methods on the 14 dimensional feature-set. We start by summarizing the characteristics of the feature-set and data preparation step. We then present the basic machine learning algorithms and initial tests of these algorithms on the full feature-set. Finally we present the feature selection experiment and illustrate experimental results.

5.3.1 Dataset, Classes and Features

The dataset we use is that also used by Costa et al. (Costa et al. 2008), and is available at (Epilepsy.uni-freiburg.de 2007). We use the data for patient 2, comprising around 135 minutes of EEG recordings, including 3 seizures, with ~30—40 minutes of data leading up to, and ~10 minutes following each seizure. The data available at the given URL is preprocessed from the raw EEG data into 14 features at each timepoint according to steps in chapter 4.

Each data instance therefore represents a 5-second time period, and is pre-classified into one of four distinct classes: *inter-ictal*, *pre-ictal*, *ictal*, and *post-ictal.* In the epilepsy literature, 'ictal' refers to a seizure, and as we have indicated, classification and evaluation are centred on the Sensitivity and Specificity of the classifier in relation to predicting the *pre-ictal* class.

5.3.2 Algorithms and initial tests

We apply a Multi-Class Support Vector Machine (MC-SVM – as supplied in the Matlab Spider library (People.kyb.tuebingen.mpg.de n.d.)), and an EANN (standard neural network trained by an evolutionary algorithm) to Costa et al's 'Single' patient scenario. Our main aims are to investigate feature selection and advance prediction, but we begin by showing in Table 5.3 the results of MC-SVM on a single feature, namely, Signal Energy (see section 2.2.1). We perform this test in order to establish whether using the energy feature alone, we can achieve the same results as the 14 feature experiments of Costa et al. The results reveal that all three measures of Specificity, Sensitivity and Accuracy, averaged over 10 random split runs, are relatively low, at Accuracy less than 50%.

	Specificity	Sensitivity	Accuracy
MC-SVM	60.877	57.587	46.511

Table 5.3 Test results of training MC-SVM on signal energy, averaged over 10 runs.

We now present results with MC-SVM and EANN (Table 5.4), for comparison with the representative results from Costa et al. shown in Table 5.2. In each case these are the means (with standard deviations in parenthesis) of 10 runs on different random 70%/30% splits of the data.

	MC-SVM	EANN	Costa et al.
Specificity	99.8%	95.2%	99%
Sensitivity	80.3%	97.8%	97%
Accuracy	96.4%	97.0%	97%

Table 5.4 MC-SVM and EANN results on the 'Single' Experimental scenario against Costa et al.'s Recurrent Network.

The EANN is clearly competitive with the results from Costa et al, while the MC-SVM also performs well but with a less than ideal average for Sensitivity. Table 5.5 presents the t-test analysis of MC-SVM against EANN. Though the results show significant differences between the two, we make no particular claims about the comparative performance of these algorithms with each other or with those tested by Costa et al, since the amount of data involved is too small to base firm conclusions. However, we note that the performance of the EANN is clearly promising for future study, while both

the SVM and EANN seem suitable for use in our main experiments, which follow, and which are aimed at investigating feature selection and advance prediction in this context.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
MCSVM vs. EANN	5.43	1.27	-5.98	-4.16	-12.66	9	0.00

Table 5.5 The t-test for 10 runs of MCSVM and EANN, calculated using the S1-score.

In Table 5.6 we compare the results of the single feature MC-SVM (Table 5.3) to that of the 14 feature MC-SVM, in order to analyse whether using Costa et al.'s 14 features significantly improves prediction. The mean S1-score of the two experiments differs by 30.407%, with our implementation of Costa et al.'s 14-dimensional feature-set performing significantly better than the benchmark 1-dimensional feature-set (p=0.000). From these results, we can conclude that using additional features yields a significantly better performance outcome.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
1-d vs. 14-d	30.407	1.832	-30.831	-28.210	-50.950	9	**0.000**

Table 5.6 The t-test for 10 runs of MCSVM on a single feature (Signal Energy) and on Costa et al.'s 14 features, calculated using the S1-score.

5.4 Experiments: Feature Selection

We explore two feature selection methods here, available in the Matlab Spider toolbox (People.kyb.tuebingen.mpg.de n.d.): namely *clustub* and *mutinf*. In each case, the feature selection method takes the training dataset and ranks the features in order of importance. Clustub uses spectral clustering for this (related to principal components analysis), while mutinf uses mutual information (essentially an information-theoretic measure of the amount of information each term contains about the classes in the data).

For each combination of feature ranking method (clustub or mutinf), learning algorithm (MC-SVM or EANN) and number of features *N* (from 12 to 2 in steps of 2), the following was repeated 10 times for different randomised 70/30 splits of the classes:

1. Obtaining a feature ranking on the training set
2. Run the learning algorithm on the top *N* features

For each combination of feature ranking method, learning algorithm, and number of features, we recorded the mean and standard deviation of these 10 runs, and the results are recorded in Table 5.7.

	MC-SVM		**EANN**	
	Sensit.	Spec.	Sensit.	Spec.
Clustub (12)	75.9% (11.81%)	99.7% (0.20%)	79.5% (10.1%)	100% (0.0%)
Clustub (10)	52.1% (23.89%)	99.9% (0.17%)	58.3% (13.7%)	100% (0.0%)
Clustub (8)	13.8% (18.32%)	99.9% (0.13%)	21.8% (11.1%)	99.6% (0.1%)
Clustub (6)	3.7% (8.31%)	99.9% (0.14%)	10.0% (16.2%)	100.0% (0.0%)
Clustub (4)	0.3% (0.93%)	100.0% (0.00%)	6.3% (1.2%)	100.0% (0.00%)
Clustub (2)	0.5% (1.17%)	100.0% (0.11%)	1.7% (1.8%)	100.0% (0.11%)
Mutinf (12)	78.3% (5.15%)	99.8% (0.19%)	86.4% (6.2%)	100% (0.0%)
Mutinf (10)	79.2% (7.39%)	99.7% (0.20%)	85.1% (6.7%)	100% (0.0%)
Mutinf (8)	82.0% (6.93%)	99.7% (0.24%)	87.0% (6.9%)	100% (0.0%)
Mutinf (6)	60.6% (15.21%)	99.8% (0.16%)	62.6% (9.2%)	100% (0%)
Mutinf (4)	21.8% (4.54%)	98.4% (0.61%)	28.6% (7.5%)	97.2% (0.8%)
Mutinf (2)	9.9% (12.90%)	98.8% (1.55%)	11.7% (15.2%)	96.4% (1.9%)

Table 5.7 Seizure detection results for two different approaches to data preparation

We also recorded the sets of feature rankings for each combination (since the training-sets changed between trials, feature rankings may also vary). Table 5.8 shows the mean rank for each of the features (referred to by the category and ID provided in Table 5.2), arising from rankings by the mutinf method.

As we can see in Table 5.7, Specificity remains very robust, however few features we seem to choose, but there is a clear pattern of variation with Sensitivity values. Clustub seems particularly poor at ranking features in such a way that the higher ranked choices will work well together in this task. For the moment, we will focus on the mutinf results. This shows a peak in Sensitivity at 8 features (where the Sensitivity value seems slightly better than when using the full set of 14 features), with then a sharp drop in performance as we reduce from 6 features to 4. As we saw in Table 5.4, the EANN approach and MC-SVM achieve similar performance, but the EANN seems to have the edge. Although the relatively high standard deviations suggest that the differences are not statistically significant.

Feature	Mean rank from mutinf
Accumulated energy	1.0
Energy STE 4 (50—100Hz)	3.4
Energy LTE 1 (0—12.5Hz)	4.1
Lyapunov exponent	5.6
Energy STE 3 (25—50Hz)	5.8
Energy STE	6.8
Energy LTE 2 (12.5—25Hz)	7.3
Energy LTE 4 (50—100Hz)	7.6
Energy level	8.7
Energy LTE 3 (25—50Hz)	9.6
Energy STE 2 (12.5—25Hz)	10.0
Energy STE 1 (0—12.5Hz)	11.0
Energy LTE	11.3
Correlation dimension	13.0

Table 5.8 Feature rankings averaged over 10 randomised training-sets. 1 is highest, 14 is lowest rank.

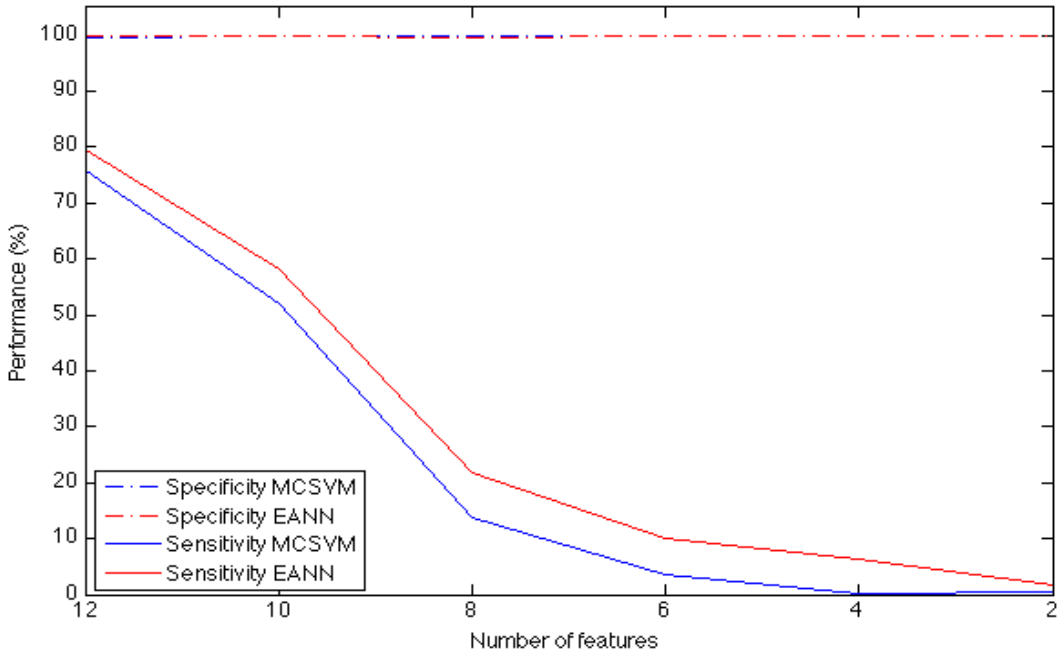


Figure 5.2 Epileptic Seizure Prediction results for EANN and MCSVM using **Clustub** stepwise dimensionality reduction - The plot displays the Sensitivity and Specificity for various numbers of features, averaged over 10 randomised runs.

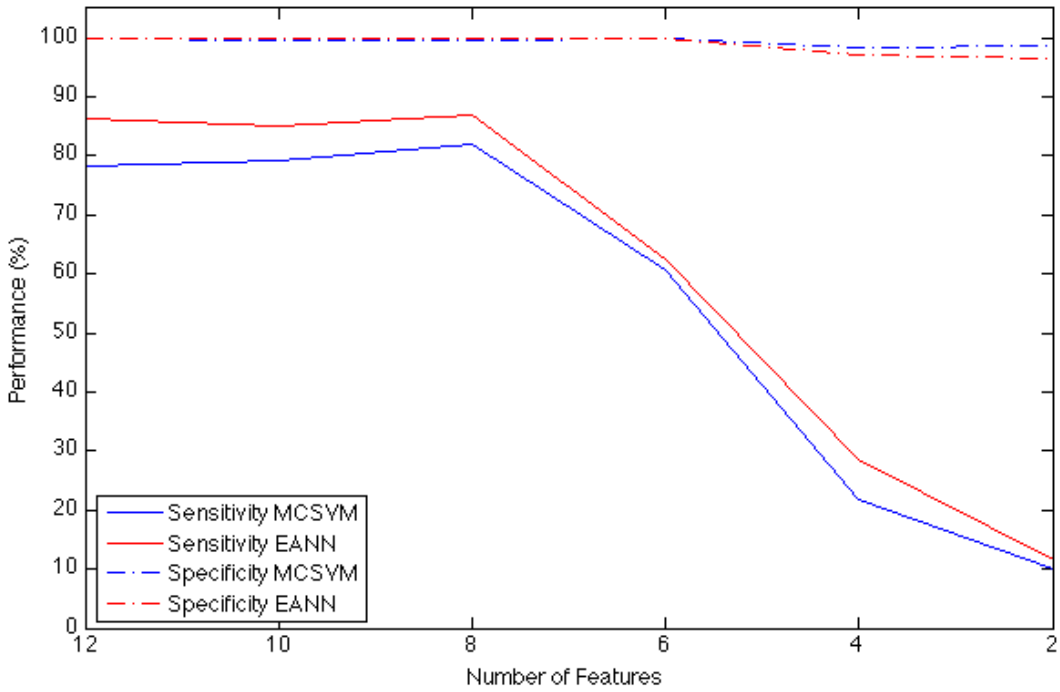


Figure 5.3 Epileptic Seizure Prediction results for EANN and MCSVM using **Mutinf** stepwise dimensionality reduction - The plot displays the Sensitivity and Specificity for various numbers of features, averaged over 10 randomised runs.

More interesting are perhaps the feature rankings shown in Table 5.8. The 'Accumulated energy' feature is very clearly dominant and predictive, while, of the nonlinear dynamics features, the maximum Lyapunov exponent is clearly valuable, but correlation dimension seems to contribute very little. Meanwhile, short-term (in this case, measured over 7 seconds) and high frequency energy attributes tend to rank higher than long term energy attributes, especially longer term high frequency ones.

5.5 Experiments: Advance Seizure Prediction

In this section we study the efficacy of making predictions well in advance of he feature onset. We start by describing how the experiment was implemented and continue by presenting important results from the experiment.

5.5.1 Implementation

The algorithms used for this experiment is similar to that described in 5.3.2.

So far, both here and in Costa et al, we have used the dataset as provided, in which, in connection with three seizures, 180 feature vectors classified as *pre-ictal*, varying in timestamp from 0 seconds (i.e. immediately preceding) to ~5 minutes before the onset of the ictal period. In order to investigate advance prediction, we presently work with adjusted datasets, in which we either remove or reclassify some of the feature vectors whose timestamps lead up to an ictal period.

We refer to either *del* or *rename* experiments. In a *del N* experiment, the *N* minutes of data (*N* x 12 feature vectors) immediately before the seizure are removed from the data, and, where this leaves fewer than 24 *pre-ictal* vectors before a given seizure, a number of appropriate *inter-ictal* vectors are reclassified to ensure that 2 minutes' worth of *pre-ictal* vectors exist. This means that a correct classification of *pre- ictal* is predicting the seizure between *N* and *N*+2 minutes in advance. In a *rename N* experiment, the situation is the same as that in a *del N* experiment, with the exception that no data are deleted – instead the *N* minutes of *pre-ictal* data are renamed as *ictal*.

In these experiments we use all 14 features, and again we use both MC-SVM and EANN as described in section 5.3.2.

5.5.2 Results

The results (Table 5.9) are average test-set results over 10 random 70/30 splits of the data, shown with standard deviation in parenthesis.

As Table 5.9 shows, Specificity values remain quite robust throughout these experiments (i.e. there is a low number of false positives), while Sensitivity values drop sharply, for both *del* and *rename* when *N* is 2, however these grow sharply again and peak at *N*=8 for both *del* and *rename*. Considering just the MC-SVM results, the *N*=8 cases show that prediction of a seizure 8—10 minutes in advance seems possible with sensitivities better than those achieved with (effectively) *N*=0 (Table 5.4). However, again we note that the standard deviation is relatively high, and no statistically significant conclusions can yet be made. Nevertheless, given the use of the Costa et al. feature-set, it seems that patient- specific classifiers could be trained that achieve prediction of seizures 8—10 minutes in advance with high Sensitivity (missing few real seizures) and high Specificity (few false alarms).

	MC-SVM		**EANN**	
	Sensit.	Spec.	Sensit.	Spec.
del/2	6.9% (14.9%)	100.0% (0.05%)	27.8% (11.1%)	95.0% (3.1%)
del/4	60.1% (9.13%)	99.9% (0.15%)	68.10% (7.3%)	97.1% (1.5%)
del/6	67.9% (10.95%)	99.5% (0.54%)	72.7% (8.1%)	99.5% (0.4%)
del/8	89.2% (6.97%)	99.3% (0.40%)	94.8% (4.4%)	98.0% (0.4%)
del/10	79.2% (9.18%)	99.4% (0.37%)	86.1% (8.9%)	98.1% (0.5%)
rename/2	29.1% (6.62%)	99.7% (0.25%)	35.4% (7.8%)	97.8% (1.1%)
rename/4	52.3% (14.79%)	99.3% (0.51%)	58.1% (10.2%)	98.0% (0.7%)
rename/6	63.1% (10.05%)	98.6% (0.59%)	70.2% (8.2%)	97.8% (0.8%)
rename/8	82.9% (11.70%)	99.0% (0.66%)	88.8% (7.8%)	97.2% (0.8%)
rename/10	62.6% (16.50%)	99.4% (0.46%)	66.1% (12.5%)	99.1% (0.7%)

Table 5.9 Advance Prediction results for two different approaches to data preparation

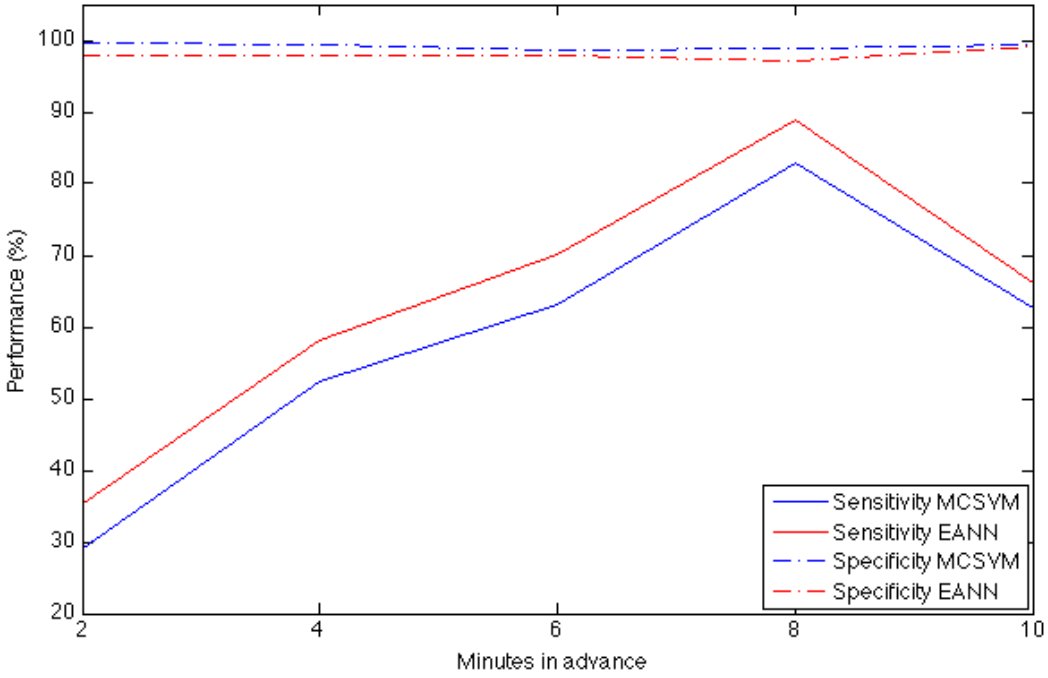


Figure 5.4 Advance Prediction results for EANN and MCSVM using **Rename** Data Preparation method - The plot displays the Sensitivity and Specificity for various advance time-steps, averaged over 10 randomised runs.

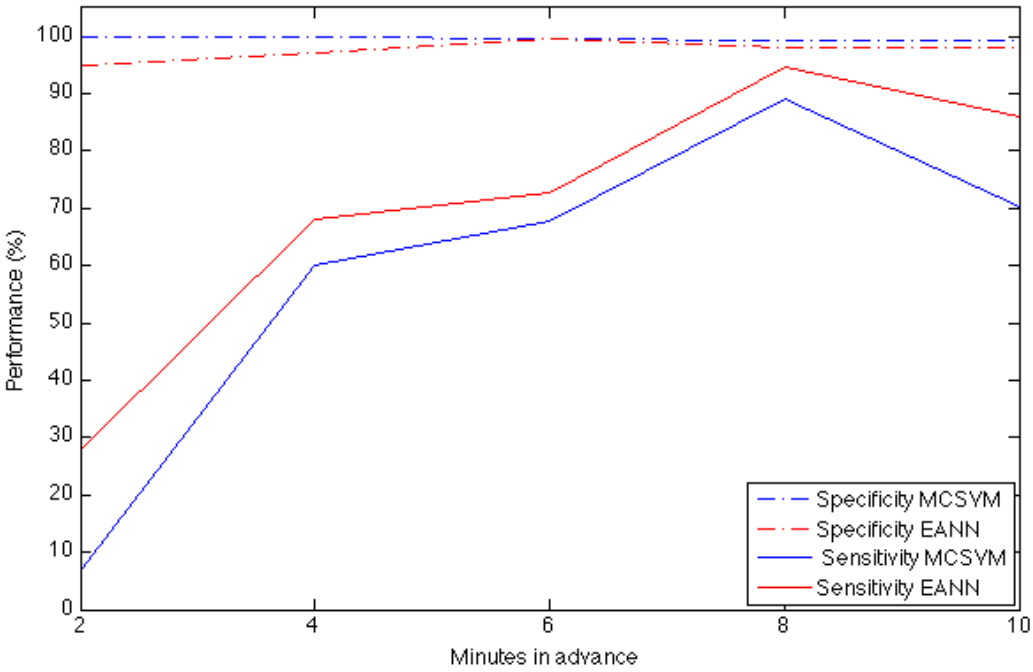


Figure 5.5 Advance Prediction results for EANN and MCSVM using **Delete** Data Preparation method - The plot displays the Sensitivity and Specificity for various advance time-steps, averaged over 10 randomised runs.

5.6 Further Discussion

Building on the promising work by Costa et al. (Costa et al. 2008) using 14 features extracted from the EEG, part of our study has involved investigating these features, to begin to assess their relative importance, as well as the potential for building classifiers that use fewer features (which carries the advantage, among others, of less bias towards over-fitting). We found that the use of mutual information was particularly well-suited to feature selection in this task, and it seems that promising results can be obtained, especially with the EANN, with only 8 of the 14 features. Analysis of feature rankings by *mutinf* showed particular prominence for features 1, 8, 9 and 14 (see Table 5.2). This suggests future research could concentrate on examining additional such features, respectively concerning the total accumulated energy in the signal, high-frequency aspects measured over short-term windows, low-frequency attributes measured over longer-term windows, and more metrics associated with Lyapunov exponents.

Other promising future work could explore wrapper techniques that attempt to derive an appropriate feature subset in tandem with the machine learning process.

The advance prediction results were similarly interesting. It seems clear that, among the 14 features used, detectable seizure-predictive patterns are present 8—10 minutes, as well as earlier, in advance of an ictal period. Both the MC- SVM and the EANN seem capable of constructing useful classifiers for 8—10 minutes advance prediction, with promising Sensitivity levels, at least in this single-patient context. However, more experimentation is required, with more patient data, in order to further validate these findings.

As discussed in the comprehensive recent survey by Mormann et al. (Mormann et al. 2006), the literature on seizure prediction studies is beset with a number of issues that confound progress, such as no standard approaches to experimental setup, large variation in the number and types of patients' studies, and in the amount of data (EEG and otherwise) available per patient. Also, validation (i.e. obtain results for predictors on data unseen during training) is very often not done at all, since not enough data are available. Naturally, more sharing of data, and experimental design protocol, must be encouraged; whereas, reporting of results obtained on unseen data should typically be mandatory.

5.7 Summary

In this chapter we reviewed relevant results from the Costa et al. study where a 14 dimensional feature-set of a single patient from the Freiburg EEG Database was implemented in a number of experimental conditions with several Artificial Neural Networks. We evaluated the feature-set with two other classifiers, namely Multi-Class Support Vector Machine (MC-SVM) and Evolved Neural Network (EANN). The benchmark values on the 'single' patient scenario were similar to those produced by Costa et al.

We further evaluated the features with two feature selection methods: Clustub and Mutinf. We found that with a well-chosen reduced feature-set (using mutual information), promising results can be obtained with only 8 of the 14 features. Further analysis showed that the accumulated energy in the signal, the maximum Lyapunov exponent, as well as measures of high-frequency signal components measured over short term windows, seem most promising for future research into accurate advance prediction models.

In addition, we implemented an Advance Prediction algorithm where the prediction window was stretched over pre-determined timepoints. We observed that, using either a Multi-Class Support Vector Machine (MC-SVM) or an Evolved Neural Network (EANN), reasonable Specificity and Sensitivity could be achieved for prediction 8--10 minutes in advance of the seizure onset. Indications are that the EANN performance is preferable for advance prediction, however the results so far do not support this with statistical significance.

These results have served to indicate that we can achieve similar or better results to Costa et al. (Costa et al. 2008) using a similar (and hence impoverished) experimental regime. In future chapters we explore more comprehensive scenarios to validate these findings and make many new observations in the context of multiple patients and other scenarios and rigorously enhance and evaluate the Feature selection and Advance Prediction experiments.

Chapter 6

Feature Selection and Dimensionality Reduction

In this chapter, we re-visit the importance of effective feature engineering in our seizure detection problem, drawing from our preliminary feature selection experiment presented in chapter 5. From the results in chapter 5, we concluded that using the correct feature selection method, we are able to produce a significantly smaller subset of features, for which the performance measures of the full feature-set are maintained. We also concluded that the contribution of certain features to the success of our seizure detection model is less than others.

In this chapter we present a number of experiments to further evaluate the established conclusions, under exhaustive experimental conditions. We hope to achieve a better understanding of the role of different features in the performance of our classifiers, and to determine the optimum feature settings under which, the performance of the model is at its highest value. We also aim to extend our feature-set based on heuristic results of experiments presented in this chapter, and further evaluate the performance of this extended feature-set. By determining the most effective features, we are able to build classifiers of increased efficacy and to clarify the role of EEG channels and features in successful seizure classification.

This chapter presents 3 experiments. All experiments were carried out on Patient-Files from the Freiburg EEG Database (as detailed in chapter 4), in a single-patient mode, where the classification model is built from, and tested on a single patient. Section **6.1 motivates** us on why feature selection is important in machine learning problems, in particular, epileptic seizure detection studies. In **experiment I**, we apply the feature selection algorithms on each patient under **default experimental settings** presented earlier in chapter 5. The outcome of this step is a ranking of each feature based on algorithm-specific criteria; this is discussed in section **6.2.3**. The **ranking table** will be further used in the same experiment to perform a **stepwise dimensionality reduction** on segments of each patient's EEG which contain ictal (seizure) states recorded from a single focal channel (similar to that seen in chapter 5). In **experiment II** we perform the same feature selection and dimensionality reduction steps as experiment I, this time on an extended feature-set which is derived from all 6 recorded

EEG channels of each patient from the Freiburg EEG Database, referred to as **Multi-Channel Patient-Files**. In **experiment III**, we heuristically **extend our feature-set,** building on results of earlier experiments of this chapter. This is to comprise additional features, of similar characteristics to those features with the highest performance outcome. We further analyse the performance of our classifiers using the extended feature-set. In section **6.5**, we **discuss the outcome** of the experiments and we present possible **conclusions** drawn from our results in section **6.6**.

6.1 Motivation

As described in chapter 3, epileptic seizure detection and prediction from EEG recordings has been the focus of many studies in the field of computational neurology. The unpredictable nature of the seizure can impose potential risk for the individual with epilepsy. Therefore, the automatic detection of the seizure at the time of its occurrence or seconds before the neuronal onset, can give rise to timely intervention, minimising the risks involved.

The raw data recorded from EEG channels prior to pre-processing, is merely voltage outputs from each channel. This means that the data from the Freiburg EEG Database, in its raw format, has six features corresponding to the 6 recording channels. The small number of features and the large number of input data are disproportionate. Some studies have carried out automatic seizure detection algorithms on such data, mainly to showcase the power of the underlying machine learning algorithm, discarding the extensive feature engineering body of work, which could lead to potential algorithmic improvement (e.g. (Santaniello et al. 2011)). Amongst the studies that use feature engineering, the majority have i) used either non, or simple machine learning algorithms ii) have not conducted sufficient validation in terms of Sensitivity and Specificity iii) have evaluated the features individually (uni-variate analysis), instead of combining multiple features.

The problem in question encompasses the elements of both machine learning and feature selection problem by treating the question as a machine learning problem, we find the best learning model which yields the highest performance outcome on unseen data, and by posing the question as a feature selection problem, we are able to use the extensive body of work behind EEG feature engineering to further optimise the

learning algorithm. In other words, we seek to further improve seizure classification by optimising and improving the combination of features used in the learning algorithm. We aim to optimise the detection of seizure states through heuristic exploration of the default features from (Costa et al. 2008), and the introduction of some new features. The outcome of this chapter is the determination of best feature combinations per patient in addition to an optimum overall combination of features, considering the results from all patients. The latter will further be used for the extraction of additional features

6.2 Experiment I: Dimensionality Reduction on Single EEG Channel

This section comprises a series of seizure prediction and feature reduction experiments, under the simplest experimental conditions, namely patient specific training, validation and test sets, single-channel feature-set and ictal datasets from the Freiburg EEG Database. The aim of this set of experiments is to evaluate the impact of reducing the feature-set on the learning performance, on a larger number of patients; in this case, all the patients from the Freiburg EEG Database. We aim to see whether general patterns, such as optimum number of features and highly ranked features, emerge among the population of tested patients. The outcome of this series of experiments is a set of feature rankings per Patient-File in addition to the performance measures of the respective learning models constructed at each dimensionality reduction step.

6.2.1 Methods For Single-Channel Dimensionality Reduction

In this section we present the data preparation and implementation steps undertook for conducting the stepwise dimensionality reduction experiment on single-channel Patient-Files from the Freiburg EEG Database.

Data Preparation of Single-Channel Data For Dimensionality Reduction

The data used in this experiment are from the Freiburg EEG Database. The unprocessed data are originally in ASCII format, comprising signal voltage recorded from 6 incoming channels at 256 Hz. Each patient has numerous ASCII files, which were organised based on the recorded segment and the incoming channel. These files were prepared according to steps described in chapter 4. The preparation steps resulted

in 21 Excel and Matlab Patient-Files, each of which contain 14 extracted features of three main categorisations: signal energy, wavelet transforms and non-linear dynamics. A list of the engineered features is presented in Table 6.1 for ease of reference.

Concept	*Features*
Signal Energy	Accumulated energy
	Energy level
	Energy variation (short term energy (STE))
	Energy variation (long term energy (LTE))
Wavelet Transform	Energy STE 1 (0Hz − 12.5Hz)
	Energy STE 2 (12.5Hz − 25Hz)
	Energy STE 3 (25Hz − 50Hz)
	Energy STE 4 (50Hz − 100Hz)
	Energy LTE 1 (0Hz − 12.5Hz)
	Energy LTE 2 (12.5Hz − 25Hz)
	Energy LTE 3 (25Hz − 50Hz)
	Energy LTE 4 (50Hz − 100Hz)
Nonlinear system dynamics	Correlation dimension
	Max Lyapunov Exponent

Table 6.1 Original 14 features extracted from EEG Channels – The feature-set is based on the work of Costa et al. (Costa et al. 2008).

The features were calculated using a moving window analysis for each 5-second block of data, all of which were labeled with the ictal state of the patient. The seizure-state labels take values of 1 - 4, which respectively represent inter-ictal, pre-ictal, ictal and post-ictal states of the brain. Each prepared file holds 1 hour of data per seizure, comprising all 4 classes. The Patient-Files contain 1 to 5 seizure recordings for each patient. The rendered 14 features were extracted from a single focal EEG channel for each patient. The 5 remaining channels were not used in this experiment.

Implementation of Dimensionality Reduction on Single-Channel Data

This Dimensionality Reduction study is composed of a series of segregated experiments conducted on each individual Patient-File, as shown in Figure 6.1.

Each preprocessed Patient-File was separately normalised to values in the range [0, 1] for ReliefF feature selection algorithm, and was split into 70% training-set and 30% test-set, using a random seed permutation. The feature selection method, namely ReliefF, was implemented on each training-set for a total of 10 runs. The outcome of the total runs of each feature selection algorithm was an $f \times 10$ matrix of rankings where f is the total number of features, and each row is a ranking $r = 1 \ldots f$, where $r = 1$ accounts for the best feature and $r = f$ accounts for the worst feature. Best and worst are determined based on the ranking criteria of ReliefF feature selection algorithm, as described in section 2.3. The rankings for each feature were averaged over 10 runs, and the features were sorted according to the ascending average rank. Table 6.2 lists an example of such rankings for Patient-File 2.

	ReliefF Ranks
Accumulated energy	1
Energy level	8
Energy STE	7
Energy LTE	4
Lyapunov exponents	2
Correlation dimension	9
Energy STE 1	10
Energy STE 2	5
Energy STE 3	3
Energy STE 4	6
Energy LTE level 1	11
Energy LTE level 2	14
Energy LTE level 3	13
Energy LTE level 4	12

Table 6.2 Rankings of the 14 features of Patient-File 2 ReliefF feature selection method – The features were based on (Costa et al. 2008) and were obtained from the default EEG channel for the patient.

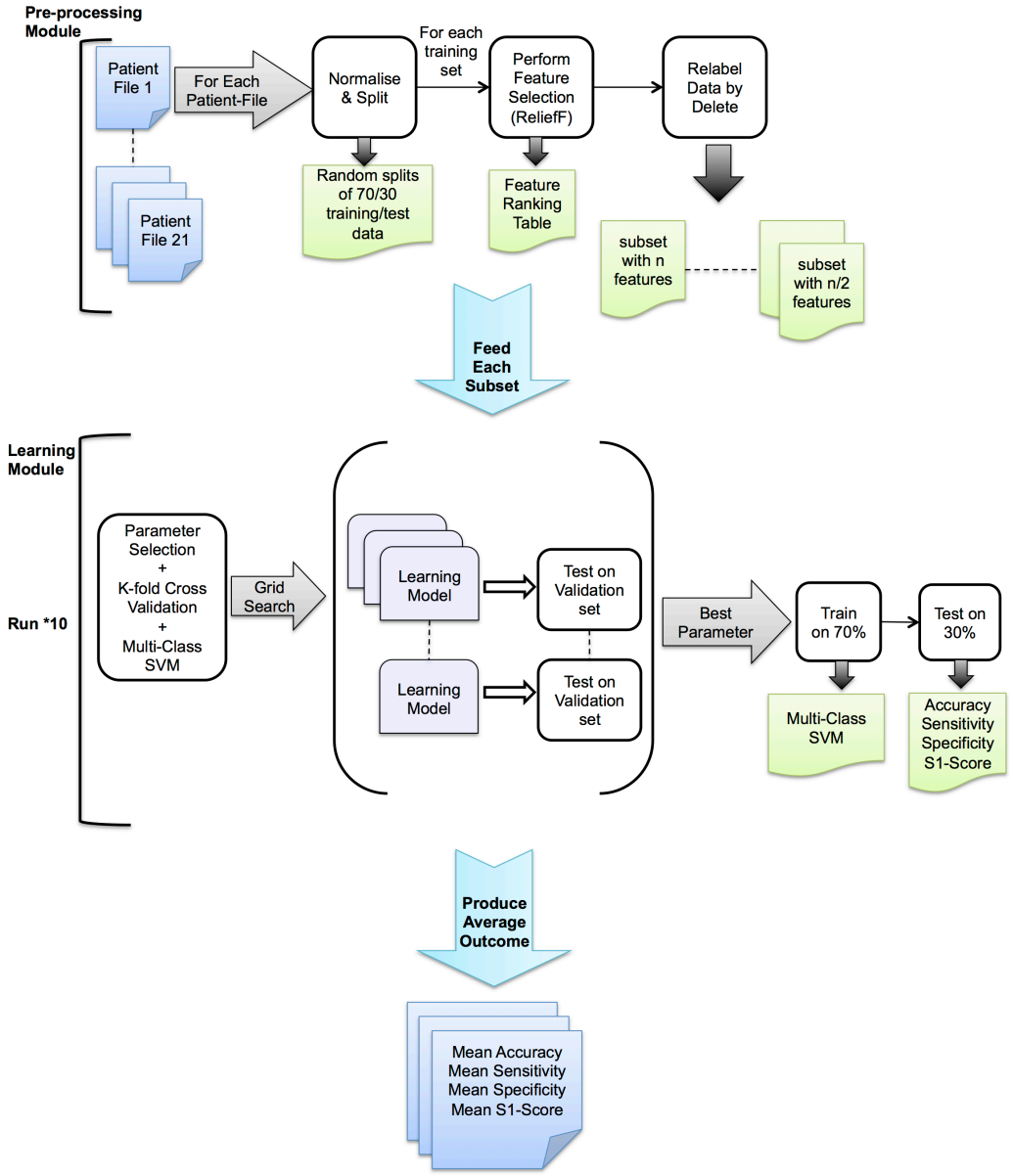


Figure 6.1 the architecture of the Dimensionality Reduction Experiment – The system consists of a Pre-processing Module and a Learning Module. The data preparation and initial experimental setup takes place in the Pre-processing Module, which varies for each experiment. This is separated from the learning and classification task in the Learning Module, which remains unchanged for the main part of the experiments.

For each Patient-File, the constructed ranking table was used to reduce the feature subset in a stepwise manner. For each ranking table of size f, n features were removed at each step, for $s = f/2$ number of steps. In this particular experiment, $f = 14$ and $s = 7$. The n features removed at each step were those with the lowest ranks. This resulted in $d = f/2$ number of training-sets and test-sets. Each training-

file subset was separately fed into a learning module, where the training-set was further divided into random permutations of training-set (90%) and validation set (10%) over several folds for parameter selection purposes.

The Multi-class Support Vector Machine (See Chapter 2) was used in the learning module in order to learn mappings from training features, and use them to classify the blocks of test and validation data into inter-ictal (1), pre-ictal (2), ictal (3) and post-ictal (4) states.

The Multi-class SVM classifier was implemented using the LIBSVM software package for Matlab (Chang & Lin 2011; Csie.ntu.edu.tw 2013) , in which a 'one-against-one' approach (Knerr et al. 1990; Kreßel 1999) was used, where for each k number of labels k $(k\text{-}1)/2$ classifiers are constructed. Error in learning was penalised (Chang & Lin 2011) based on weights specified for each class:

$$\begin{aligned} W_1 &= 1 \\ W_2 &= \text{int}\_ictal / pre\_ictal \\ W_3 &= int\_ictal / ictal \\ W_4 &= int\_ictal / post\_ictal \end{aligned} \qquad (1)$$

The learning algorithm penalises the misclassification of each label in accordance with the corresponding weight W, where W_1 is the weight for inter-ictal misclassification; W_2 is for pre-ictal, W_3 for ictal and W_4 for post-ictal misclassification. Using misclassification weights is especially useful when working with unbalanced datasets, where the occurrence of a certain class is more probable than the others. By carefully choosing the weights, misclassification errors can be avoided. In our dataset, class 1, the inter-ictal class, had a higher frequency than the other 3 classes. For each hour of ictal data, there were approximately 5 minutes of seizure data, 5 minutes of pre-ictal and 5-minutes of post-ictal data, yielding a highly imbalanced dataset where ~75% of the dataset (45 minutes per 1 hour of ictal data) consisted of inter-ictal data with little detectible neuronal abnormality. We therefore set W_1 to 1 and set all the other weights as $W_{class} = size(\text{int}\_ictal) / size(class)$, in an effort to avoid inaccurate classification and performance outcomes.

The RBF kernel was used with the Multi-class SVM. The two RBF parameters *C* and γ were chosen using a grid search. The grid search was implemented in a 10 fold cross-validation (explained in chapter 2) to search through a grid of parameters. When the cross-validation was complete, the parameters with the highest CV Accuracy were returned. In these experiments, we used a 10 fold CV with a nested grid search for finding the best C and γ values. Since the complete grid search is computationally expensive, we search for parameters in the following two steps:

1- A 'coarse' parameter grid is searched for 'better' regions

2- After identifying these better regions, a finer grid is used to find the best (C,γ).

After conducting step 1, we concluded that the 'better' regions for C and γ values are:

1- For experiments training on single patients or several patients up to a combination of 12, the grid search was conducted on the log2 of the kernel parameters: log2C is in the range [8, 16] in intervals of 4; log2γ is in the range [0, 10] in intervals of 2.

2- For experiments training on multiple patients where the combination of patients is above 12, log2C is in the range [8,16] with intervals of 4; log2γ is in the range [4,8] with intervals of 2.

This experiment yielded $f/2$ final Multi-class SVM classification models for each Patient-File, parameters of which were derived from 10 random seed cross validation classifiers.

After constructing the classifiers, they were then tested on their corresponding unseen test data, and classification results were produced in terms of Accuracy, Sensitivity (or recall), Specificity (or precision) and S1-Score.

There were two main outcomes for this experiment: i) The performance of the classifiers over different stages of dimensionality reduction, and over distinct control variables, ii) a feature-ranking table for each Patient-File which consisted of the ordering of importance of the features, determined by ReliefF feature selection method. Another worthy of note outcome is the identification of the ranking of 14 features

averaged over the performance of all patients, which will be the basis of a number of experiments in chapters 6 and 7.

6.2.2 Results of Single-Channel Dimensionality Reduction

In Figures 6.3 we see the results of the stepwise Dimensionality Reduction performed on each individual Patient-File, comprising features from a single channel recording. Figures 6.3 presents average performance of the classifier at each reductive step. The performance measures are namely Accuracy, Sensitivity, Specificity, and S1-Score, averaged over all corresponding Patient-File classifications.

Each of the Patient-File classifiers was constructed from a different subset of features, depending on the feature-ranking produced by the feature selection algorithm. In Table 6.5 we have listed the ranking for each feature, averaged over all Patient-Files. We present results of the stepwise dimensionality reduction classification experiments, irrespective of the specific features subset used by each classifier.

ReliefF Dimensionality Reduction

In Figure 6.3 and Table 6.3, we see the results of the stepwise dimensionality reduction using the ReliefF feature selection method. All the values in this plot are evidently high. Specificity starts at 98.89% at $f = 14$ and picks up at $f = 12$ with 99.35%. The high value is maintained and the maximum is hit at $f = 8$ with 99.47%, until it gradually declines in transition from $f = 8$ to $f = 6$ and $f = 6$ to $f = 4$, until the minimum is hit at $f = 2$ with 92.72%. Accuracy follows the same pattern as Specificity, rising at $f = 12$, peaking at $f = 8$ with 97.5%, decreasing until it sharply hits the minimum at 86.97% from transition from $f = 4$ to $f = 2$. Sensitivity also follows a similar pattern to that of Accuracy and Specificity, although variability between steps is more prominent.

The Sensitivity values start at $f = 14$ with 88.5% and increase in transition from $f = 8$ to $f = 6$. Sensitivity increases to 94.06%, (from 94% at $f = 8$) unlike the gradual decrease seen in the same step transition in Accuracy and Specificity. The decline from $f = 6$ to $f = 4$ is also sharper followed by a smoother slope from $f = 4$ to $f = 2$. The minimum Sensitivity is however higher than the minimum Accuracy at 88.65%. The result of S1-Score is similar to Sensitivity but with smoother transitions particularly from $f = 6$ to $f = 4$ and $f = 2$. The curve starts at 93.32%, inclines through $f = 12$, 10, 8 and 6, hits the maximum at $f = 8$ with 99.5%, at which point it declines until it hits the

minimum at 90.31%. The mean S1-Score values in steps f = 12, 10, 8 and 6 are above the general mean value for the S1-Score plot, while the benchmark value (f = 14) and f = 2, 4 are less than the mean. This indicates the improvement of the classification with feature counts 12, 10, 8 and 6, which mutually yield high values among all 4 measures.

	ACC	f	SP	f	SS	f	S1	f
min	86.97	2	92.72	14	88.50	2	90.31	2
max	97.50	8	99.47	8	94.06	8	96.64	8
f full	96.06	14	98.89	14	88.50	14	93.32	14
f = 2	86.97	2	92.72	2	88.65	2	90.32	2
mean	95.37		98.22		91.59		94.67	
median	97.08		99.21		92.12		95.57	
mode	86.97		92.72		88.50		90.32	
std	3.79		2.45		2.42		2.31	
range	10.54		6.75		5.56		6.33	

Table 6.3 Summary of important data statistics from the stepwise **dimensionality reduction on 18 single-channel** patients using **ReliefF**.

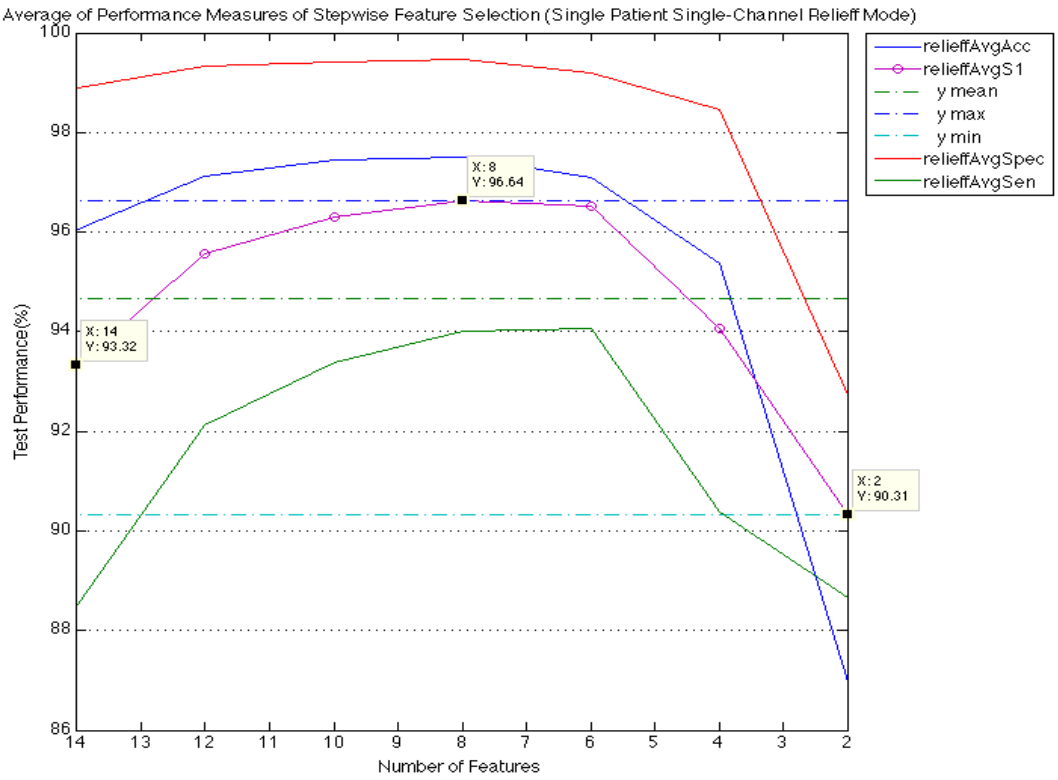


Figure 6.3 Summary of stepwise dimensionality reduction on 18 **single-channel** patients using **ReliefF**.

Inter-Patient Variability

Figures 6.4 shows the box and whiskers diagram of the S1-Score results of the ReliefF stepwise dimensionality reduction experiment, averaged across all single-channel Patient-Files. In the mRMR set of results, the highest variability can be seen at $f = 2$, with right skewness indicated by the median being in the upper half of the box. From this, it is apparent that the higher quartile is densely populated, denoting high variability in the lower quartile. Patient-Files 3, 10 and 14 are deemed as outliers in most reduction steps and are therefore removed from the average summary results. The other boxes in the plot are mostly parallel and the median lines are in the middle of the boxes indicating a normal distribution within the 25% and 75% quartiles. The short top whiskers in $f = 6, 8, 14, 12$ indicates a long tale for S1-Score Patient-File distribution at these feature counts. The boxes at $f = 12, 10, 8, 6$ are particularly short, indicating less variability among S1-Scores of the Patient-File population in these regions. This confirms that the high results seen in the average summary result plots are not arbitrary and are reflective of the majority of the Patient-Files. The variability amongst different steps of the experiment is also low, in particular among $f = 12, 10, 8, 6$. The box at $f = 2$ is the tallest among all other steps, although the median is towards in the upper half of the box, indicating that the upper quartile is densely populated with higher performance Patient-Files. In addition to Patient-Files 3, 10 and 14, which were removed from the average summary analysis, Patient-Files 9 and 15 are also marked as outliers for some reductive steps. However, we have not excluded them from the summary analysis as they do not appear as outliers in $f = 14$, which is our benchmark for these sets of experiments.

For each patient the feature-set with the maximum S1 score was identified and tested against the full feature-set for the same patient using a t-test. The results are presented in Table 6.4. The t-test for those patients whose maximum S1 was produced by the full set of features is eliminated from the table. The results show that the best feature-set for 10 out of 16 patients is indeed better than the full feature-set for each patient. In Table 6.5, the mean S1-score of the best feature-set of all patients is examined against the mean S1-score of the full feature-set. The results reveal that the feature-set with the highest S1-score selected by feature-reduction is significantly better than the full feature-set ($p=0.012$).

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1	1.113	3.937	-4.530	1.102	-1.377	9	0.202
Pat 2	13.471	4.190	17.078	23.072	15.151	9	**0.000**
Pat 3	2.330	7.568	-15.194	-4.367	-4.087	9	**0.003**
Pat 5	18.039	2.802	-3.742	0.267	-1.961	9	0.081
Pat 6	1.718	3.946	-15.076	-9.009	-9.156	8	**0.000**
Pat 7	4.035	3.144	-4.985	-0.486	-2.751	9	**0.022**
Pat 8	3.534	2.551	-3.414	0.236	-1.970	9	0.080
Pat 9	13.795	3.956	-11.760	-6.100	-7.138	9	**0.000**
Pat 11	0.157	2.818	-4.979	-0.947	-3.326	9	**0.009**
Pat 12	6.940	1.834	-1.372	1.252	-0.103	9	0.920
Pat 13	1.622	3.497	-3.110	1.893	-0.550	9	0.596
Pat 15	46.493	5.813	-10.561	-2.244	-3.483	9	**0.007**
Pat 16	0.208	2.187	-3.231	-0.103	-2.410	9	**0.039**
Pat 17	9.750	3.144	-4.985	-0.486	-2.751	9	**0.022**
Pat 18	0.305	2.627	-2.937	0.822	-1.272	9	0.235
Pat 20	11.949	3.905	-13.547	-7.961	-8.709	9	**0.000**

Table 6.4 Each row represents a t-test for each patient. The t-test compares the S1-score of the 10 runs of the full feature-set against the best feature-set. The best feature-set was determined by finding the feature-subsets with the highest S1-score averaged along 10 runs. The best feature subset is different for each patient. Those patients for which the best feature-set was the full feature-set were omitted from the table. The values in bold indicate $p \leq .05$ and are considered statistically significant.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1:21	6.450	10.752	-11.345	-1.556	-2.749	20	**0.012**

Table 6.5 The best mean S1-score of all patients is examined against the mean S1-score of the full feature-set, using a paired t-test.

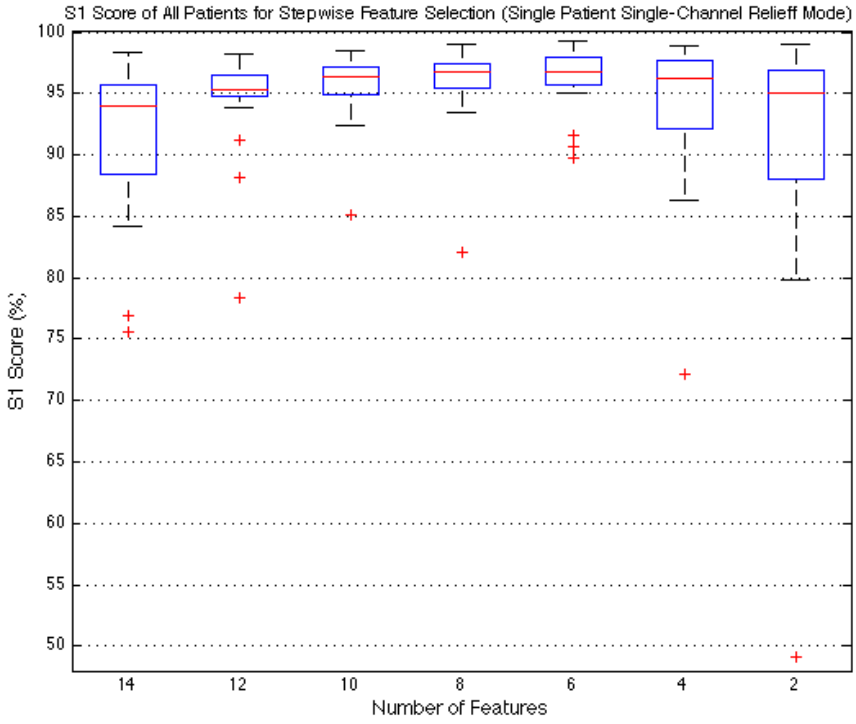


Figure 6.4 The Box and Whiskers diagram of stepwise dimensionality reduction on 21 **single-channel** patients using **ReliefF**.

6.2.3 Feature Ranking Table

The single-channel stepwise dimensionality reduction of each patient-File was conducted based on a ranking table produced by ReliefF. A ranking table was generated for each Patient-File, resulting in a total of 21 tables, a sample of which was presented in Table 6.2. The rankings generated for each patient varies from those of other patients. In Table 6.6, we see the average ranking of each feature over all Patient-Files for both feature selection methods.

We can clearly see that, based on average performance, accumulated energy is listed as the most important feature. In chapter 5 we used the Mutinf feature selection method on a single-channel Patient-File (see section 5.4). We used the feature ranking produced by Mutinf in a similar stepwise dimensionality reduction experiment. The outcome of the experiment clearly revealed an improvement on all 3 performance measures when number of features was reduced to 8. This was followed by a minor decline in the performance (Accuracy, Sensitivity and Specificity) of both EANN and MC-SVM classifiers in transition to feature 6, and a more drastic decrease from feature

4 onwards. This outcome hinted the importance of the top 8 features in improving classification performance. We further expanded on this in experiment I, where we used a different classifier, and a different feature selection method, on all Patient-Files extracted from the Freiburg EEG Database to further evaluate the validity of our conclusion. In section 6.3.2 we observed the outcome of this experiment where the mean performance measure (averaged over all Patient-Files) revealed a dip in the performance from feature 6 onwards, for both feature selection methods. We can clearly conclude that the features ranked in the 6 – 8 tier are of significant importance to the performance of the classifiers. This effect is not solely based on the number of features, as there is no linearity between the number of features and declination in performance. We assume that the intrinsic characteristics of the features are also of high importance as the performance ascends monotonically until feature 6; after which point, the performance monotonically drops.

Feature Type	*Features*	*ReliefF (Avg. Rank)*
Signal Energy	Accumulated energy	**2.95**
	Energy level	8.48
	Energy variation (STE)	8.62
	Energy variation (LTE)	**4.90**
Wavelet Transform	Energy STE 1	**8.00**
	Energy STE 2	**6.76**
	Energy STE 3	**7.67**
	Energy STE 4	**7.86**
	Energy LTE 1	**5.57**
	Energy LTE 2	**7.24**
	Energy LTE 3	**7.33**
	Energy LTE 4	**6.24**
Nonlinear system dynamics	Correlation dimension	10.95
	Max Lyapunov Exponent	12.43

Table 6.6 Rankings of the 14 features averaged over all single-channel patients from ReliefF feature selection method. The table is organised based on the category of the features. The rankings in bold are those of values ≤8 which will further be used in the feature extension experiments.

In Table 6.6 those features with mean rankings of ≤ 8 are emphasised in bold. In general, most wavelet transform features are in the high rank. Non-linear dynamics are in the lower rank tier according to ReliefF. Accumulated energy has the highest ranking, while other signal energy features are deemed important for ReliefF. This observable impact of feature rankings and consequences of removing features from the feature-set, provides us with invaluable insight into both the suitability of the feature selection method for our particular learning problem and the role of features in improved detection performance. We use these observations, in future sections, to further expand our feature-set in an effort to i) improve classification performance ii) further identify the most suitable type of features for the seizure detection and prediction problem.

6.2.4 Discussion on Single-Channel Dimensionality Reduction

In this experiment we fitted Multi-class SVM models to several variations of the original feature vector ($f = 14$) of each Patient-File. We reduced the size of the feature vector using ReliefF, during a pre-processing step prior to classification. We used 3 different measures of performance to analyse the outcome of each reductive step, these are, Accuracy, Sensitivity and Specificity. Values of the performance measures were averaged at each step, over all Patient-Files, excluding the detected outliers. While Accuracy and Specificity remained high throughout each reductive step, Sensitivity appeared to variably fluctuate from step to step. We used the S1-Score as the main comparative performance measure as it is the harmonic mean of both Sensitivity and Specificity.

The ReliefF feature selection indicates an improvement in classification performance for feature subset size of 12 to 6, with the maximum performance at feature-set of size 6. This is consistent with the results produced in chapter 5.

Another interesting point is the overall performance of ReliefF, which is evident of the power of good feature selection. We can clearly see that performance improves at reduced feature-sets (12-6), and even though it decreases at $f = 4$ and minimises at $f = 2$, the discrepancy between $f = 2$ and $f = 14$, which is the full feature-set, is vary low.

The optimum number of features, according to the feature selection method is 6 where accumulated energy holds the highest ranking. The performance remains high for $f>2$, (above 78%). By carefully using a suitable feature selection method, we were

able to reduce the computational time for learning without loosing significant performance. This is a rule, which can be generalised for all patients.

6.3 Experiment II: Dimensionality Reduction on All EEG Channels

In experiment I, we examined the stepwise Dimensionality Reduction conducted on single-channel Patient-Files resulted in better results than the full feature-set. We also saw that performance measures, though variable through different reduction steps, were still considerably high; in some cases higher than the benchmark values. This is indicative of the carrying effects of features on the performance outcome of seizure prediction. In this experiment, we further expand on the results of experiment I, in order to obtain feature rankings on all 84 features, extracted from 6 recorded EEG channels. We expect to see higher feature rankings at some steps, as there are more features to choose from, increasing the likelihood of finding suitable features. In addition, we aim to reveal the effect of a larger feature-set on single-patient classification models.

6.3.1 Methods for Multi-Channel Dimensionality Reduction

In section 6.2, we discussed the preparation procedure of the single-channel Patient-Files. Moreover, we reviewed the extraction process of the 14 features from Patient-Files in chapter 4. For this series of experiments, we extend each Patient-File to include an extended number of features. These extended properties are extracted from the remaining 5 EEG channels, using procedures described in chapter 4. This process expands the size of the feature-set for each patient 6 folds, resulting in an $m \times 84$ feature-vector per patient. Pre-processing and training models of the now extended number of features, demands for greater computational cost. For this reason, we modularised the experiments such that each training block could be constructed on several machines in parallel, effectively reducing the computation time. We used Matlab pooling to run each stage of the ReliefF experiments per patient-module, independent of the other components. These experiments were carried out on a cluster of 8-core, 64bit CentOS machines, with each machine running 8 modules synchronously. The same dimensionality reduction algorithms described in section 6.2.2 of this chapter were used to conduct this set of experiments.

6.3.2 Result of Multi-Channel Dimensionality Reduction

In Figures 6.5-6.6 we can see the results form the stepwise dimensionality reduction experiments on multi-channel Patient-Files. The multi-channel Patient-Files comprise 84 features, as opposed to the 14 features of the default Patient-File datasets. Figure 6.5 displays the summary performance results of the classifiers, averaged across all Patient-Files, reducing 2 low ranked features at a time, $f = f - 2$, by ReliefF feature selection methods. The detailed summary of the results is presented in Table 6.7.

	ACC	f	SP	f	SS	f	S1	f
min	78.47	2	92.23	2	75.61	2	81.78	2
max	94.11	32	99.37	54	92.21	76	94.92	54
f full	93.15	84	98.86	84	88.57	84	92.56	84
f = 2	78.47	2	92.23	2	75.61	2	81.78	2
mean	92.63		98.67		89.25		92.78	
median	93.21		98.95		89.79		93.19	
mode	78.47		92.23		75.61		81.78	
std	2.44		1.12		2.96		2.29	
range	15.64		7.14		16.60		13.14	

Table 6.7 Summary of important data statistics from the stepwise dimensionality reduction on 18 **Multi-Channel** patients using **ReliefF**.

Results reveal that the mean S1-Score oscillates at intervals of 3, 4 and 5. Specificity is high, starting at 98.86%, with a maximum 99.37% at $f = 54$. The measure remains high with oscillations along the steps, displaying variability from step to step, denoting the importance of features involved. The performance oscillates along a steady line until $f = 12$, after which point it declines at a higher rate until it hits the minimum at $f = 2$ at 92.23%.

Accuracy also has oscillations throughout different steps; however, variability is lower compared to Sensitivity, Specificity and S1-Score. Sensitivity displays a great deal of variation among different steps of dimensionality reduction. The S1-Score curve mimics the trend seen in Sensitivity at a higher range due to high values for Specificity. S1-Score starts at 92.56% at $f = 84$, goes through a number of oscillations around a steady line, and peaks at $f = 54$, with 94.92% which is higher than our benchmark value. It gradually decreases, still oscillating around a line, until it hits the minimum at $f = 2$ with 81.88%. The mean value for S1-Score is 92.78% indicating that the majority of the experimental steps have an average result above the mean.

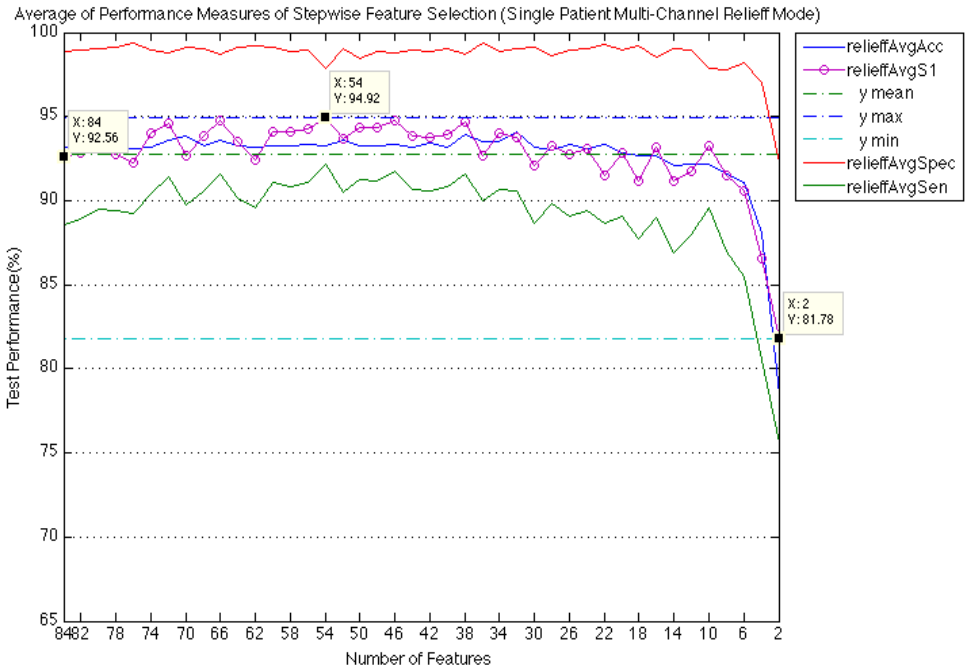


Figure 6.5 Summary of stepwise dimensionality reduction on 18 Multi-Channel patients using ReliefF.

Inter-Patient Variability

In Figure 6.6, we see the box and whisker diagram of the population of Patient-File S1-Scores over the different dimensionality reduction steps of this experiment. In the box and whisker plot we see far less variability among patients (until $f = 8$) as well as low variability among different steps, with $f>8$ being in the >90% range. The boxes in $f \leq 8$, particularly for $f = 2$ and $f = 4$, are much longer, indicating variability in the S1-Score of the Patient-Files, and hence, suggesting inconsistency in the suitability of the experimental setup for patients. Though population is sparse at these points, the boxes are symmetrical, indicative of a normal distribution, with a long left start tail at $f = 4$ and $f = 12$.

For each patient the feature-set with the maximum S1 score was identified and tested against the full feature-set for the same patient using a t-test. The results are presented in Table 6.8. There were no patients whose best performance was produced by the full feature-set (84). The results show that the outcome of the selected best feature-set was significantly better that the full feature-set, only for a small proportion of the patients (8 out of 21). In Table 6.9, the mean S1-score of the full multi-channel feature-set (84) is examined against the mean S1-score of the full single-channel feature-set (14) for all

patients. The results reveal that there are no significant differences between the two feature compositions, suggesting that adding more channels does not significantly improve the performance outcome. In Table 6.10 the mean S1-score of the best feature-set of all patients in the multi-channel setting is examined against the mean S1-score of the best feature-set of all patients in the single-channel setting. The results reveal the best S1-score of the multi-channel experiment is not significantly different from that of the single channel experiment, suggesting that generating 14 features across all channels does not necessarily result in the selection of a more powerful feature-set.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1	0.400	1.443	-0.800	1.265	0.508	9	0.623
Pat 2	2.534	3.140	-3.319	1.173	-1.081	9	0.308
Pat 3	10.650	6.038	6.565	15.203	5.701	9	**0.000**
Pat 4	1.279	9.209	-8.918	4.258	-0.800	9	0.444
Pat 5	1.525	1.750	-0.731	1.773	0.942	9	0.371
Pat 6	1.864	3.744	2.589	7.946	4.449	9	**0.002**
Pat 7	2.011	2.272	0.528	3.779	2.997	9	**0.015**
Pat 8	7.503	13.288	-10.232	8.779	-0.173	9	0.867
Pat 9	2.051	2.612	-0.782	2.955	1.315	9	0.221
Pat 10	6.168	3.722	11.941	17.266	12.407	9	**0.000**
Pat 11	0.036	2.399	-3.084	0.348	-1.803	9	0.105
Pat 12	2.864	3.673	-4.637	0.618	-1.730	9	0.118
Pat 13	3.142	2.798	-4.222	-0.218	-2.509	9	**0.033**
Pat 14	1.805	1.703	-1.300	1.136	-0.152	9	0.882
Pat 15	0.777	3.106	-16.493	-12.049	-14.531	9	**0.000**
Pat 16	0.858	1.432	-1.871	0.178	-1.869	9	**0.094**
Pat 17	5.939	2.962	4.551	8.789	7.122	9	**0.000**
Pat 18	0.704	1.850	-0.518	2.129	1.377	9	0.202
Pat 19	0.542	10.939	-12.741	2.909	-1.421	9	0.189
Pat 20	1.203	3.526	-2.874	2.170	-0.316	9	0.759
Pat 21	0.518	1.548	-0.625	1.590	0.985	9	0.350

Table 6.8 Each row represents a t-test for each Multi-channel patient. The t-test compares the S1-score of the 10 runs of the full feature-set against the best feature-set. The best feature-set was determined by finding the feature-subsets with the highest S1-score averaged along 10 runs. The best feature subset is different for each patient. The values in bold indicate $p \leq .05$ and are considered statistically significant.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1:21	3.359	14.900	-10.142	3.423	-1.033	20	0.314

Table 6.9 The mean S1-score of the full multi-channel feature-set of all patients is examined against the mean S1-score of the full single-channel feature-set, using a paired t-test.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1:21	0.502	8.020	-3.149	4.152	0.287	20	0.777

Table 6.10 The best multi-channel mean S1-score of all patients is examined against the best single channel mean S1-score, using a paired t-test.

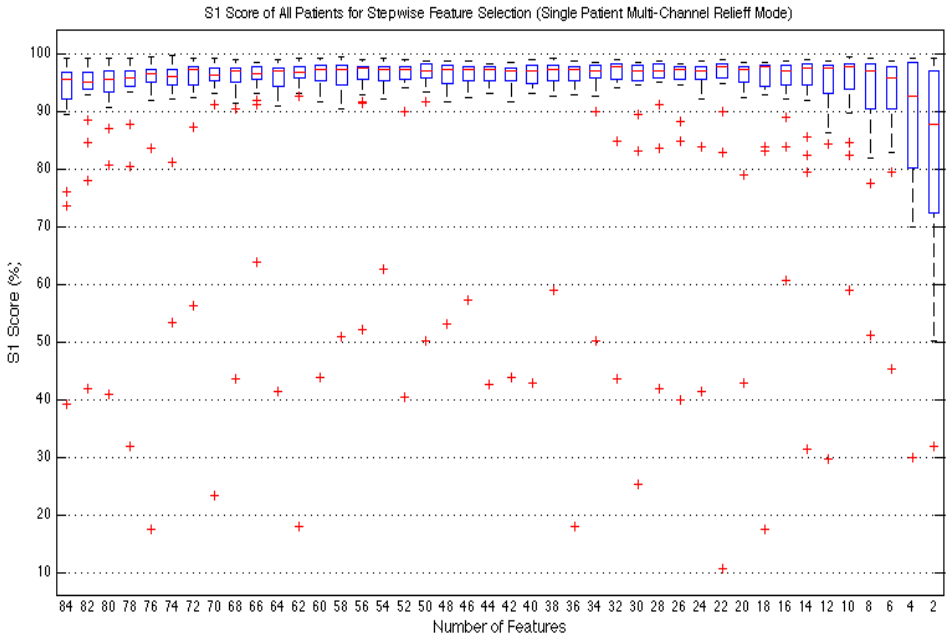


Figure 6.6 The Box and Whiskers diagram of stepwise dimensionality reduction on 21 Multi-Channel patients using ReliefF.

6.3.3 Multi-Channel Feature Selection and Rankings of Features

In chapter 5, we established the importance of features in terms of discriminatory power for features of rankings ≤ 8 (see section 5.4). In experiment I of this chapter (section 6.3), we further evaluated this hypothesis by modifying experimental conditions to include all Patient-Files. We identified the discriminatory power of the top 6 features, on the mean performance measure of stepwise dimensionality reduction.

In this experiment, we expanded our dimensionality reduction experiment to include all 6 channels for each of the Patient-Files from the Freiburg EEG Database. This resulted in 84 total features per patient. In Table 6.11 we have included the ranking table averaged across all Patient-Files, in ascending order of feature ranks. The first column corresponds to the feature number (1−84). The Feature Name column indicates the type of feature from a list of features listed in Table 6.1, regardless of the channel the feature is extracted from. This can help identify clusters of features that are higher up in the ranking table. The channel column holds the channels each feature comes from. For instance, feature 22 corresponds to the Short Term Energy in 12.5Hz − 25Hz frequency band (STE 2) incoming from EEG channel 2. Identifying channels of each feature ranking will be useful when identifying the most predictive channel per patient. The Average Rank column holds the mean rank of the feature across the entire population of patients.

In the rankings produced by ReliefF, we can see that the two non-linear features have the lowest mean rankings over all channels. The lowest ranked features of Energy STE and Energy Level which also had the lowest rank in the single-channel experiment continue to hold the lowest ranking in the multi-channel experiment, regardless of the channel they are extracted from. This is while the signals received from the focal electrodes contain more prominent seizure activity than those produced by extra-focal (Figure 6.7) channels. This suggests that regardless of how useful a channel is in terms of discrimination power, it is still not as discriminatory as the type of the feature engineered from that channel. In other words, feature type, overrules seizure channel. This reveals a robustness in the rankings made by the feature selection methods and that the rankings produced in the single-channel experiment are not arbitrary. In fact, the lowest ranking features remain the same across several channels from several patients.

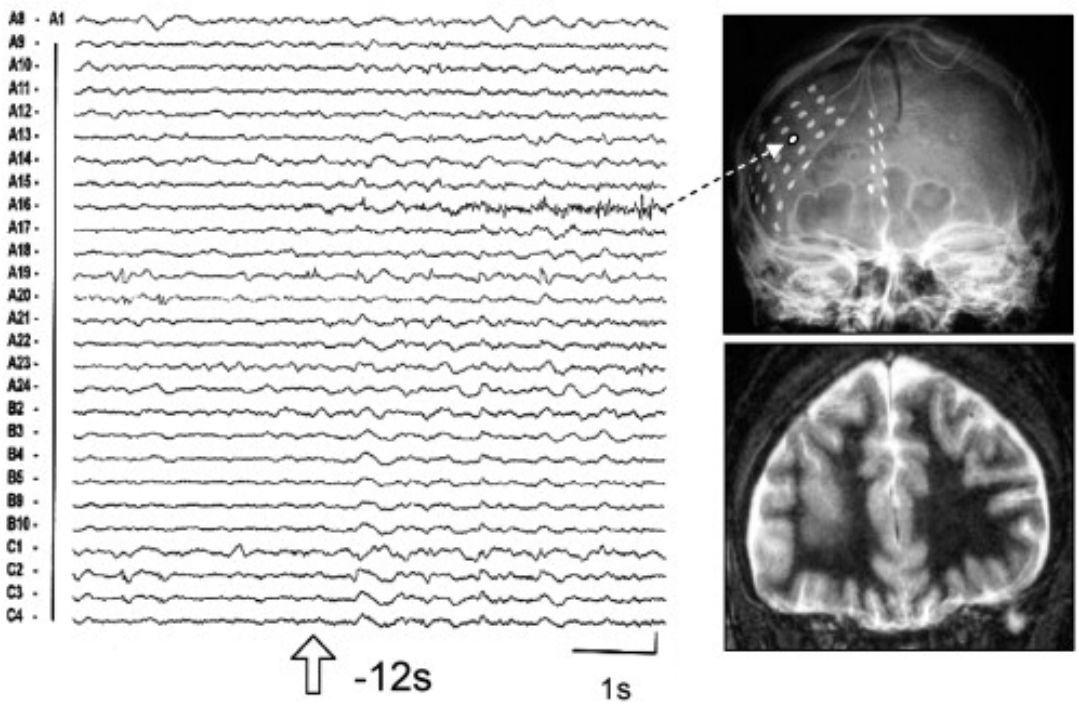


Figure 6.7 Invasive EEG of a patient with Focal Epilepsy: Abnormal activity is most prominent in the seizure onset location distinguished with dashed line (Noachtar & Rémi 2009).

Feature No.	Feature Name	Channel	Avg. Rank	Standard dev.
22	STE 2	2	25.29	±19.26
29	Accumulated energy	3	25.29	±22.13
43	Accumulated energy	4	25.91	±22.95
8	STE 2	1	26.95	±18.89
15	Accumulated energy	2	27.00	±25.08
23	STE 3	2	28.48	±21.40
46	Energy LTE	4	28.91	±21.33
24	STE 4	2	29.48	±19.55
32	Energy LTE	3	30.29	±20.52
1	Accumulated energy	1	30.67	±26.00
9	STE 3	1	30.86	±20.58
25	LTE 1	2	30.91	±17.58
21	STE 1	2	30.95	±18.43
71	Accumulated energy	6	31.43	±23.90

Table 6.11 Properties of 14 features from the 84 Multi-Channel feature-set that hold the highest rankings according to ReliefF feature selection method. The rankings were averaged across all 21 patients from the Freiburg EEG database. Listed are the feature number in the dataset, the feature name irrespective of channel, the channel the feature was obtained from, the average ranking and the standard deviation of the rank. The full ranking table is listed in Appendix A.

The highest ranked features for ReliefF are STE 2, Accumulated Energy, STE 3, Energy Variation LTE, STE 4, LTE1 and STE1. The higher rankings seem to be mostly from the focal channels. As the rankings decrease, the population mostly comprises extra-focal channels (see Appendix A).

6.3.4 Discussion on Multi-Channel Dimensionality Reduction

In this experiment, we extended our 14-feature single-channel feature-set to incorporate EEG recordings of all 6 channels, comprising 3 focal channels and 3 extra-focal channels. The focal channels are those that are nearest to the seizure foci of the brain (see sections 3.1.1 and 3.2.1) and are expected to display stronger seizure signals than those of the extra-focal channels, due to their close proximity to the seizure onset region. We fitted several Multi-class SVMs to each Patient-File and monitored the performance as we reduced the number of features 2 at a time, using the ReliefF feature selection methods.

We can see that the feature selection method performs well for f>14, maintaining a high performance for as few as 6 features. The extended number of features produced high performance outcomes, but the output does not significantly vary from the single-channel experiment, with the exception of the smaller feature subsets. This may be due to the fact that the focus channel is a high-resolution representation of the seizure state of the brain and yields the most discrimination between the seizure and non-seizure states. The extra-focal channels however, may produce useful information about the brain in general, but discriminate less between the seizure and non-seizure states, potentially introducing unnecessary noisy and correlated data. The extra channels may however prove to be useful when used in multi-patient models, where they could provide data to further generalise the dataset.

The mean ranking tables revealed that the least important features are mainly of the same type, extracted from all 6 channels, indicating that the intrinsic characteristic of a feature most frequently overrules the locality of the channel it is coming from.

6.4 Experiment III: Extension to Feature-set and Dimensionality reduction

In experiment I, we used the feature selection algorithm in identifying the ranking of features for each Patient-File based on the intrinsic attributes of the data; ReliefF ranks

the features based on conditional dependencies between features. After a generic analysis was conducted over the ranking tables produced for all patients, the top 8 ranking features were identified which heuristically led us to developing an extended feature-set. This new extension contains features that are related to the top ranked features from experiments I and II. This section covers the steps taken in determining the new features, the description of the extended feature-set and an experiment evaluating the effects of this new feature-set in a dimensionality reduction setup, ranging from the full feature-set down to a feature subset containing as few as two elements.

6.4.1 Extending The Feature-set

As seen earlier in section 6.3, one of the most interesting outcomes of these experiments was the ranking table produced for each multi-channel Patient-File. Drawing from the feature ranking tables constructed in experiment I and II, we presented the 14 highly ranked features averaged across all patients, to establish the representative top 14 features produced by ReliefF.

We used the ranking tables from experiment II independent of the channels they were extracted from, in order to explore the possibilities of extending the feature-set. We saw that signal energy features produced the highest outcome. We will not expand the non-linear dynamics features as i) they are computationally expensive and take longer to compute compared to uni-variate linear methods that can be computed in real-time ii) we chose to expand on the highly ranked features of ReliefF and since non-linear features were ranked poorly by ReliefF, we do not expand on non-linear features.

We used this information to extract a further 20 features from each EEG channel, adding up to 20×6 features for each of the patients in the dataset, a list of which is presented in Table 6.12.

In our single-channel and multi-channel study of dimensionality reduction, we saw that accumulated energy had the highest average ranking. Signal Energy was also ranked highly by ReliefF in both single-channel and multi-channel analyses. The results from our study, in line with results of previous literature (Mormann et al. 2005), suggest that linear features, particularly of the signal energy category, have performed well in other experimental settings. Therefore, we decided to expand our feature-set to incorporate more features of this category.

Feature Type	*Features*
Spectral Edge Frequency (SEF)	SEF STE SEF LTE SEF STE Median SEF LTE Median
Statistical Moments	Mean STE Mean LTE Skewness STE Skewness LTE Kurtosis STE Kurtosis LTE
Spectral Band Power (SBP)	SBP STE Delta SBP STET Theta SBP STE Alpha SBP STE Beta SBP STE Gamma SBP LTE Delta SBP LTE Theta SBP LTE Alpha SBP LTE Beta SBP LTE Gamma

Table 6.12 List of 20 extended features – The features were heuristically obtained based on the performance of previous feature selection experiments. The Feature Type column lists the general category of the features and the Feature column lists the respective features derived in each category. The 20 features are extracted across all 6 channels resulting in a total of 120 new features.

Statistical moments

The statistical moment is a quantitative measure, which characterises the distribution of a set of points (Mormann et al. 2007). In analysing time series $\{x_i\}$, statistical moments provide various representations of the amplitude values of the time series. There are four statistical moments:

Mean of a signal {xi} is the **first** moment, denoted with

$$\mu = \frac{1}{N-1}\sum_{i=1}^{N} x_i \qquad (2)$$

Variance is the **second** moment;

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 \tag{3}$$

Skewness of a distribution is the **third** central moment. The skewness of a symmetric amplitude distribution is 0. When the distribution has a heavier tail to the left, the skewness is negative and when the tail is heavier on the right, the skewness is positive. Skewness of $\{x_i\}$ is calculated as:

$$\chi = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{\sigma}\right)^3 \tag{4}$$

Kurtosis is the **fourth** central moment, which measures the relative peakness or flatness of an amplitude distribution. The kurtosis is always strictly positive due to the power of 4 in the expectation.

$$\kappa = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i}{\sigma}\right)^4 - 3 \tag{5}$$

The second statistical moment (i.e. Variance) was part of the original feature-set (see section 4.2.3), which received high ratings particularly from ReliefF feature selection experiments. We calculated the remaining three statistical moments over a moving window analysis of two varying sliding windows: Short-Term Energy (STE) window of 9-second length and Long-Term Energy (LTE) window of 180-second length. The moments were calculated for 5-second epochs.

Spectral Band Power

In studies of neuronal signals, particular attention is dedicated to the analysis of 5 main frequency ranges of the power spectrum, where useful information of the signal is captured (Mormann et al. 2005). Associated with these frequency ranges are the power spectral bands of $\alpha, \beta, \gamma, \delta, \vartheta$ that have been used in classic EEG Analysis:

$$\delta = \frac{1}{P}\sum_{f=0.5Hz}^{4Hz} p_f \quad (6) \qquad \vartheta = \frac{1}{P}\sum_{f=4Hz}^{8Hz} p_f \quad (7) \qquad \alpha = \frac{1}{P}\sum_{f=8Hz}^{13Hz} p_f \quad (8)$$

$$\beta = \frac{1}{P}\sum_{f=13Hz}^{30Hz} p_f \quad (9) \qquad \gamma = \frac{1}{P}\sum_{f=30Hz}^{48Hz} p_f \quad (10)$$

where P is the total power of the signal.

Spectral Edge Frequency

The Spectral Edge Frequency (Stanski et al. 1984) is a measure that characterises the power distribution, which is mostly distributed in the frequency range of 0 – 40 Hz. This measure is defined as the minimum frequency up to which, 50% of the spectral power within this frequency band is contained in the signal:

$$f_{50} = \min\left\{ f \left| \sum_{v=0.5Hz}^{f} p_v > P_{40Hz} \bullet 0.50 \right. \right\} \qquad (11)$$

6.4.2 Methods of Dimensionality Reduction on The Extended Feature-set

The extended feature-set developed and presented in this chapter, is used as the new dataset for this experiment. From each channel, a further 20 features were extracted, increasing the number of features per channel to 34, and growing the feature-matrix to incorporate $m \times 204$ properties. This increase in the size of the feature matrix requires additional computation time and power. The experiment was distributed over a cluster of 8 core 64bit CentOS machines, with each machine in the cluster running 8 Matlab pools in parallel. By modulating the experiment in several smaller workloads, with single runs of training and classification per patient in each module, every classification model was only implemented in one thread, hence eliminating the need to distribute the model construction over different workstations.

6.4.3 Results of Dimensionality Reduction on Extended Feature-set

In Figures 6.8 – 6.9 we see the outcome of the stepwise dimensionality reduction phase, summarised over the classifiers trained on Patient-Files comprising a combination of the multi-channel features-set as well as the extended feature-set, described in the previous section.

In Figure 6.8 and Table 6.13, we present the outcome of the experiment using ReliefF as the underlying feature selection method. As we can see, the overall performance of all 4 measures has improved compared to that of the single-channel and multi-channel features. Specificity has an extremely high value throughout the steps, with maximum value of 99.68%, mean of 99.51% and standard deviation 0.70%. Sensitivity however gradually and monotonically decreases for f <16, until the

minimum of 92.54% is hit at $f = 2$. The value of Accuracy is initially quite high at 94.35% for $f = 204$, after which point it gradually increases, with dips and peaks along the way, hitting a maximum of 98.3%, maintaining a mean of 96.58% and standard deviation of 1.93%. The Accuracy value gradually decreases after $f = 28$, and for $f < 10$ the value decreases at a higher rate, until it hits the minimum of 80.81% at $f = 2$. Both S1-Score and Sensitivity display a similar behaviour to that of Accuracy with the overall trend of gradual increase, followed by sudden decrease down to $f = 2$, but with greater variability among different steps.

	ACC	f	SP	f	SS	f	S1	f
min	80.81	2	92.54	2	71.48	2	78.63	2
max	98.3	38	99.68	50	95.3	190	97.38	50
f full	94.38	204	99.6	204	82.57	204	89.41	216
f = 2	80.81	2	92.54	2	71.48	2	78.63	2
mean	96.58		99.51		90.5		94.36	
median	96.73		99.63		91.06		94.82	
mode	80.81		92.54		71.48		78.63	
std	1.93		0.71		3.99		2.85	
range	17.49		7.14		23.82		18.75	

Table 6.13 Summary of important data statistics from the stepwise **dimensionality reduction** on the 18 patients with extended **204-dimensional** feature-set using **ReliefF**.

The value for S1-Score starts at $f = 204$ with 89.41% and gradually increases and hits the maximum at $f = 50$ with 97.38%, at which point there are no more increases; the value is maintained for a number of steps with less variation. After $f = 28$, values start to increase gradually, and for $f < 16$ values drop monotonically until the minimum is hit at $f = 2$ with 78.63%. The S1-Score has mean value of 94.35% with a standard deviation of 2.84%. For $14 \leq f \leq 134$ S1-Scores are above the mean and the remaining steps are below the mean. Sensitivity is in the range [71.48%, 95.30%] with mean 90.5% and standard deviation 3.98%, indicative of high performance with little variability among different steps.

Inter-Patient Variability

In Figure 6.13, we see the box and whisker diagrams of the S1-Score of classifiers trained on Patient-Files with extended feature-sets, against the stepwise reduction of features using the ReliefF algorithm. We observe less variability for $8 \leq f \leq 190$, which indicates that performance is consistently high among all Patient-Files. In the

cases where we have longer boxes, the variation is relatively low. The majority of the boxes in the range $16 \leq f \leq 200$ display low variation among different steps within this range. The plot displays the previously established set of outliers (patient 3, 10 and 14) and some new ones such as patients 18, 8, 1 and 2, where seemingly the high number of features does not produce the same high values for these particular patients. However, the S1-Scores produced are still high and the same gradual increase in the values can be observed among the outliers. For $14 \leq f \leq 54$, the outliers have higher values than other steps. Patient-File 8 in particular responds poorly to the reduction of features to f <18, with values as low as 0.58%. This shows once again that the effect of dimensionality reduction on classification is patient-specific; certain features can enhance the classification for some Patient-Files while worsening the performance in others.

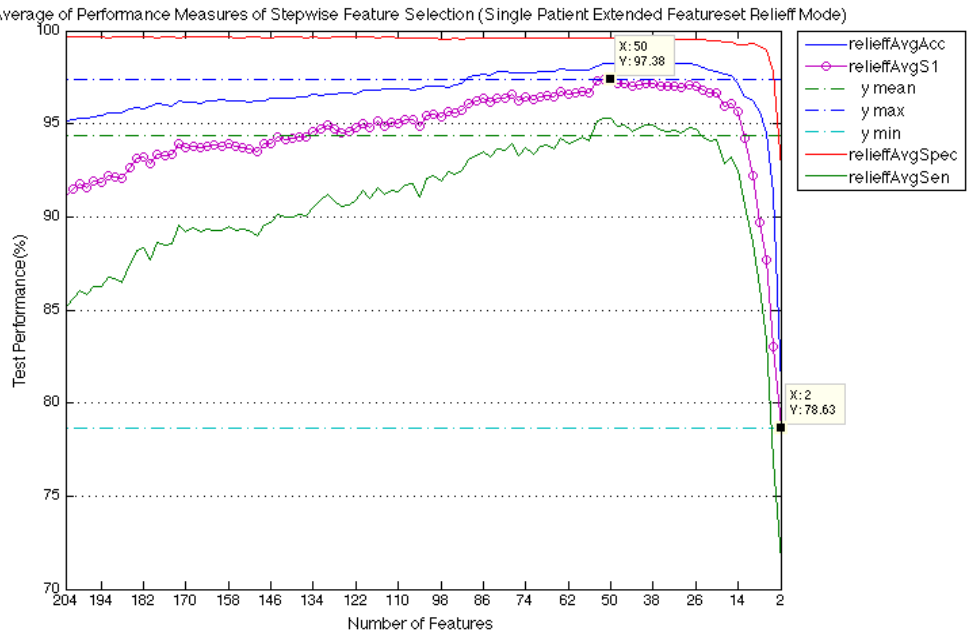


Figure 6.8 Summary of stepwise dimensionality reduction on 18 **Multi-Channel Extended Feature-Set** patients using **ReliefF**.

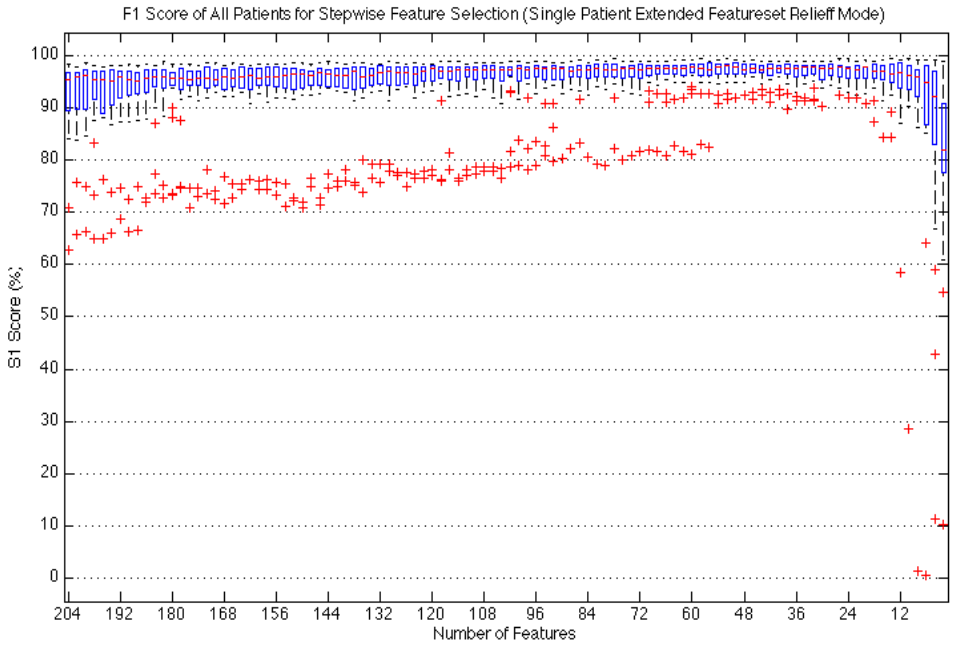


Figure 6.9 The Box and Whiskers diagram of stepwise dimensionality reduction on 21 **Multi-Channel Extended Feature-Set** patients using **ReliefF**.

For each patient the feature-set with the maximum S1 score was identified and tested against the full feature-set for the same patient using a t-test. The results are presented in Table 6.14. There were no patients whose best performance was produced by the full feature-set (204). The results show that the outcome of the selected best feature-set was significantly better that the full feature-set for a greater proportion of the patients (13 out of 21) in contrast to the multi-channel experiment. In Table 6.15, the mean S1-score of the full multi-channel feature-set (204) is examined against the mean S1-score of the full single-channel feature-set (14) for all patients. The results reveal that there are no significant differences between the two feature compositions. This was expected as the full extended feature-set in itself does not have an impact on performance outcome, though combined with feature reduction steps, an improved outcome is expected. In Table 6.16 the mean S1-score of the best feature-set of all patients in the extended-feature setting is examined against the mean S1-score of the best feature-set of all patients in the single-channel setting. The results reveal the best S1-score of the extended features-set is indeed significantly different from that of the single-channel experiment ($p=0.019$), suggesting that a subset of features from the extended feature-set has the edge over the best single-channel feature-set.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1	10.755	1.844	8.216	10.854	16.355	9	**0.000**
Pat 2	31.559	5.913	24.468	32.928	15.347	9	**0.000**
Pat 3	2.627	3.198	-4.000	0.575	-1.694	9	0.125
Pat 4	0.350	2.160	-0.677	2.413	1.271	9	0.236
Pat 5	3.305	1.616	3.139	5.451	8.405	9	**0.000**
Pat 6	15.373	2.726	-3.415	0.485	-1.699	9	0.123
Pat 7	2.374	1.611	-0.106	2.199	2.054	9	0.070
Pat 8	3.183	6.036	-9.195	-0.560	-2.556	9	**0.031**
Pat 9	2.734	2.326	1.263	4.590	3.979	9	**0.003**
Pat 10	2.121	2.240	0.766	3.970	3.344	9	**0.009**
Pat 11	1.838	3.786	-12.129	-6.713	-7.869	9	**0.000**
Pat 12	1.997	3.266	-2.556	2.117	-0.213	9	0.836
Pat 13	27.813	7.293	22.485	32.920	12.011	9	**0.000**
Pat 14	6.863	2.904	2.927	7.082	5.449	9	**0.000**
Pat 15	1.384	2.334	-0.543	2.796	1.526	9	0.161
Pat 16	4.051	3.166	-1.076	3.453	1.187	9	0.266
Pat 17	4.416	2.214	4.731	7.898	9.021	9	**0.000**
Pat 18	10.365	2.859	13.206	17.297	16.870	9	**0.000**
Pat 19	7.842	2.434	-1.512	1.971	0.298	9	0.772
Pat 20	0.802	0.976	-2.857	-1.461	-6.999	9	**0.000**
Pat 21	1.424	2.387	-4.778	-1.362	-4.067	9	**0.003**

Table 6.14 Each row represents a t-test for each extended-feature-set patient. The t-test compares the S1-score of the 10 runs of the full feature-set against the best feature-set. The best feature-set was determined by finding the feature-subsets with the highest S1-score averaged along 10 runs. The best feature subset is different for each patient. The values in bold indicate $p \leq .05$ and are considered statistically significant.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1:21	0.582	14.575	-7.217	6.052	-0.183	20	0.857

Table 6.15 The mean S1-score of the full extended feature-set of all patients is examined against the mean S1-score of the full single-channel feature-set, using a paired t-test.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
Pat 1:21	0.950	1.712	-1.729	-0.170	-2.542	20	**0.019**

Table 6.16 The best extended-feature mean S1-score of all patients is examined against the best single channel mean S1-score, using a paired t-test.

6.4.4 Rankings for Features of the Extended Feature-set

At the start of this section, we introduced 120 new features, (20 across 6 channels). We used the new features along with the old features in a stepwise dimensionality reduction experiment presented earlier in this section. From the results, we saw that Specificity remained exceptionally high throughout reduction stages while the three measures of Accuracy, Sensitivity and S1-Score started at lower values and gradually increased to above the mean, at varying reduction steps, where they remained notably steady until the decline towards the later reduction steps (generally $f \leq 8$), and hitting the minimum at $f = 2$. The results clearly indicated an improvement in the performance measures as features were reduced (mean generally around $f = 160$). We present the feature rankings produced by ReliefF on the 204 dimensional feature-set in Table 6.17.

The feature-set was expanded by 257.14% from 84 to 204 features per Patient-File. In the previous section, we saw that the S1-Score for rises above the mean at $f = 144$. According to the ranking tables, out of the 120 new features introduced, 31% were in the lower range (below the mean) and 69% were in the higher range (above the mean), where performance was steadily high. Out of the 84 original features, 40% were in the lower range and 63% were in the higher range. This indicates that a greater percentage of the new features are in the high performance range while a greater percentage of the original feature-set appear in the lower performance range. The proportion of new features and original features are quite similar but the new feature-set seems to have the edge.

Weighted Frequency of High-Rank Features

In Figure 6.10 we present the weighted frequency of the top 84 features determined by the ReliefF algorithm. For each general feature (independent of the channel), the

ranking power of each feature was treated as the weight of that instance of the general feature. Therefore, for each feature we have

$$b = \sum_{j=1}^{i} \left(f_j \times weight \right) \qquad (12)$$

where b is the weighted frequency, j is a value from 1 through 84, f is the frequency of the feature and *weight* is the ranking-power of the respective feature.

From the graph it is clear that the number of new features in the top 84 ReliefF experiment is higher than the number of original features. The highest rankings correspond to SEF LTE, Accumulated Energy, SBP LTE Delta and SPB LTE Theta. In ReliefF, the top 6 features are Spectral Edge Frequency (Long term Energy window) from all 6 incoming channels, followed by accumulated energy, Signal Band Power (Long term Energy) from the Delta and Theta bands.

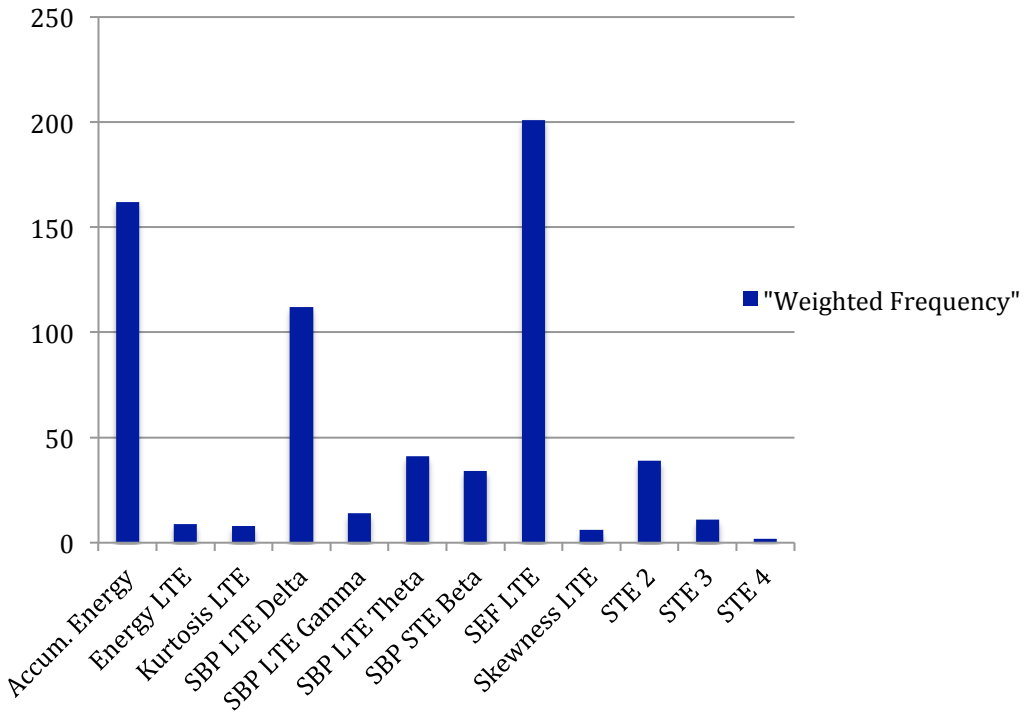


Figure 6.10 Weighted Frequency of top 84 **ReliefF** features – The weighted frequency of features within the top 84 range of the ranking table produced by ReliefF feature selection were calculated based on their ranking and frequency, independent of the channel they came from.

6.4.5 Discussion on Feature Extension and Dimensionality Reduction

In this experiment we expanded our feature-set heuristically and based on the outcomes of experiment I and II of this chapter. We identified the signal energy feature category as the most powerful among the feature categories. We further expanded our feature-set to include an additional 20 features per channel, adding up to 204 features in total across all channels. The new features were uni-variate signal measures. These were Spectral Band Frequency over 5 frequency band of Delta (0.1 – 4 Hz), Theta (8 – 15 Hz) Alpha (8 – 15 Hz), Beta (15 – 30 Hz) and Gamma (30 – 100 Hz), Spectral Edge Frequency and Statistical Moments of Mean, Skewness and Kurtosis. These features were calculated in Short Term Energy window (STE) and Long Term Energy window (LTE) in a moving window analysis. The new dataset, comprising 84 original features and 120 new features was then used in a stepwise dimensionality reduction experiment, where the ReliefF feature selection method was used in a pre-processing step to rank the features, according to which 2 of the lowest ranking features were removed from the dataset at each step. This resulted in the creation of 102 feature subsets and classification models per Patient-File. The results of this were presented in section 6.4.3. Specificity remained high throughout the reduction steps, while Accuracy, Sensitivity and S1-Score seemingly improved by reducing features to an optimum number. The relatively poorer performance in the higher number of features suggests a degree of variance introduced by some of the additional features, resulting in an over-fitted classification. This problem was resolved for ReliefF at $8 < f \leq 144$. The reduction of features at $f = 8$ produced lower than mean S1-Scores which were relatively high, after which point, for $2 \leq f \leq 6$ the value monotonically and rapidly dropped. This is similar to what we have seen so far in previous experiments of this chapter, where performance drops for reductions of $f < 8$. The variability of S1-Score across patients at the higher performance steps was very low, indicating the high performance range of features yielded high performance for all Patient-Files.

We also showed that the newer features had a higher weighted frequency in the top 84 features. The variations of Accumulated Energy still remained high in the rankings, confirming the importance of this feature. Spectral Band Power and Spectral Edge Frequency features were also amongst the highest rankings. The high rankings produced for these top features and the high performance outcome of stepwise feature-

reduced classification, proves the discriminatory power of these features in classifying seizure states across all patients.

The rankings also revealed that on average, a higher percentage of new features comprised the high performance range ($f \leq 144$) and a lower percentage was in the low-performance range, where S1-Score was below the mean. The percentage of original features in the high-performance range is lower compared to that of the newer features, indicating that the newer features have an edge over the original features. This is further backed up with the classification conducted on various feature subsets created from such feature rankings.

From the results we can additionally conclude that, since performance is maintained throughout the high-performance range ($8 \leq f \leq 144$ for ReliefF), that with as few as 8-10 carefully-selected features (amongst which are a blend of original and new features), using ReliefF feature selection methods, we are able to maintain the high performance for most Patient-Files.

6.5 General Discussion on Feature Selection and Dimensionality Reduction

This chapter provided a comprehensive series of experiments, which evaluated the performance of our seizure detection algorithm on several feature-sets. The results revealed that Accuracy and Specificity were consistently high with little variation amongst experiments, with the exception of feature-set $f = 2$. Sensitivity and S1-Score however varied across experiments and feature subsets.

Several feature subsets of the various experiments yielded a performance higher than the benchmark value, which was the mean S1-Score of the original feature-set. This is represented in Figure 6.11 with a solid red line.

The results of the single-channel and multi-channel analysis were used to heuristically expand the feature-set to incorporate an additional 20 features per EEG channel. The results reveal that S1-Score clearly improves with the new feature-set but only for smaller subsets and not for the full range of features. This was expected as the full feature-set may have introduced redundancy and noise, besides providing numerous discriminatory features from which the feature selection methods can choose from, resulting in a better feature selection outcome. Performance is above the benchmark for

feature subsets of [8, 144] for ReliefF. The highest S1-Score is attained at $f = 50$ with 97.38%.

The analysis of the high-performance range of the extended features of ReliefF also revealed that the range was densely populated with elements of the extended feature-set, in particular Spectral Band Power and Spectral Edge Frequency. From the original feature-set, Accumulated Energy was still highly ranked. Both multi-channel and extended feature-set experiments revealed that the quality of the feature overrules the locality of the channel, particularly in higher rankings; this implies that, with high quality features, it is worthwhile to use all channels (focal and extra-focal) as it provides higher resolution of the respective powerful feature.

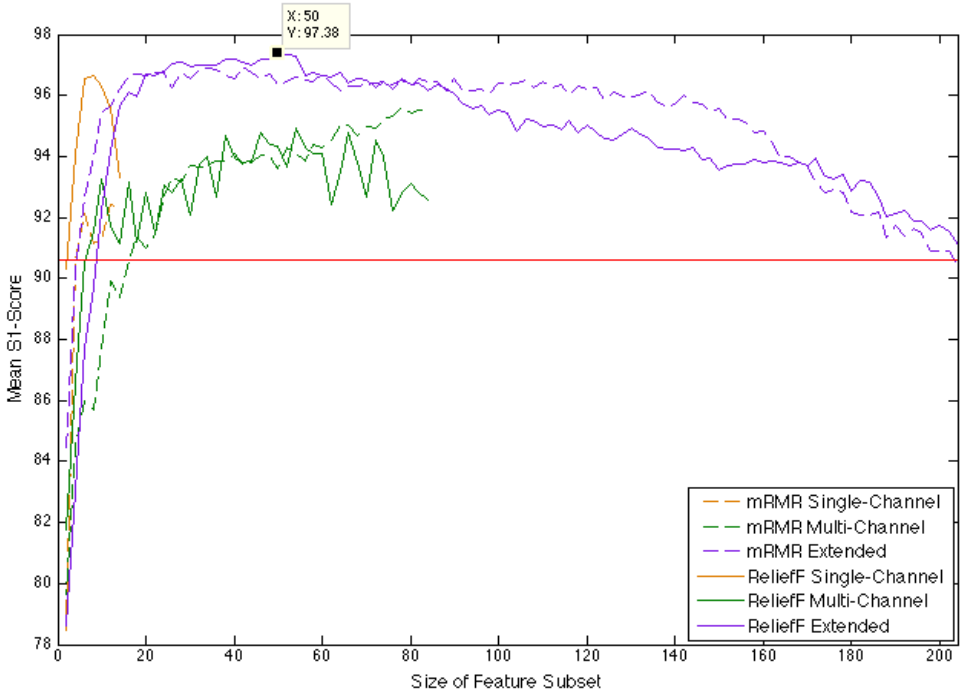


Figure 6.11 Summary of all dimensionality reduction experiments conducted on 18 patients form the Freiburg EEG Database – The reported results are for average S1-Score through respective reduction steps of each experiment. The plots with the same color represent the same feature-set variation: The dashed line marks experiments using mRMR while the solid line represents experiments based on ReliefF. The solid red line is the benchmark value for S1-Score (14 feature default channel).

6.6 Conclusion

This chapter has compared and contrasted the effects of using various feature settings, on the performance of seizure detection algorithms. We were able to show that by expanding the feature-set to include multiple channel recordings and several new features, after performing feature selection, we are able to attain higher performance measures than the benchmark, yielding S1-Score values as high as 97.38%.

We also observed that the highest performance outcome was obtained from a subset of the extended feature-set introduced in section 6.4. The single-channel experiments produced the second best performance, while the multi-channel feature-set led to the worst performance outcome.

The extended feature-set will later be used in parts of chapter 7 and chapter 8, in order to improve the results of certain experiments as well as to further prove the enhanced quality of the extended feature-set over the original feature-set.

Chapter 7
Predicting Epileptic Seizures in Advance

In this chapter, we evaluate the predictability of the epileptic brain under various experimental conditions. In chapter 5, we presented results of preliminary seizure prediction experiments in a 10-minute time frame on a single patient from the Freiburg EEG Database. The results revealed that, using MC-SVM and EANN, we are able to achieve improved evaluation measures, such as Accuracy, Specificity and Sensitivity at 6-8 minutes prior to the seizure onset. The promising results of this experiment are indicative of significantly predictable seizure patterns prior the onset of a patient's seizure.

This chapter aims to further expand the preliminary seizure prediction experiment, in order to better evaluate the possibility of accurate advance prediction under various experimental conditions, including the use of more patients and a modified feature-set. The main questions we want to answer is whether predictability is a general property of all seizure types, whether the time-frame in which a seizure is predicted bears any meaningful information and more importantly, whether advance predictability can be improved under different experimental conditions.

This chapter presents **4 advance prediction experiments**, all of which are carried out in a single-patient mode, meaning that the predictive models are trained and tested on individual patients. In section **7.1** we **motivate** predicting epileptic seizures advance the onset. In **experiment I** we analyse the performance measures of classifiers trained on patient-specific **single-channel** data. **Experiment II** further expands the feature-set in order to accommodate linear and non-linear measures from all **six recorded channels**. In **experiment III** we assess the performance of predictions made on an **extended feature-set**. **Experiment IV** performs prediction experiments on a smaller **subset of the extended feature-set** introduced in chapter 6. In section **7.6**, **implications** from these experiments are **discussed** and we **conclude** in section **7.7**.

7.1 Motivation

As described in detail in chapter 3 (section 3.3), seizure detection and prediction from EEG recordings has been the focus of research in this field. The unpredictable nature of

seizures can impose potential risks for the individual with epilepsy. Therefore, the automatic detection of the oncoming seizures, shortly before the actual onset, can give rise to timely intervention, minimizing these risks. In this chapter, the focus is on **advance seizure prediction**. In other words, the detection of the occurrence of a seizure, more than 5 minutes in advance of the actual seizure onset. Seizure pre-cursors were found 6 second prior the onset (Rogowski et al. 1981; Salant et al. 1998) and 1-minute prior the onset (Siegel et al. 1982) in early studies of seizure prediction. (Mormann et al. 2005) evaluated 30 features in various prediction windows of length 30 to 240 minutes in advance of seizure onset. These early studies statistically prove the existence of seizure markers in advance of the actual onset, but have failed to predict seizures for new and unseen EEG data. Later algorithmic approaches produced 68% Sensitivity and 0.15 false positive rate 72 minutes prior seizure onset (Chaovalitwongse et al. 2005) trained and tested on 3-14 day recordings of 10 patients. (Iasemidis et al. 2005) predicted seizure activity, 78 minutes in advance with 91% Sensitivity and 0.15 false positive rate, on the continuous EEG of only two patients which is a small population of patients, and therefore, findings are not deemed conclusive. More over, the Sensitivity and Specificity of both of these studies are low and are therefore not suitable for clinical application.

The correct implementation of advance seizure prediction could change the course of therapeutic plans to incorporate immediate medical strategies (Morrell 2006; Stein et al. 2000). In addition to enabling potential medical advancements in Anti-Epileptic Therapy, predicting seizures can enrich our understanding of the epileptic brain: why and how seizures occur. Successful and accurate advance prediction could further confirm previous findings that the occurrence of a seizure is more than a mere burst of energy and is in fact the result of neuronal activity unfolding through time.

7.2 Experiment I: Advance Seizure Prediction on Single-Channel EEG

The aim of this experiment is to assess the predictability of all patients from the Freiburg EEG Database based on a single-channel feature-set, over an extended time-frame compared to the one presented in the preliminary experiments in chapter 5.

7.2.1 Methods for Single-Channel Advance Seizure Prediction

In this section, we present the methods used in data preparation and implementation of our experiment.

Data Preparation for Advance Seizure Prediction on Single-Channel Data

The data used in this experiment are derived from the Freiburg EEG Database (Epilepsy.uni-freiburg.de 2007). The ASCII files in the dataset were prepared according to steps mentioned in chapter 4. The preparation steps resulted in 21 Excel and Matlab Patient-Files, each of which contained 14 extracted features (Table 6.1), a timestamp for each 5 second blocks of data and a state label indicating the ictal status of the brain. The state label can take values of 1 through 4, which respectively represent inter-ictal, pre-ictal, ictal and post-ictal states of the brain. Each prepared file holds 1 hour of data per seizure, comprising the labeled data. The Patient-Files contain 1 to 5 seizure recordings for each patient.

Implementation of Advance Seizure Prediction for Single-Channel Data

The advance prediction experiment is composed of a series of segregated experiments conducted on each individual Patient-Feature file. The pre-processed Patient-File is split into 70% training-set and 30% test-set using a random seed; this happens in the pre-processing module described in section 6.2.1 and as depicted in Figure 7.1. The normalized training-set and test-set are fed into the Learning Module (Figure 7.1) similar to steps described in 6.2.1, where a separate training model is built for that corresponding training data.

After carrying out parameter selection and constructing the Multi-class SVM following steps previously described in section 6.2.1, the trained model is then tested on the unseen data, producing results in terms of Accuracy, Sensitivity and Specificity.

For each Patient-File, the described process is implemented in 10 runs; each run training a new classifier on a different combination of training and test set. The procedure illustrated thus far is similar to that of the seizure prediction seen in (chapter 6). In order to predict seizures more than 5 minutes in advance of onset, these modules are further integrated into a new algorithmic framework.

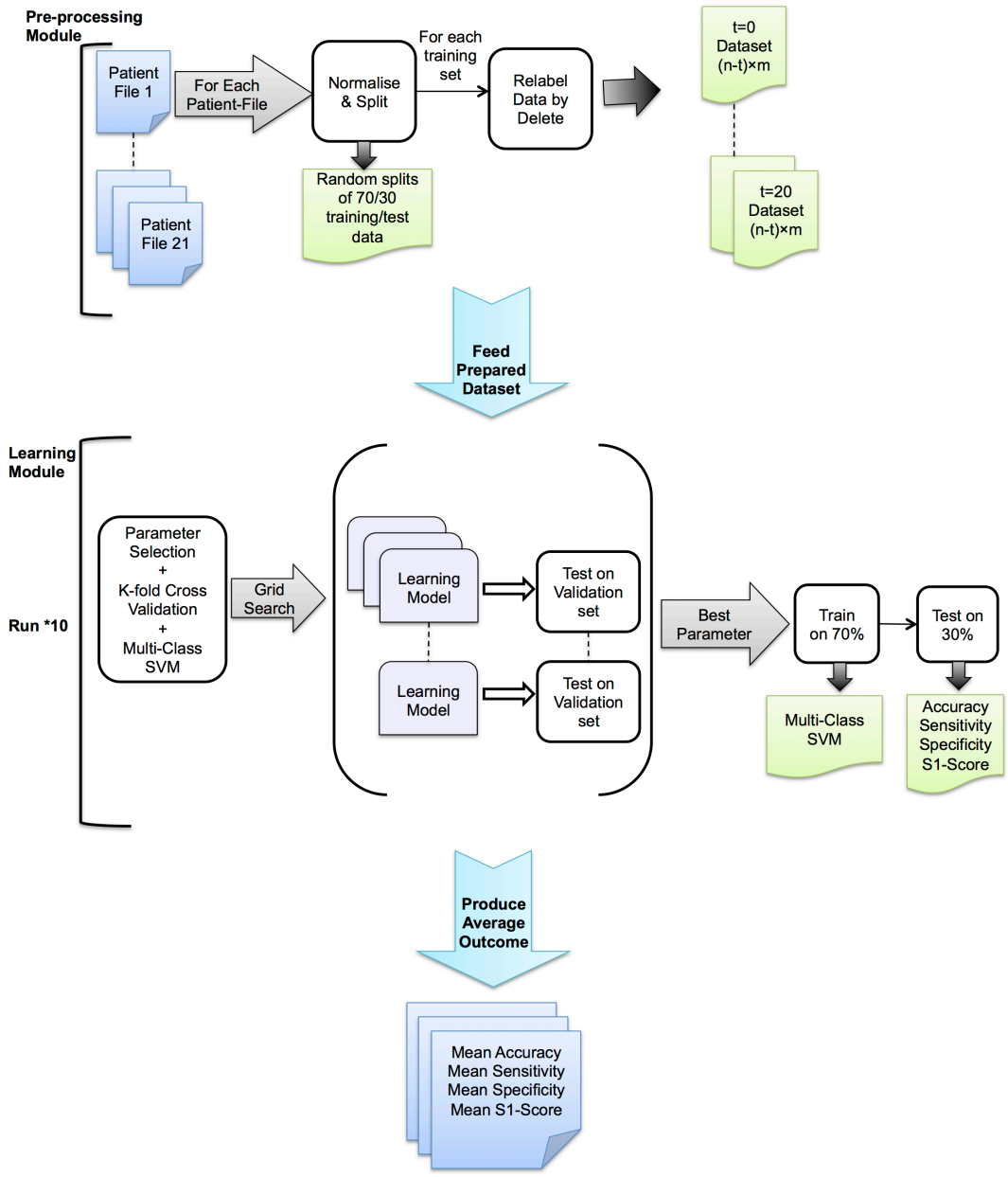


Figure 7.1 - The architecture of the Advance Seizure Prediction Experiment – The system consists of a Pre-processing Module and a Learning Module. The data preparation and initial experimental setup takes place in the Pre-processing Module, which varies for each experiment. This is separated from the learning and classification task in the Learning Module, which remains unchanged for tor the main part of the experiments. We use the term 'predictor' to refer to this prediction architecture.

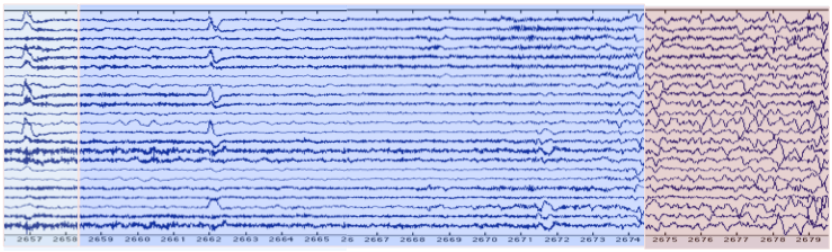
In this framework we redefine the advance prediction of a seizure as a classification problem, by sliding the pre-ictal window back in time. By moving the pre-ictal data in pre-defined intervals of time *t*, we are able to use the prediction algorithm described in

chapter 6, which is in effect time $t \times 5$ minutes prior the actual onset. In order to achieve this behaviour we propose the following approach:

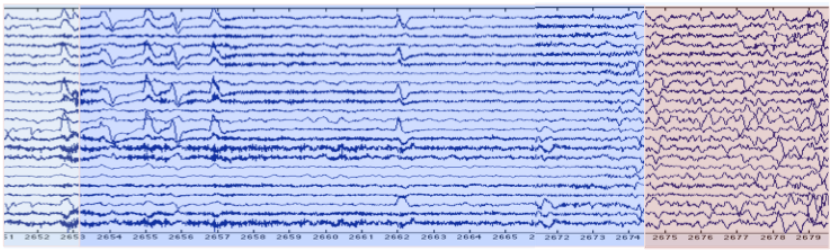
Advance Prediction by Removing Pre-Seizure Data

In this approach we slide the pre-ictal data back in time by effectively extending it to point $t_0 - t$. In order to achieve this, non-seizure data preceding the seizure onset is clipped out in windows *t* of length 1 (minute) and non-ictal data of length *t* preceding the original pre-ictal data is re-labeled as pre-ictal; this in fact decreases the length of our non-seizure data. The correct classification of data points by our model implies the detection of data in advance. For instance time frame $t = 1$ denotes that the data 1 minute prior the seizure onset deleted, and the new pre-ictal window starts at 1 minute prior the seizure onset. If our model successfully classifies the data in time frame *t* prior the seizure as pre-ictal, this means that data in time frame t contains markers that indicate the occurrence of future seizures. We expand time frame t in 1-minute increments and up to a maximum of 20 minutes $t_0, t_1, t_2, \ldots, t_{20}$ where t_0 is the seizure onset. The reason we have selected 20 as the upper bound of our advance time frame is purely due to data availability. Our ictal data as described in chapter 4 only contains 1-hour recordings per patient; out of which 20 minutes was the maximum we could go back without interfering with the inter-ictal or post-ictal stages of other seizures. Figure 7.2 illustrates this algorithm in 6 steps for t_0 through t_5. This predictive algorithm is referred throughout this thesis as the Delete algorithm.

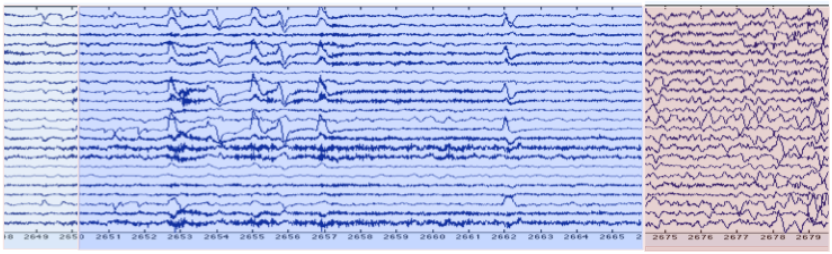
$t = 0$
$D = m \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = x \times n$



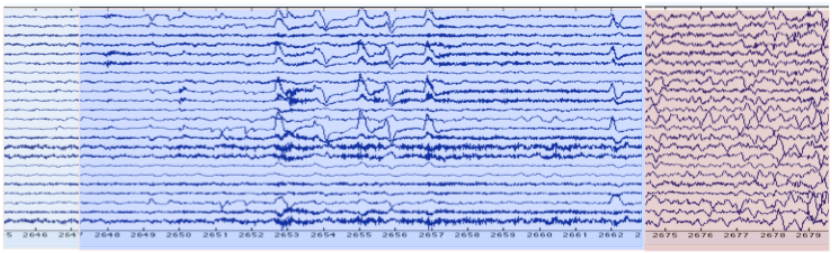
$t = 1$
$D = (m - t) \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



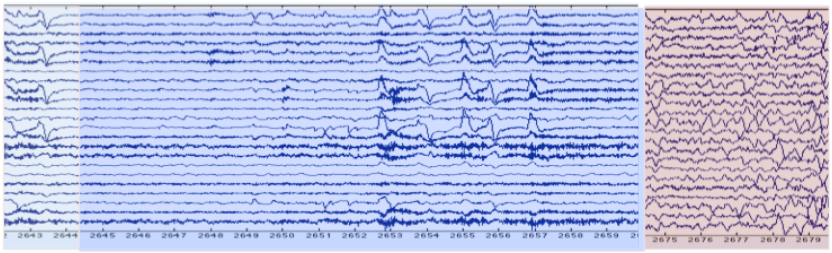
$t = 2$
$D = (m - t) \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



$t = 3$
$D = (m - t) \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



$t = 4$
$D = (m - t) \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



$t = 5$
$D = (m - t) \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$

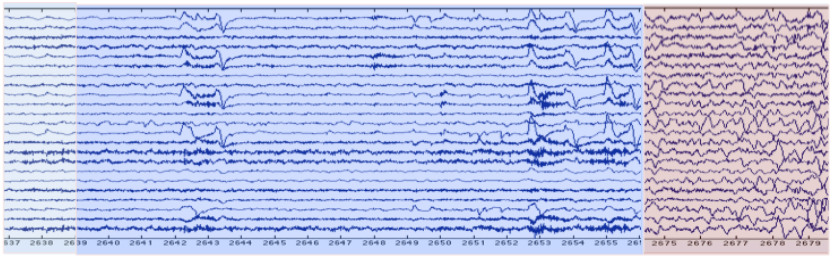


Figure 7.2 Prediction Simulation by Deleting Data – The top image displays the default EEG status where the ictal data is preceded by Pre-ictal data. In this illustration we display up to 5 steps of data manipulation. At each step *t,* the ictal window is pushed back for a fixed interval by deleting the preceding data, which is either pre-ictal or inter-ictal. For each stage of the Delete process, updates to time-step t, ictal length, pre-ictal length, inter-ictal length and overall dimensionality of the dataset is indicated.

Minutes in advance	#instances	Minutes in advance	#instances	Minutes in advance	#instances
0	2049	7	1797	14	1545
1	2013	8	1761	15	1509
2	1977	9	1725	16	1473
3	1941	10	1689	17	1437
4	1905	11	1653	18	1407
5	1869	12	1617	19	1383
6	1833	13	1581	20	1359

Table 7.1 The number of instances of patient 2at each step of advance prediction dataset modification by Delete.

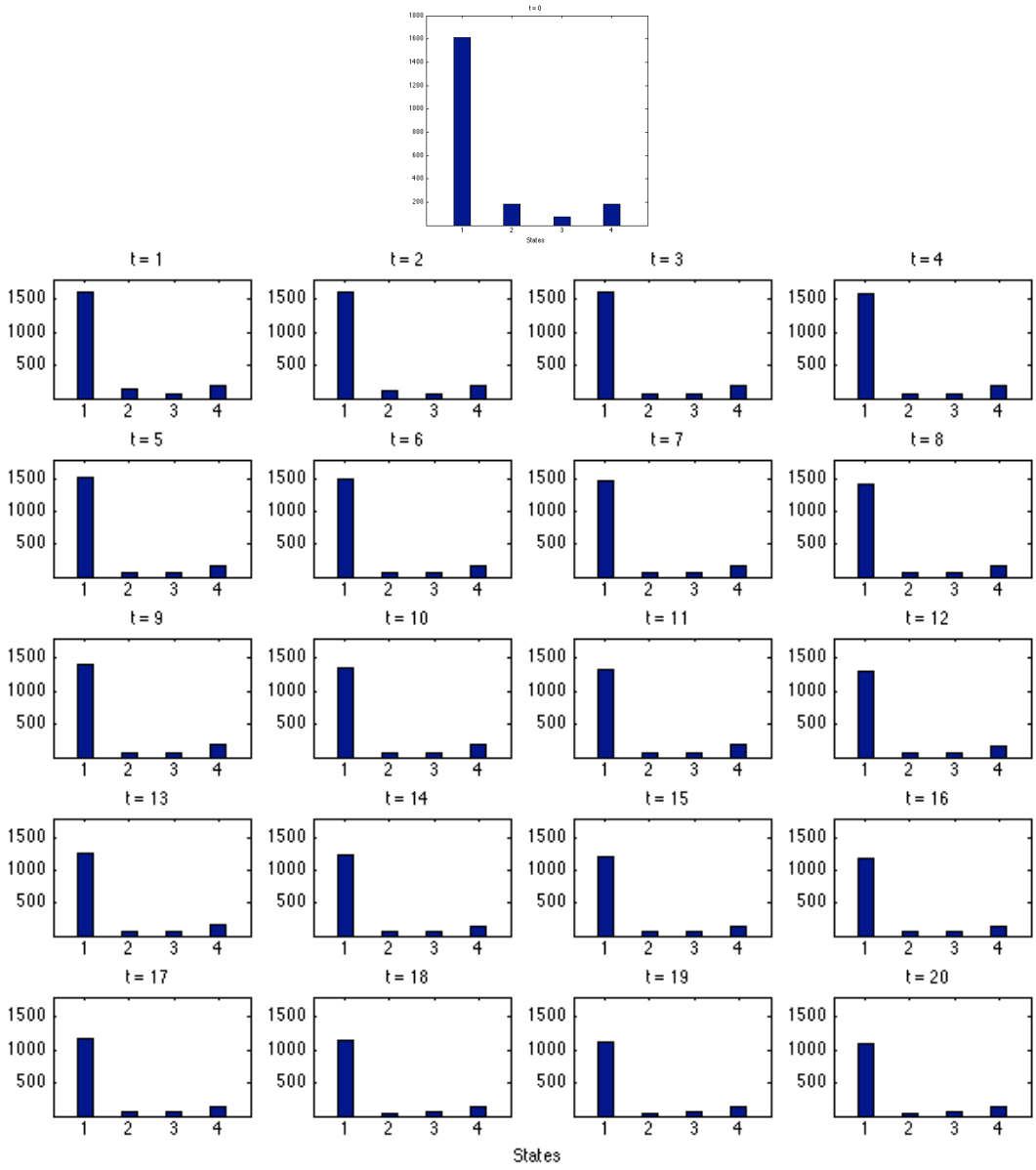


Figure 7.3 The class distribution of patient 2 along all steps of the Delete algorithm. The top histogram presents the class distribution of the unchanged patient-file (t = 0). The results reveal that overall, class distribution varies by a little amount, at each step.

7.2.2 Results of Single-Channel Advance Seizure Prediction

In the experiments of this chapter, we only consider the average performance of the predictor (Figure 7.1) over all Patient-Files, as the performance measures for most patients are quite similar.

Figure 7.4 shows the average performance of all patients over time. The performance measures are Accuracy (the percentage of correctly classified instances), Sensitivity (the capacity of identifying positive cases), Specificity (the capacity of correctly identifying negative cases) and S1-Score (weighted harmonic mean of precision and recall). In the majority of seizure prediction studies, Sensitivity is the main determinant performance measure as it tends to be more variable under experimental conditions in comparison to the other measures: Accuracy and Specificity. Additionally, when it comes to the trade-off between Sensitivity and Specificity, maintaining a high Sensitivity is favored in seizure prediction studies, as Sensitivity to seizure occurrence is more important than false alarms, for real-life application. Most classification models are tuned to improve the Accuracy; therefore this measure is most likely to remain stable over different experimental conditions. As mentioned in chapter 2, we pay particular attention to S1-Score, as it is our main performance criteria in comparing learning models. The S1-Score measure combines the results of both Sensitivity and Specificity, calculating a harmonic mean of both measures.

	ACC	t	SP	t	SS	t	S1	t
min	95.75	0	98.77	0	80.11	4	88.33	4
max	96.28	12	99.35	15	88.02	0	92.99	0
t = 0	95.75	0	98.77	0	88.02	0	92.99	0
t = 20	95.81	20	99.04	20	85.90	20	91.88	20
mean	96.07		99.14		84.10		90.79	
median	96.09		99.13		84.29		90.85	
mode	95.75		98.77		80.11		88.33	
std	0.18		0.15		2.32		1.42	
range	0.53		0.58		7.91		4.67	

Table 7.2 Summary of important data statistics from the stepwise advance seizure prediction by **Delete** performed on 18 **single-channel** patients – The minimum, maximum, mean, median, mode, standard deviation and range of the four measures are listed. Additionally, the four measures at the seizure onset (t = 0) and at the largest prediction window (t = 20) have been listed. The 't' column for each measure lists the timepoint of the corresponding statistical measure. E.g. the minimum value for Accuracy is at timepoint 0 with value 95.75%.

In Figure 7.4 and Table 7.2 we can see that Accuracy and Specificity are steady overall throughout different stages of prediction. Due to the low variation in Specificity, Sensitivity appears to strongly influence the variability of S1-Score, resulting in the strong resemblance of S1-Score and Sensitivity curves. The values of S1-Score are, however, higher than those of Sensitivity due to the consistently high values of Specificity throughout the different prediction windows. The starting value for S1-Score is 92.99% at timepoint t_0, which is merely the result for automatic seizure onset detection. The minimum S1-Score value is 88.33% at t = 4 minutes prior to seizure onset and the highest value is 92.99% at time t = 0 followed by other timepoints with values $=\sim t_0$, namely t = 13 and t = 14. The S1-Score values are within the range [88.33%, 92.99%] with a mean value of just under 90.79%. The majority of timepoints t within the range $t_9 \leq t \leq t_{20}$, are above the mean value for S1-Score.

Inter-Patient Variability

In Figure 7.5 we see the distribution of the performance (S1-Score) of each patient predictor over time. The performance measure in each box represents a *collection* of Patient-File S1-Scores as opposed to an averaged S1-Score for all patients used thus far in the analysis of our results. The boxes indicate the higher and lower quartiles of the population of patient predictors. The length of whiskers was calculated from the inter-quartile range and the line in the middle represents the median. Any points outside the span are deemed as outliers.

The figure shows the population of patients' S1-Scores against the prediction time frames for the Delete predictor. The box plot shows some outliers at timepoints t = 1, 7, 8, 11,19, 20. The 4 patients recurring as outliers to the population are patients 3, 10, 14 and 20, with patient 14 having the highest recurrence as an outlier. Patient 14 is also an outlier at time t_0, suggesting that the classifier does not train well on this particular patient. Boxes at times t = 0, 1, 10, 11, 13, 14, 15, 16, 17 and 18 have a relatively low variability, whereas the S1-Score at timepoints t = 3, 5, 7, 8, 19 have higher levels of variability.

Our important moments from the previous diagram were t = 0, 4 and 12. The population at these timepoints has low variability, which indicates that most patient predictors produce similar results at these timepoints. At t_0 population is symmetric, indicating a normal distribution of values within the inter-quartile range. At t_{12}, the

median is not centred in the middle of the inter-quartile range and has a shorter upper-quartile. This indicates that the population of S1-Scores of Patient-Files is slightly skewed to the left, showing that the majority of the sample population is within the range [91%, 94%] and the lower quartile is highly variable in comparison. It is therefore safe to assume that the peaks and dips respectively at moments 4 and 12 are valid for the majority of predictors.

Table 7.3 summarises the paired t-test of several predictive timepoints of all patients. The comparison of $t = 0$ and $t = 20$ does not show any significant differences which suggests that seizures can indeed be predicted 25 minutes in advance with minimum loss of performance. We also examine $t = 13$ which is a local maxima in Figure 7.4, against two low values at $t = 4$ and 5. The t-test comparison indicates that although $t = 0$ and 20 are not significantly different, other time-points in between have significant differences among them, as also seen in Figure 7.4.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
T=0,T=20	0.899	4.594	-1.192	2.990	0.897	20	0.380
T=13,T=4	4.982	6.439	2.052	7.913	3.546	20	**0.002**
T=13,T=5	3.402	6.464	0.459	6.344	2.412	20	**0.026**

Table 7.3 The mean S1-score of all patients in several time-points is examined in a paired t-test. The test examines t=0 vs. t=20, t=13, vs. t=4 and t=13 vs. t=5. The values in bold indicate $p \leq .05$ and are considered statistically significant.

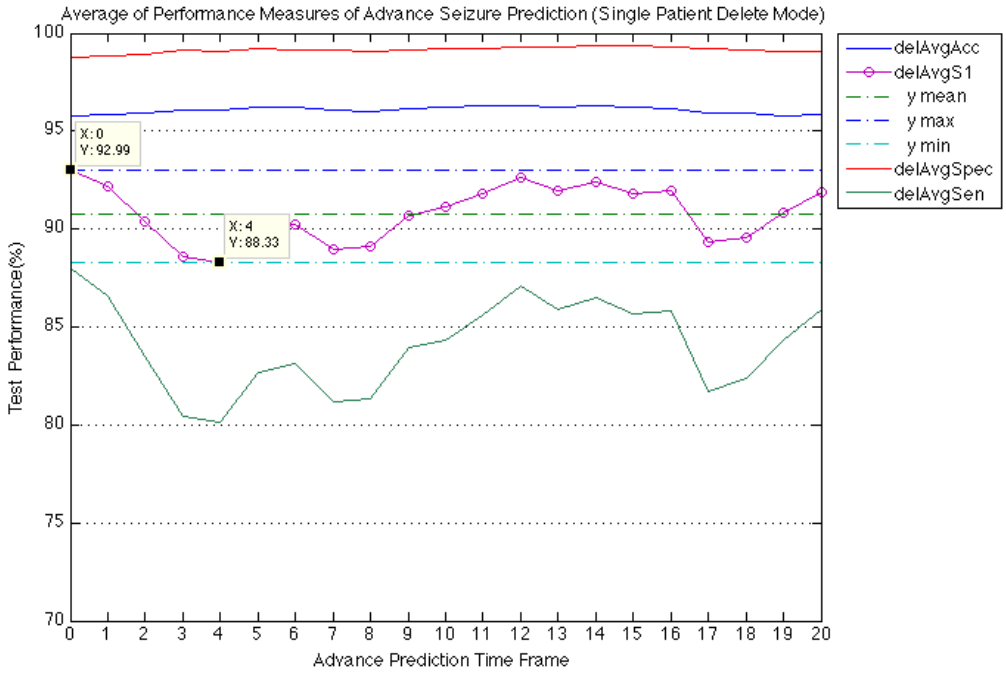


Figure 7.4 Summary of stepwise advance prediction by **Delete** on 18 **single-channel** patients – The plot shows Accuracy, Sensitivity, Specificity and S1-Score averaged across all 18 patients at each prediction time-step. The plot also displays the minimum, mean, maximum and full feature-set values for the S1-Score measure.

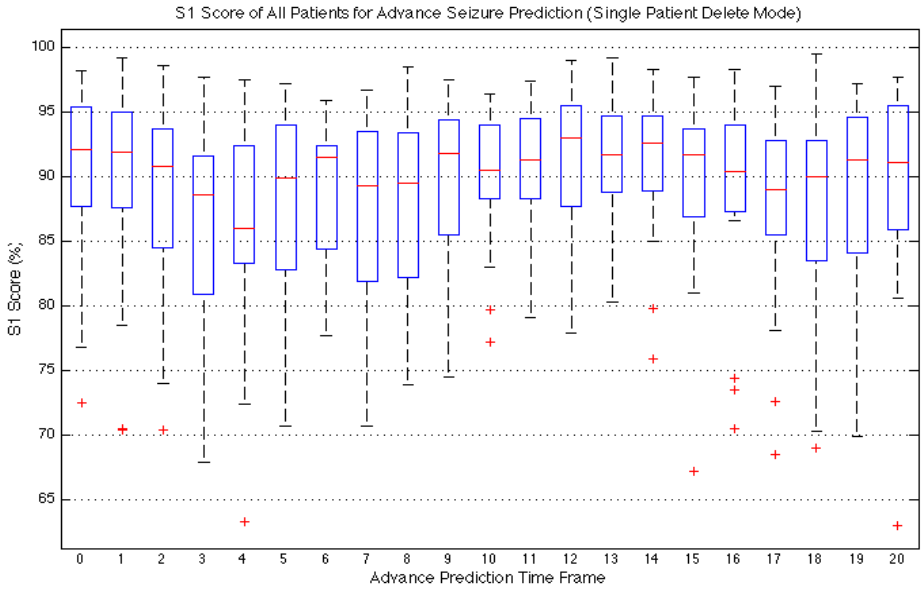


Figure 7.5 The Box and Whiskers diagram for stepwise advance prediction by **Delete** on 18 **Single-Channel** patients – The boxes at each interval display the distribution of average S1-Score of each of the 21 patients at each advance prediction time-step.

7.2.3 Discussion on Single-Channel Advance Seizure Prediction

In this experiment we used the Delete algorithm to evaluate the prediction performance of seizures several minutes in advance. The method was used to generate subsets of datasets in which pre-ictal data was pushed back for increments of t = 1 to t = 20 minutes. This resulted in 21 data subsets for each patient. The subsets were then used in training and testing of Multi-class SVMs. The performance averaged over all Patient-Files revealed high Accuracy and Specificity for all prediction windows. S1-Score and Sensitivity were relatively variable within a high range of values.

In Delete, the pre-ictal window was pushed back in time while deleting the original pre-ictal and inter-ictal data to simulate advance prediction. This entailed that the seizure activity in ictal state remained intact while the size of the dataset was reduced each round.

The performance was highly variable among some intervals while it remained smooth for others. Prediction at 20 minutes in advance was better than prediction at 4 minutes in advance, for both predictive methods. In the case of Delete, performance was almost as high as t_0 for $11 \leq t \leq 16$ with a low standard deviation among patients, revealing that those moments are particularly of significant predictive value. This is indicative of detectable seizure activity in this time-range, which potentially has similar predictive power to the ictal activity at the onset.

7.3 Experiment II: Advance Seizure Prediction - All EEG channels

In experiment I of this chapter, the feature-set of each Patient-File was extracted from the recordings of the first EEG channel; in all cases, this channel is one of the three focal channels in the corresponding signal recording ASCII file. The aim of this experiment is to repeat experiment I on an extended feature-set composed of all 6 channels, and to assess the performance of the patient predictor under the new conditions.

7.3.1 Methods for Multi-Channel Advance Seizure Prediction

In chapter 4 we discussed the preparation procedure of the Patient-Files. Moreover, we reviewed the extraction process of the 14 features from each corresponding Patient-file. For this series of experiments, we extend each Patient-File to include an extended

number of features. These extended properties are extracted from the remaining 5 EEG channels, using procedures described in chapter 4. This process expands the size of the feature-set six fold for each patient, resulting in an $m \times 84$ feature-vector per patient. Processing and training models of the now extensive number of features demand greater computation time and power. For this reason, we modulated the experiments such that each training block could be constructed on several machines in parallel, to reduce the computation time. We used Matlab pooling to run each stage of the Delete and Rename experiments, separately for each patient predictor. These experiments were carried out on a cluster of 8 core 64bit CentOS machines, with each machine running 8 predictors synchronously. The same predictive algorithms described in section 7.2.1 were used to conduct this set of experiments.

7.3.2 Results of Multi-Channel Advance Seizure Prediction

In Figure 7.6 and Table 7.4 we see the various performance measures of the multi-channel advance seizure prediction experiment, averaged over the entire population of patient predictors (except for predictors of patients 3, 10, and 14 which were removed as outliers). These performance measures are Accuracy, Sensitivity, Specificity and S1-Score. We can see that the Accuracy and Specificity of our classification is consistently high over timepoints [0, 20], with values of ~99% for Specificity and ~97% for Accuracy. In other words, seizure prediction of up to 20 minutes in advance is highly and consistently accurate and specific to seizure occurrence. Sensitivity on the other hand is low relative to the values of Accuracy and Specificity, and in the range of ~[80%, 84%]. Due to the consistency of Specificity over time, Sensitivity is the main determinant of the shape of the S1-Score plot, however, the high values of Specificity are reflected in the higher range of values of the S1-Score plot. The maximum S1-Score is 95.66% at timepoint t_0, which implies that detection in this case has a higher average S1-Score than advance prediction. In other words, prediction does not improve at any point beyond the seizure onset. The minimum is at timepoint t = 16 with an 87.9% value, which is considerably high relative to the maximum at t_0. The overall shape of the S1-Score is relatively smooth with low variability, particularly after the initial dip subsequent to t_0. The mean value of S1-Score is 90.23%, indicating a relatively high performance over time.

	ACC	t	SP	t	SS	t	S1	t
min	96.97	19	99.17	19	79.35	16	87.90	16
max	97.79	1	99.49	8	92.27	0	95.66	0
t = 0	97.73	0	99.36	0	92.27	0	95.66	0
t = 20	97.02	20	99.18	20	80.98	20	88.85	20
mean	97.45		99.37		83.07		90.23	
median	97.46		99.39		82.41		89.77	
mode	96.97		99.17		79.35		87.90	
std	0.25		0.09		3.29		2.00	
range	0.82		0.32		12.92		7.76	

Table 7.4 Summary of important data statistics from the stepwise advance seizure prediction by Delete performed on 18 Multi-Channel patients.

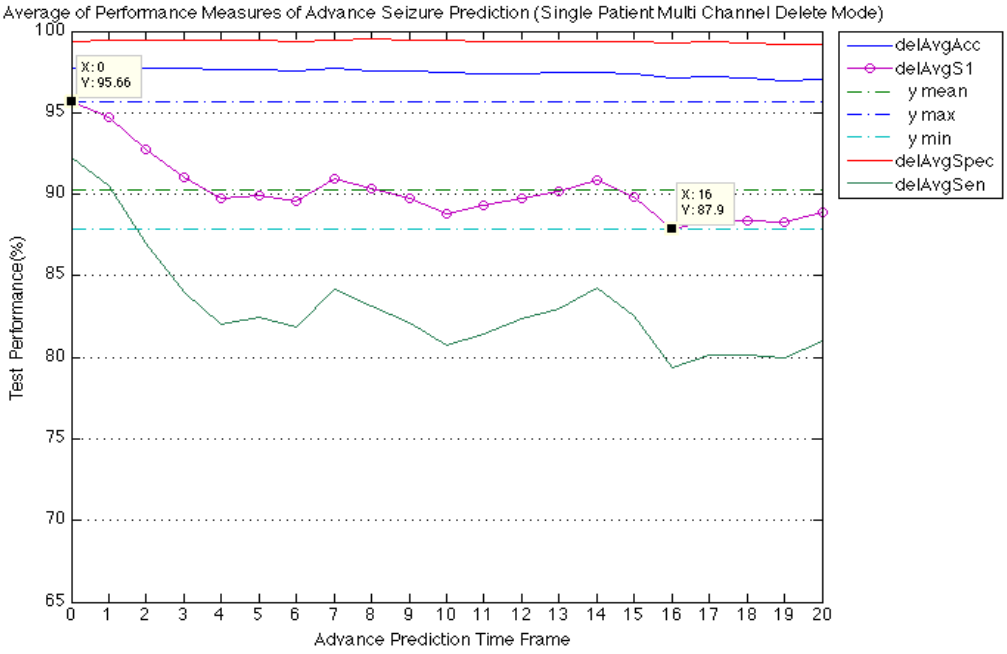


Figure 7.6 Summary of stepwise advance prediction by Delete on 18 Multi-Channel patients.

Inter-Patient Variability

The box and whisker diagram of the average S1-Score of the population of multi-channel patient predictors over time for the experiments is respectively illustrated in Figure 7.7. In the Delete experiment, patients 3, 10 and 14 remain as outliers, along with the occasional appearance of patients 9 and 15 outside the span of the whiskers. The latter two however, were not be excluded from the average performance analysis as they are not extreme instances and have only become outliers in the course of the change of experimental settings (which was expected). The highest performance value

with the least variability among different patients is at time t_0 as anticipated. The majority of boxes appear to be symmetric, showing a normal density in most cases. We can also observe low variability across timepoints [4, 20]. The lowest whisker ends are at timepoints 9, 10, 12 and 20, but the long bottom whiskers in these timepoints show that there is high variability in the lower quartile of the population. The majority of the boxes have shorter top whiskers, indicating a skewness of the population to the left.

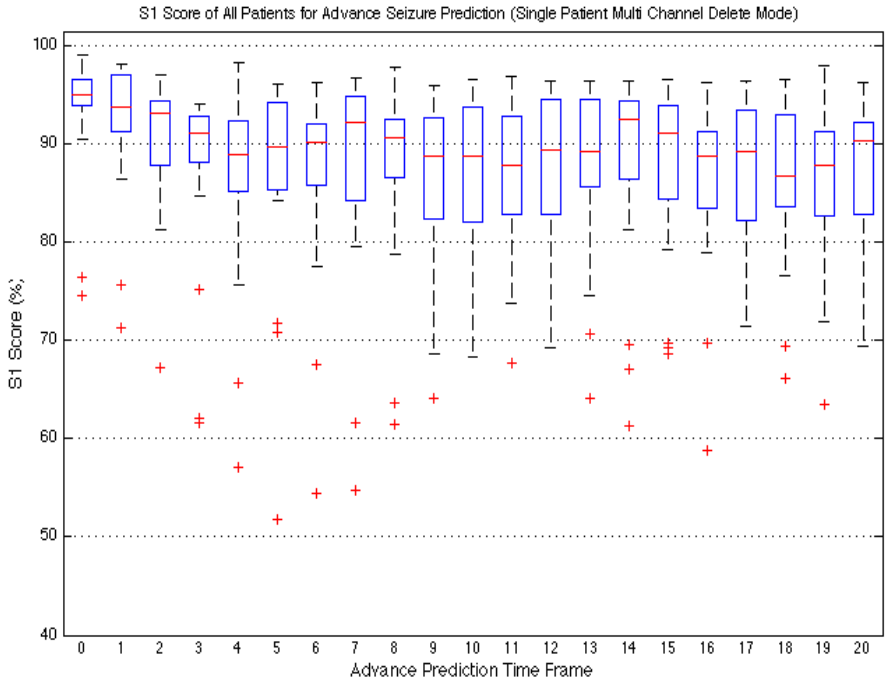


Figure 7.7 The Box and Whiskers diagram for stepwise advance prediction by **Delete** on 21 **Multi-Channel** patients.

Table 7.5 summarises the paired t-test of several predictive timepoints of all patients in the multi-channel predictive experiments. The comparison of $t = 1$ and $t = 0$ shows significant differences which confirms the observed variation seen at these moments in Figure 7.6. We also examine $t = 14$ vs. $t = 16$ and $t = 4$ vs. $t = 7$ which are respectively local maxima and minima in Figure 7.7. The results of the t-test suggest that there are no significant differences between these moments which in turn entails that the timepoints in the multi-channel setting do not particularly differ from one another, giving a more constant outcome compared to the single channel scenario.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
T=0,T=1	1.186	2.184	0.192	2.180	2.488	20	**0.022**
T=16,T=14	1.879	7.606	-5.341	1.583	-1.132	20	0.271
T=4,T=7	1.269	5.333	-1.158	3.697	1.091	20	0.288

Table 7.5 The mean S1-score of all multi-channel patients in several time-points is examined in a paired t-test. The test examines t=0 vs. t=1, t=16, vs. t=14 and t=4 vs. t=7.

7.3.3 Discussion on Multi-Channel Advance Seizure Prediction

In this experiment, we utilised all 6 EEG channels by extracting the 14 features of Table 6.1 from each channel; this resulted in 84 features for each Patient-File. The S1-Score dropped at $t>0$ and never picked up for either of the predictive algorithms. Timepoints 1-3, 7 and 14 were of the high performance predictive moments for both algorithms.

The results indicated that, given the current feature-set, extracting features from all focal and extra-focal channels does not improve advance prediction. This may be due to the redundancy of some features, particularly those extracted from the extra-focal channels where seizure activity may not be as prominent as other channels. It can also be due to over-fitting caused by the introduction of new features. Although advance prediction is not improved by the introduction of more channels, it is indeed less variant among time-steps.

7.4 Experiment III: Advance Seizure Prediction - Extended Feature-Set

In section 6.4 we presented an extended feature-set for each patient, the derivation of which was based on extensive feature selection experiments presented in the same chapter. In this experiment, we assess the predictability of the seizure state in each Patient-File, on an extended feature-set with up to 204 extracted features, ranging over all 6 EEG channel recordings. The outcome of this experiment will reveal the effect of an extended feature-set, comprising new and original features on the predictability of the seizure state of the brain.

7.4.1 Methods for Advance Prediction of Seizures on Extended Feature-set

The extended feature-set developed in chapter 6 (see section 6.4) is used as the new dataset for this experiment. From each channel, a further 20 features were extracted, increasing the number of features per channel to 34, and growing the feature-vector to incorporate $f \times 204$ elements. This increase in the size of the feature matrix demands additional computational cost. The experiment was distributed over a cluster of 8 core 64bit CentOS machines, with each machine in the cluster running 8 Matlab pools in parallel. By modulating the experiment in several smaller workloads of single runs of training and classification for each patient, each predictor module was only implemented in a single thread, hence, eliminating the need to distribute the model building over different workstations.

7.4.2 Results of Advance Seizure Prediction on Extended Feature-set

The four performance measures Accuracy, Sensitivity, Specificity and S1-Score of the Delete presented in Figure 7.8 and Table 7.6. Each of the curves in Figure 7.8 is a measure of the performance of prediction on the extended feature-set, averaged over all the patient predictors (except patients 3, 10 and 14 which were removed as outliers).

In results of Delete we can see divergence of the lines from the patterns we have seen throughout this chapter. The Specificity value is the highest we have seen thus far, with mean 99.70%. This is indicative of an improvement in the Specificity measure through the introduction of additional features. Accuracy is also steadily high, although the line has a downward slope towards the later timepoints. The Sensitivity is variable throughout time, as we have seen in previous results, however, the lowest values are just below 70%.

The S1-Score once again mimics the behaviour of Sensitivity, the most variable of the two measures. The maximum S1-Score is at t_0, with 88.94%, which indicates that prediction does not perform better than seizure detection. The maximum value however, is the lowest we have seen throughout the experiments in this chapter. The shape of the S1-Score is different to what we have seen in other experimental results. The initial dip is sharper, dipping at timepoint $t = 3$ at a value approximately ~9% lower. After the initial dip, the line monotonically ascends, fluctuating between local maxima and minima, until it hits a dip at timepoint $t = 18$ which is statistically the minimum value. After hitting the minimum it starts ascending until it reaches the final timepoint

at 83.05%. The S1-Score line is fairly smooth between the several dips and peaks, predominantly for timepoints 12, 13, 14, and 15.

	ACC	t	SP	t	SS	t	S1	t
min	93.32	19	99.61	19	66.90	18	78.30	18
max	94.40	5	99.78	15	81.84	0	88.94	0
t = 0	94.27	0	99.62	0	81.84	0	88.94	0
t = 20	93.49	20	99.66	20	73.39	20	82.90	20
mean	94.05		99.70		73.33		83.05	
median	94.09		99.71		73.40		83.17	
mode	93.32		99.61		66.90		78.30	
std	0.33		0.04		3.25		2.42	
range	1.09		0.17		14.95		10.63	

Table 7.6 Summary of important data statistics from the stepwise advance seizure prediction by **Delete** on 18 **Multi-Channel Extended Feature-Set** patients.

Inter-Patient Variability

In the box and whisker diagrams of the Delete experiment (Figure 7.8), the variability among distinct timepoints is quite low, except for the timepoints in the range $0 \le t \le 3$, where the initial dip seen in Figure 7.7 takes place. Most boxes are long, with some spanning a range of 20%, indicating high variability among the predictor performance at each timepoint with a centred or top median line; respectively indicating that the underlying population is symmetric or left-skewed. This means that despite the variability between some predictor values at various timepoints, the median and above the median are densely populated and variability is mainly in the lower quartile. The values are among the lowest we have seen so far for the Delete experiments, with the lowest whisker in the range of [48%, 60%]. Although outliers were removed from the average performance analysis, some of them have performed extremely poorly with values as low as 17%. The outliers for the timepoint t_0 (detection at seizure onset) are patients 13 and 2. Patients 14, 10 and 3 do not seem to appear in the outliers, suggesting that the extended feature-set may have improved their overall performance.

Table 7.7 summarises the paired t-test of several predictive timepoints of all patients in the extended feature-set predictive experiments. The comparison of $t = 1$ and $t = 0$ shows significant differences which confirms the observed variation seen at these moments in Figure 7.7. We also examine $t = 20$ vs. $t = 18$ and $t = 8$ vs. $t = 3$ which are respectively local maxima and minima in Figure 7.7. The results of the t-test

suggest that there are no significant differences between these moments, which in turn entails that the timepoints in the extended feature-set setting do not particularly differ from one another. The extended feature-set, as with the multi-channel setting, produces a more constant outcome compared to the single channel scenario.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
T=0,T=1	2.733	2.819	-4.017	-1.450	-4.443	20	**0.000**
T=20,T=18	2.694	7.478	-6.098	0.710	-1.651	20	0.114
T=8,T=3	1.939	9.165	-2.233	6.111	0.969	20	0.344

Table 7.7 The mean S1-score of the extended feature-set (204) for all patients in several time-points is examined in a paired t-test. The test examines t=0 vs. t=1, t=20, vs. t=18 and t=8 vs. t=3. The values in bold indicate $p \leq .05$ and are considered statistically significant.

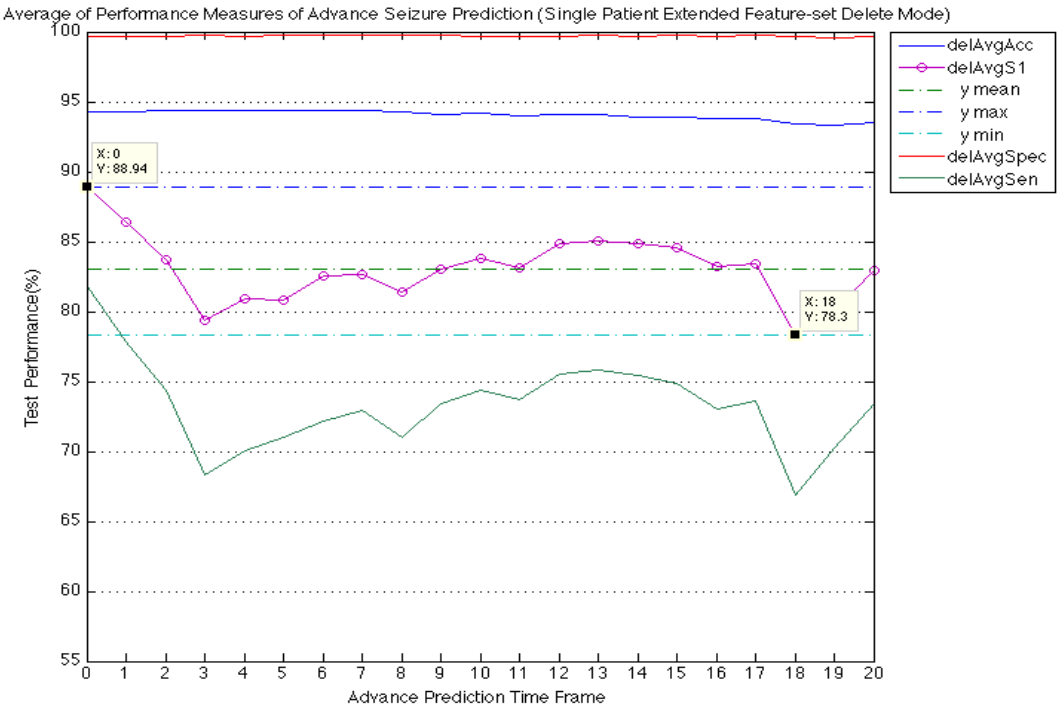


Figure 7.8 Summary of stepwise advance prediction by **Delete** on 18 **Multi-Channel Extended Feature-Set** patients.

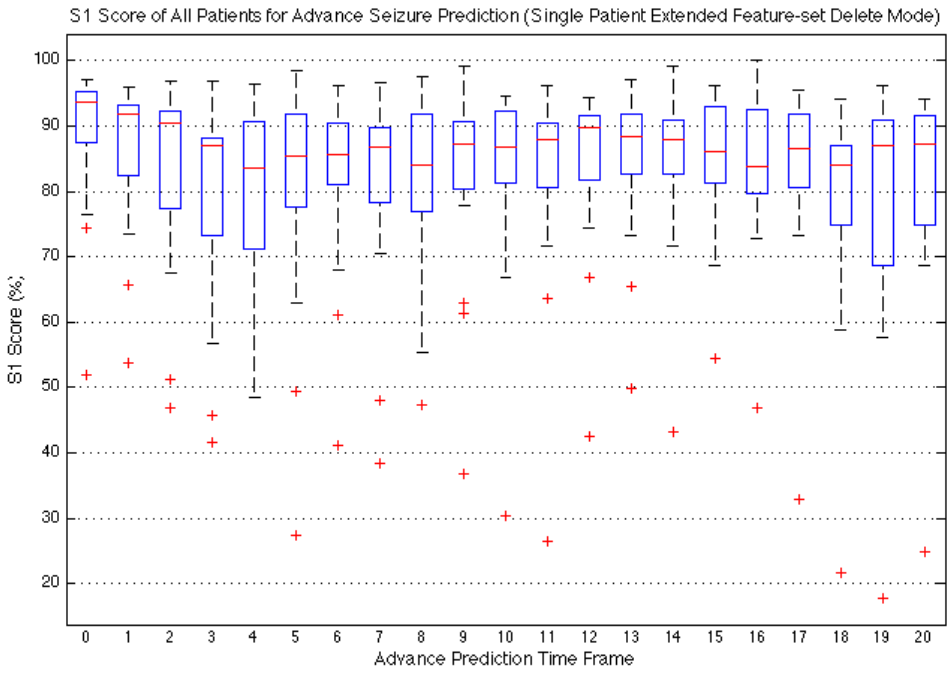


Figure 7.9 The Box and Whiskers diagram for stepwise advance prediction by **Delete** on 21 **Multi-Channel Extended Feature-Set** patients.

7.4.3 Discussion on Advance Prediction of Seizures on Extended Feature-set

In this experiment, we used a 204 dimensional feature-set with a mixture of the original feature-set (Table 6.1) and new features (Table 6.12) introduced in chapter 6. We saw in experiment II of this chapter that the introduction of additional features extracted from all recorded EEG channels did not improve advance prediction. In this experiment we used a far larger feature-set which is yet again derived from all 6 EEG channels, but with 20 additional features introduced per channel.

The Delete algorithm produced a similar S1-Score trend to that seen in the previous experiment, this time with higher oscillations between time steps. The overall spectrum of values was not necessarily poor, with S1-Score within the range [78.30%, 88.94%]. This reveals that i) regardless of changes to the feature-set, a common trend can be observed for the S1-Score over time-steps throughout all experiments seen thus far, supporting the hypothesis that seizure activity markers exist several minutes before the actual onset. ii) The use of more features does not improve the predictability of our model, though the outcome is still relatively high. The full effect of introducing new

features is not prevalent in this experiment as it is over-shadowed by the noise induced by the high number of features. The effects of a reduced subset of the new feature-set will be evaluated in a later section of this chapter.

7.5 Experiment IV: Advance Seizure Prediction on Subset of Extended Feature-Set

In section 6.4 we expanded our original feature-set to comprise 20 additional features per EEG channel. In section 5 of this chapter we looked at the performance of our advance predictor algorithm on the extended feature-set and discovered that performance at the full set of extended features did not necessarily improve prediction. In this section, we conduct advance prediction on a limited subset of the extended feature-set ranked highest by our feature selection method (explained in 6.2) in order to verify subsequent improvements in perdition.

7.5.1 Methods for Advance Prediction on Subsets of Extended Feature-set

The extended feature-set engineered in section 4 of chapter 6 is used in this experiment. These features were heuristically selected based on previous experimental outcome and literature review; they included features related to Signal Edge Frequency, Signal Band Power and statistical moments. We then carried out stepwise feature selection on the new feature-set and found that performance of the learners enhanced on a smaller subset of features $8 < f \leq 144$ for ReliefF. We had a closer look at these features and discovered that the majority of the new features resided in the high-performance range; from which, we selected 14 top ranked features. The top 14 features were a mixture of original and new features but the new features reserved a greater proportion of the set.

We used the subset of features in this experiment in the same manner as seen in previous experiments. Each Patient-File had a set of top 14 features, though these feature-sets are different among Patient-Files. These features may be from any of the 6 recorded EEG channels, so a patient may have 6 features across all channels (e.g. Sbp Delta 1, Sbp Delta 2, Sbp Delta 3, …, Sbp Delta 6) and several features from a single channel. The feature-sets were used in the advance prediction experiment, resulting in a total of 21 experiments per Patient-File. The experiments were parallelised over a

cluster of thirty 8-core 64bit CentOS machines using Matlab parallel pooling. The implementation was similar to that of other experiments in this chapter.

7.5.2 Results for Advance Prediction on Subsets of Extended Feature-set

The performance of the Delete predictive algorithm on the top 14 ReliefF features are depicted in Figure 7.10 and summarised in Table 7.8. Accuracy, Specificity, Sensitivity and S1-score measures are averaged over all patients except for those removed as outliers. Accuracy and Specificity are consistently high with values respectively in the ranges [97.41%, 97.88%] and [99.36%, 99.67%]. These values are, so far, the highest within the non-expanded dataset. Sensitivity and S1-score are also within a high range with prominent variation between time-steps. Sensitivity is within the range [88.95%, 93.47%] and standard deviation of 1.29%, which is indicative of the lowest variability of Sensitivity observed in presented experiments. S1-Score is also within an exceptionally high range [93.79%, 96.30%] and a very low standard deviation of 0.77%. The S1-Score curve starts at t_0 with 96.18% and rises a little at t = 1, **96.30%,** which is so far, the highest observed value for advance prediction, and suggests that prediction at t_1 leads to a better performance than t_0 which is the onset detection timepoint. It then decreases until it hits a dip at t = 5 with 93.92% which is yet again, very close to the value of the minimum at t = 10, 93.79%. The value then rises to a peak at t = 8 with 96.13%. After this point it oscillates around the mean, hitting the minimum at t = 10 along the way. It stabilises for a short period over t = 13, 14, 15 at the mean value 94.98%. Then the value rises briefly at t = 16 with 95.63% and wavers around the mean, until it finally hits t = 20 at 94.20%.

	ACC	t	SP	t	SS	t	S1	t
min	97.41	20	99.36	0	88.95	10	93.79	10
max	97.88	7	99.67	7	93.47	1	96.30	1
t = 0	97.52	0	99.36	0	93.38	0	96.18	0
t = 20	97.41	20	99.44	20	90.15	20	94.20	20
mean	97.68		99.55		91.14		94.98	
median	97.73		99.56		91.07		95.01	
mode	97.41		99.36		88.95		93.79	
std	0.14		0.08		1.29		0.77	
range	0.47		0.31		4.52		2.51	

Table 7.8 Summary of important data statistics from the stepwise advance seizure prediction by **Delete** on 18 patients – The dataset of each patient comprises a **subset of 14 Multi-Channel Extended Feature-Set** determined by **ReliefF** feature selection method.

Inter-Patient Variability

The box plot in Figure 7.11 portrays the S1-Score of all Patient-Files across all time-steps *t*, from 0 to 20 for the ReliefF experiment. The outliers are mainly patients 3, 10 and 14, which were previously removed from the performance analysis. The shortest boxes in Delete are at $t = 0$ and 14. The boxes are mainly parallel indicating little variation between time-steps. The red line is also mainly in the middle or in the upper half of the box indicating a normal distribution of S1-Scores around the median for each time-step or a left skewed distribution where the higher value range is densely populated, indicating fewer Patient-Files in the lower ranges.

Table 7.9 summarises the paired t-test of several predictive timepoints of all patients in experiments using a subset of the extended feature-set. The comparison of $t = 1$ and $t = 0$, $t = 5$ vs. $t = 8$ and $t = 10$ vs. $t = 16$ revealed no significant differences between the performances of these extreme timepoints. This suggests that the performance is relatively constant across all predictive time-points, indicating that the use of a subset of good suitable features produces the highest S1-scores seen in Figure 7.9, which are relatively consistent across the timepoints.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
T=0,T=1	0.024	1.153	-0.501	0.549	0.095	20	0.925
T=5,T=8	1.989	5.402	-4.448	0.470	-1.687	20	0.107
T=10,T=16	1.328	3.912	-3.109	0.453	-1.555	20	0.136

Table 7.9 The mean S1-score of all extended feature-set patients in several time-points is examined in a paired t-test. The test examines t=0 vs. t=1, t=5, vs. t=8 and t=10 vs. t=16. The values in bold indicate $p \leq .05$ and are considered statistically significant.

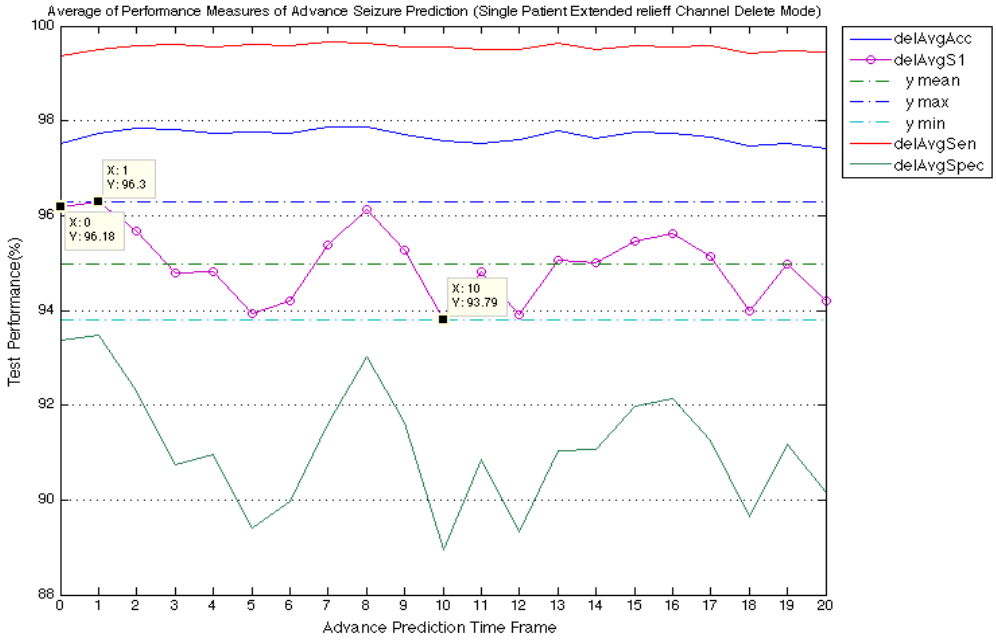


Figure 7.10 Summary of stepwise advance prediction by **Delete** on 18 patients – The dataset of each patient comprises a **subset of 14 Multi-Channel Extended Feature-Set** determined by **ReliefF** feature selection method.

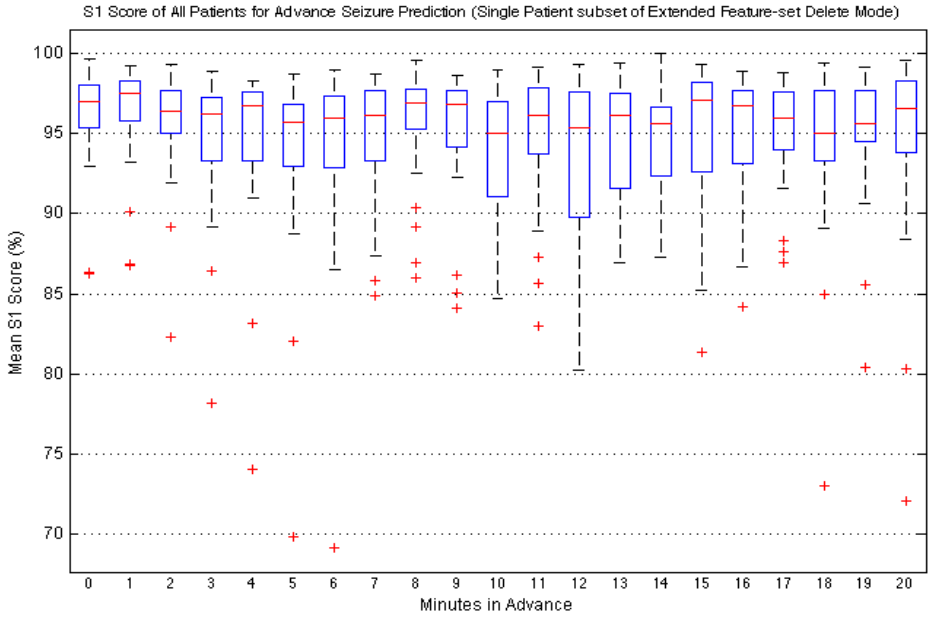


Figure 7.11 The Box and Whiskers diagram for stepwise advance prediction by **Delete** on 21 patients – The dataset of each patient comprises a **subset of 14 Multi-Channel Extended Feature-Set** determined by **ReliefF** feature selection method.

7.5.3 Discussion on Advance Prediction on Subsets of Extended Feature-set

In this set of experiments we used the extended feature-set, constructed heuristically in chapter 6, to build a smaller feature-set within the high performance range. The top 14 features, which comprised a higher proportion of new features and a smaller proportion of original features, were then used in our advance prediction algorithm.

The results revealed that the extended top 14 features performed better than the benchmark (comprised single-channel original feature-set) and the full set of multi-channel and extended features. This experiment produced the highest advance prediction performance values with S1-score of 96.30% at $t = 1$.

The results of this experiment were examined against those of the single-channel experiment in a paired t-test. Table 7.10 presents the t-test performed on the two different feature-sets for the timepoints of $t = 0$, 1, 4 and 20. The results reveal that the subset of the extended feature-set performs significantly better than that of single channel in three out of four sample timepoints.

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Stdev	Lower	Upper	t	df	Sig. (2-tailed)
T=0 Sing/ExRel	5.391	8.109	1.700	9.082	3.046	20	**0.006**
T=1 Sing/ExRel	6.527	8.792	2.525	10.529	3.402	20	**0.003**
T=4 Sing/ExRel	8.575	9.711	4.155	12.996	4.046	20	**0.001**
T=20 Sing/ExRel	4.792	11.787	-0.574	10.157	1.863	20	0.077

Table 7.10 The mean S1-score of several time-points is examined between the single-channel and extended feature-set, in a paired t-test. The test examines t=0, t=1, t=4 and t=20. The values in bold indicate $p \leq .05$ and are considered statistically significant.

The experiment revealed that time-steps 5 and 18 were in the lower performance range and moments 1, 2, 8, 16 were in the higher performance range. This consistency within the poorer and better performance range could be indicative of significant predictive value of these moments, in other words, predictive markers can be traced several minutes in advance of the actual seizure onset.

Using the extended feature-set we were able to achieve a high performance in terms of S1-score. The variation between the range of values was so little that we can

safely conclude that advance prediction is possible several minutes prior to seizure onset, given a well composed and high-quality set of features, with little compromise on Sensitivity and no compromise on Specificity and Accuracy.

7.6 General Discussion on Predicting Epileptic Seizures in Advance

This chapter provided extensive analysis on the predictability of epileptic seizures in advance of their onset. The Delete predictive algorithm was introduced and used in several experimental settings.

The Accuracy and Specificity measures remained consistently high for all timepoints and during all experiments, while Sensitivity and S1-Score values were more variable. The mean S1-score was in the worst case 71.07% which is considerably high. Advance prediction exceeded onset detection on single-channel datasets; for other experiments, the higher range advance time-step produced results very close to that of the seizure onset detection. The extended feature-set which was constructed in chapter 6 yielded the worst performance, indicating that the noise and redundancy in the data hinders advance prediction, while seizure onset detection is not considerably different from the benchmark. A subset of 14 high ranked features of the extended feature-set however, yielded the highest performance for seizure detection and advance prediction.

The predictive experiments were repeated with other experimental settings (See Appendix B), the results of which are presented in Figure 7.12. The ReliefF-Delete on a 14 dimensional feature subset of the extended feature-set produced the highest outcome with maximum advance value of 96.30% at $t = 1$ and 96.13% at $t = 8$; the latter was very close to the performance of seizure onset detection (96.18%). This outcome was followed, with little variation, by mRMR-Delete on the subset of extended feature-set. Rename-Delete on a subset of extended features had the 3rd highest performance. Best-ReliefF Delete and Best-mRMR Delete also performed exceptionally well at respectively 4th and 5th place.

Performance was relatively smooth over time for the majority of experimental settings, in particular the Delete cases of ‘Best’ channels and Extended Subsets. Time-steps 5 and 18 produced noticeably poor outcome in the majority of the cases while moments 1, 2 and 8 were amongst the highest advance prediction values in most cases.

High performance windows were not consistent among the experiments, but prediction seemed to vary in waveforms.

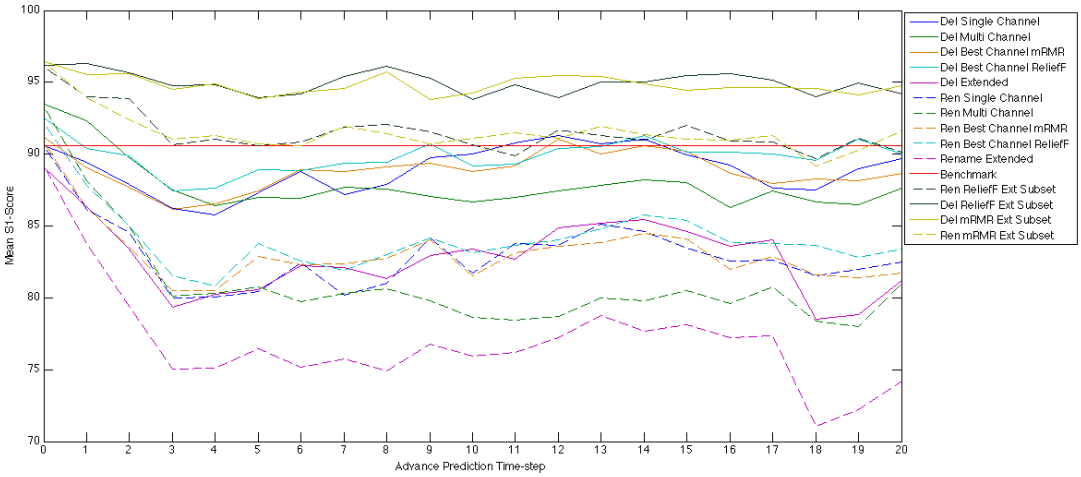


Figure 7.12 Summary of all Advance Seizure Prediction experiments conducted on 18 patients form the Freiburg EEG database – The reported results are for S1-Score over all advance prediction time-steps ranging from 0 – 20. The plots with the same color represent the same feature-set variation: The dashed line displays advance prediction by Rename and the solid line represents advance prediction by Delete. The solid red line is the benchmark value for S1-Score (14 feature default channel).

7.7 Conclusion

In this chapter we carried out a comprehensive analysis of the advance predictability of epileptic seizures. We developed the Delete predictive algorithm, which was used in a variety of experimental conditions. We concluded that predictability under optimal settings could yield a performance very similar to that of seizure onset prediction (i.e. t_1, t_2, t_8 and t_{16}). Prediction on a subset of the extended feature-set led to the highest performance with little variability among prediction windows. The multi-channel subset yielded a relatively low outcome, but the full set of extended features led to the worst performance.

The outcome of this study contributes further evidence of advance seizure predictability to the existing body of work, along with some best practices and their respective outcomes. More so, the results are statistically correct for real-life application in mind, yielding high Sensitivity and Specificity as well as Accuracy. The results also further confirm the superiority of our new feature-set over the original set of features we started out with, when used in optimal application settings.

Chapter 8
Multi-Patient Seizure Classification

The experiments presented in previous chapters addressed single-patient classification problems, where the training, validation and test-sets comprise invasive EEG recordings of individual patients. These patient-specific classifiers generally have high performance measures in default experimental setups and have thus far produced acceptable outcomes for the various experiments we have seen presented in preceding chapters. In this chapter we study the generalisation of these models against classifiers constructed on multiple Patient-Files.

In section **8.1** of this chapter we **motivate** the implementation of multi-patient classification. In section **8.2** we present results of **extensive multi-patient experiments,** where we evaluate the influences of the number of Patient-Files in the training-set, against the ability to generalise unseen data from the patients within the training-set, and unseen data from patients who were not included in the training-set. We refer to the former case (this is unseen data, but training has involved these patients) as just 'unseen-trained'; we refer to the latter case (training has involved no data at all from these patients) as 'zero-training'. In section **8.3** we review **special cases of the multi-patient classifiers** with a view to finding meaningful patterns in the classification models and their underlying Patient-Files. In section **8.4**, we present **new classification techniques** for multi-patient analysis. In section **8.5** we further **discuss the findings** of this chapter and we finally bring to a close with a **conclusion** in section **8.6**.

7.8 Motivation

In seizure detection studies presented in chapter 3, there were a few instances of multi-patient analysis studies while the main bulk of the research was focused on individualised seizure prediction. This theme is prevalent in the current trends of seizure-prediction studies. Many researchers believe that the characteristics of seizures greatly vary from patient to patient and can therefore not be generalised across multiple individuals. In addition to this, any research that has studied multi-patient seizure prediction, has produced poor outcomes compared to the individualised alternative,

further justifying the focus of research on improving the individualised seizure prediction, where the progress and potential impact seems more promising.

While the importance of improving patient-specific seizure prediction is undeniable, this should not deviate attention from the potential impact of multi-patient prediction research. In fact, evolving techniques of individualised detection can have a direct positive impact on improved design of more generic multi-patient classifiers. The reason multi-patient analysis deserves greater attention is two fold:

1) It is well understood that using more data tends to reduce the risk of over-fitting and tends to improve the quality of results on unseen data. Generally, machine learning algorithms perform better with more data. However, data at such large scales with significant potential improvement on learning is costly to gather, particularly when it is seizure data. If individualised predictors are to successfully work with the highest levels of Accuracy, an abundance of data is required which is not realistic with respect to the number of seizures a patient endures. Moreover, the number of seizures per patient is not comparable to the number of seizures that could be gathered from across patients with epilepsy. With the aid of effective multi-patient predictors, full use can be made of the extensive number of seizures accumulated across large numbers of patients in an effort to move towards better and more effective Big Data solutions.

2) Machine learning algorithms thrive on data. If we assume that enough number of seizures per patient can be obtained over time, so that classifiers improve for the individual patient, we are still faced with the problem of poor seizure management during the long period leading up to the point when enough seizure data have been collected. For a patient-specific seizure detector, a large number of seizures are required to build a classifier, and this will be from a very skewed data (i.e. training data that are heavily unbalanced in terms of ictal vs. non-ictal samples). Even then, further tuning will be required to improve the predictor with high Sensitivity and Specificity. A patient reported some two years on a trial of NeuroPace (Vachtsevanos 2003) where they endured several seizures before becoming 80-85% seizure free (Epilepsy.com 2010). Similar to most healthcare applications, seizure data are costly in terms of the quality of life of the patients. By making classifiers more generic, we can make use of existing seizure data across patients, with minimum requirement of fine-tuning on patient-specific seizures, and therefore, minimising the potential risks from triggering

seizures for data gathering purposes or poor seizure management as a results of on-going fine-tuning of the predictor.

In addition to the justification for building more powerful multi-patient seizure detectors, we also require advancement in technology to enable this. As mentioned earlier, the reported literature deems seizure detection an individualised study, due to the differences in seizure activity among patients. Furthermore, those multi-patient experiments that are reported have very poor performance in comparison to the alternative individualised predictors. This is while multi-subject research in similar application fields has performed better results. In the field of brain computer interface (BCI), (Fazli et al. 2009) improved their classifiers to allow for a better performance on unseen subjects, also known as zero-training classification, in order to eliminate the lengthy training requirement per participating subject. In the field of handwriting recognition (Lecun et al. 1998), where datasets are constructed of real or artificial multi-sourced data, results have reported errors as low as 0.23 (Ciresan et al. 2012). While these reports are mainly on Accuracy (as Sensitivity and Specificity have little implications on the performance of these particular methods), the results are very high and have been practically applied (in the case of BCI) where performance has been evaluated in terms of real-life Accuracy. Advancement in multi-source classification as such provides the suitable platform for building more powerful multi-patient seizure detectors.

In order to evaluate the generalisation ability of multi-patient predictors across a large number of patients, we start by conducting a full exhaustive analysis on all possible group settings for training-sets of the learner on the Freiburg EEG Database using a reliable machine learning algorithm, which we have used throughout this thesis. This gives us our benchmark performance on a large number of patients, which has not been evaluated at this scale in research presented in the literature. This also gives an understanding of where we stand in terms of generalisation with a powerful machine learning tool and how this generalisation is affected by changing the number of patients used in the training-set.

We then apply a number of machine learning algorithms that particularly suit multi-source problems as such, in order to study the possibility of improving the performance of multi-patient seizure detectors.

7.9 Multi-Patient Classification

So far in our experiments, we have only been concerned with patient-specific classifiers and their performance on respective Patient-Files. In this section, we review the effects of training classifiers on multiple patients. We also further evaluate the generalisation ability of these classifiers on unseen Patient-Files.

7.9.1 Data and Implementation

The data preparation for this series of experiments is similar to the default single-channel data preparation steps described in previous chapters.

In chapters 6 and 7, we introduced the core learning module, where classifiers were constructed separate from the data preparation and processing steps. In the multi-patient analysis, the same core learning module is used for classifier training and testing, though the pre-processing module, where data were prepared and segmented, varies from the pre-processing module previously presented.

All 21 Patient-Files are loaded into the pre-processing module and are normalised prior to further manipulation. The next step is the grouping of the Patient-Files with parameters g and l. The parameter g is the number of Patient-Files in the training-set and takes the value $1 \leq g \leq 20$. The case $g = 1$ denotes the single-patient default classifier used in previous chapters, and $g = 20$ is the Leave-One-Out setup, where all but one Patient-File are used to train the classifier and are ultimately tested on the single held-out Patient-File. The parameter l denotes the number of combinations of g Patient-Files required for the experiment. The default l for most experiments is 50, as calculating the entire possible combinations of 21 files with $g > 2$ is not possible in finite time. However, for $g = 1$, 2 & 20, where all possible combinations can be calculated, the values of l are respectively 21, 210 and 21. After determining values for g and l, for each combination of files in the training-group matrix, the following is executed:

1- the Patient-Files in the corresponding combination set are split into 70/30 random seed partitions.

2- The randomly selected 70% of each Patient-File from the combination set of files is added to the training-set of a learning-module and the 30% is re-labeled as the test-set under respective Patient-File names.

3- For each group of elements in the combination set, the processed training-set is fed into the learning module, where the Multi-class Support Vector Machine (described in chapter 2 and 6) is used as the classification algorithm. The parameters of the Multi-class SVM are fitted using a 5 fold cross validation (described in chapter 2).
4- After the learning model is constructed, each Patient-File including the 30% random partition of the training Patient-Files and the unseen Patient-Files are separately classified by the learner. The test result of each Patient-File is returned to the caller along with Accuracy, Sensitivity and Specificity outcome of the test classification.
5 - The outcome of the process is written to a log file; the training-set and test-sets are re-set and this procedure is repeated for *l* times.

The described procedure is repeated for each group *g*, that is, 20 times. It is worth noting that the combination set does not allow repeats, ensuring that the several occurrence of one classifier does not introduce bias when analyzing classification results.

The output of these experiments is the four performance measures of Accuracy, Sensitivity, Specificity and S1-Score of all models evaluated on each Patient-File.

7.9.2 Results for Multi-Patient Classification

The result of the multi-patient classification stages described in the previous section are summarised in Figure 8.1. On the X-axis, we see the number of single-channel Patient-Files *g* used to train the classifier at each stage of the experiment, starting from 1, indicating a single patient training mode, which we have used as the default mode thus far throughout this document. The highest value for *g* is 20, which indicates a Leave-One-Out setting for the experiment, where the classifier is trained on all Patient-Files except for a single Patient-File held out for testing. The Y-axis displays the mean S1-Score values in percentage. The S1-Score value at each stage is a mean of performance of ~50 multi-patient classifiers (~210 in the case of 2-Patient Classifiers, ~21 in the case of 1-Patient and 20-Patient classifies). The S1-Score plot solely reflects the mean result of the several trials on each mode on the 'zero-training' Patient-Files. This means that Patient-Files, which were used in training the model, were excluded from the summary results displayed in Figure 8.1 in order to reflect an unbiased result for the generalisation of each stage on unseen patients.

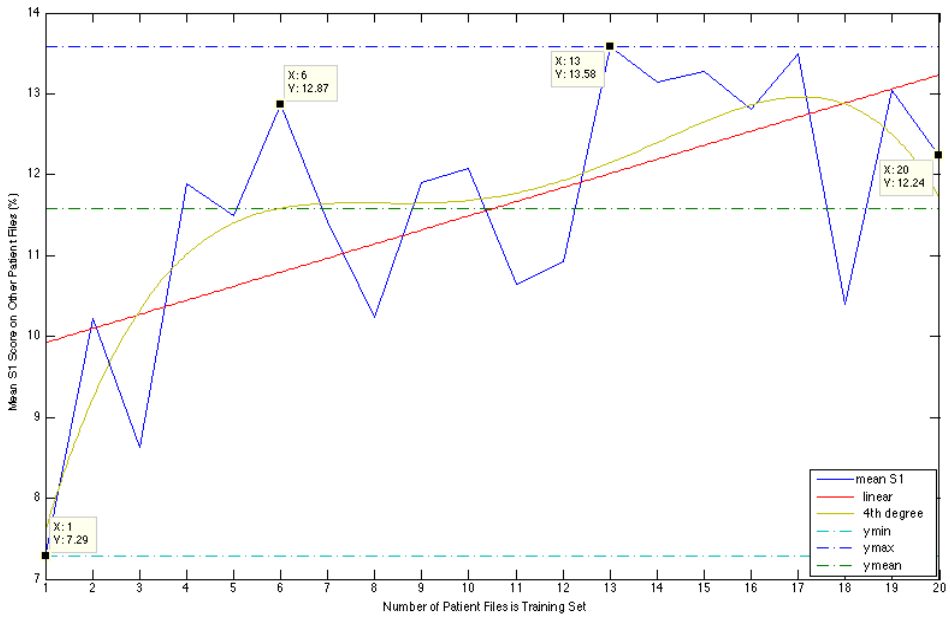


Figure 8.1 The mean S1-Score of the Multi-Patient Analysis on 'zero-training' Patient-Files: Across the X-axis, the number of Patient-Files in the training-set is displayed. The Blue line is the mean S1-Score of the classification of each group size on 'zero-training' data. The red line displays the linear fitting and the green curve is the 4th degree polynomial fitted to the main plot. Maximum, mean and minimum values of the mean S1-Score across group sizes are displayed respectively in dashed blue, green and cyan.

The results reveal that multi-patient generalisation is significantly low for training-sets comprising a single Patient-File. This is expected as the parameters in the model are fine-tuned on the single Patient-File, hence, the learner fails to correctly classify the values on other 'zero-training' Patient-Files, while yielding high performance values on the Patient-File it is trained on. However, as we increase the number of classifiers, the mean S1-Score seemingly increases, more so in some stages in comparison to others. In the case of 2 patients, the mean S1-Score rises to a higher 10.22% on 'zero-training' data. This is while the performance on the training files remains high. The S1-Score seemingly drops down at $g = 3$, however, it still remains higher than $g = 1$. From this point, there is an interesting peak at $g = 4$, where values are as high as 11.98%, followed by an insignificant dip at $g = 5$. The plot indicates another peak at $g = 6$ with a high value of 12.87%; this is approximately two times the S1-Score at $g = 1$.The performance then variably dips and peaks, not so far off the mean value at 11.58%. The maximum performance is at $g = 13$, with S1-Score of 13.58%. From this point through

to g = 17 values still remain high with minor variability. The performance then drops to below the mean at g =18 and then picks up again to higher values for g = 19 and g = 20 with respective values of 13.04% and 12.24%. From inspecting the results on the 'zero-training' Patient-Files we can identify 3 high performance regions: $g \in \{4,5,6\}$, $g \in \{13,14,15,16,17\}$ and $g \in \{19,20\}$. Performance in these subsets is higher than the 1-Patient-File classifiers, but there is little variability among the values in any specific range. The fitted linear equation shown in red displays an increase of mean S1-Score as more Patient-Files are used at a rate of $\%0.17437$. The curve reflects peaks at 2 of the maximal regions we mentioned earlier, and a projected decrease at the final maximal range. The mean of the S1-Score curve is at 11.89%, above which there is a higher density of the multi-patient classifiers.

It is worth noting that the S1-Score is a harmonic mean of the Specificity and Sensitivity of the classifiers. The Accuracy, on the other hand, is relatively higher throughout all stages (Figure 8.2), although it generally follows the pattern observed in S1-Score. Accuracy reflects a high performance in terms of an average machine learning algorithm, however, it is not sufficient for validating the efficacy of a seizure prediction model. Therefore, we look at S1-Score, which intuitively summarises the other two performance criteria of Sensitivity and Specificity

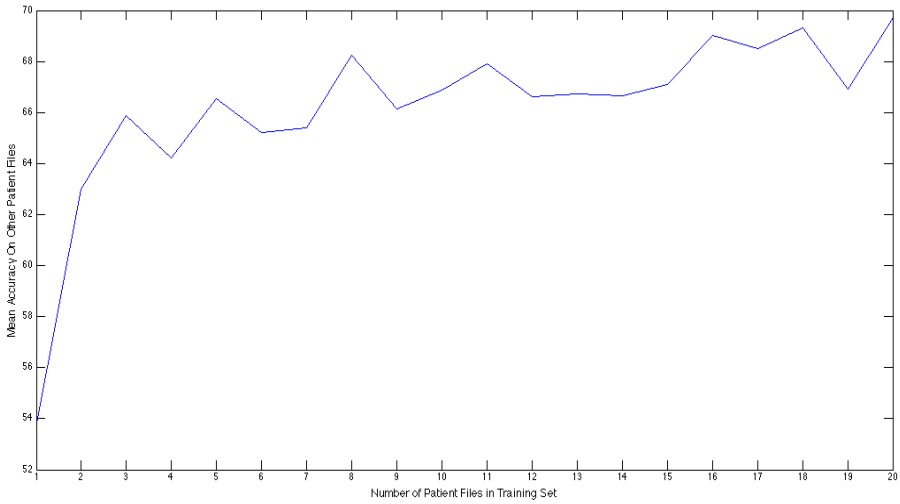


Figure 8.2. The mean Accuracy of the Multi-Patient Analysis on 'zero-training' Patient-files: The X-axis displays the number of Patient-Files in the training-set and the Y-axis displays the Accuracy averaged over all classifiers in the respective group.

The multi-patient classifiers were also used to classify unseen data from the corresponding training files. Figure 8.3 displays the classification results of 'unseen-trained' Patient-Files. The multi-patient classifiers were trained on 70% of each Patient-File in the training-set, and tested on the remaining 30%, which were unseen at the time of training ('unseen-trained'). The results in Figure 8.1 reveal an inverse trend to the zero-training Patient-Files. The highest S1-Score is at training-set of g = 1, and both linear and 4th degree polynomial trends display a monotonic decline in the performance as the size of the training-set is increased. Overall, the S1-Score decreases as g grows but it varies among several of the parameters. Some of the highest values are at g = 2, 3, 4, 5, 6, 9 which are above the mean S1-Score. The minimum value for S1-Score is 81.56% at g = 19. The plot in Figure 8.4 also displays a similar trend to Figure 8.3 for levels of Accuracy.

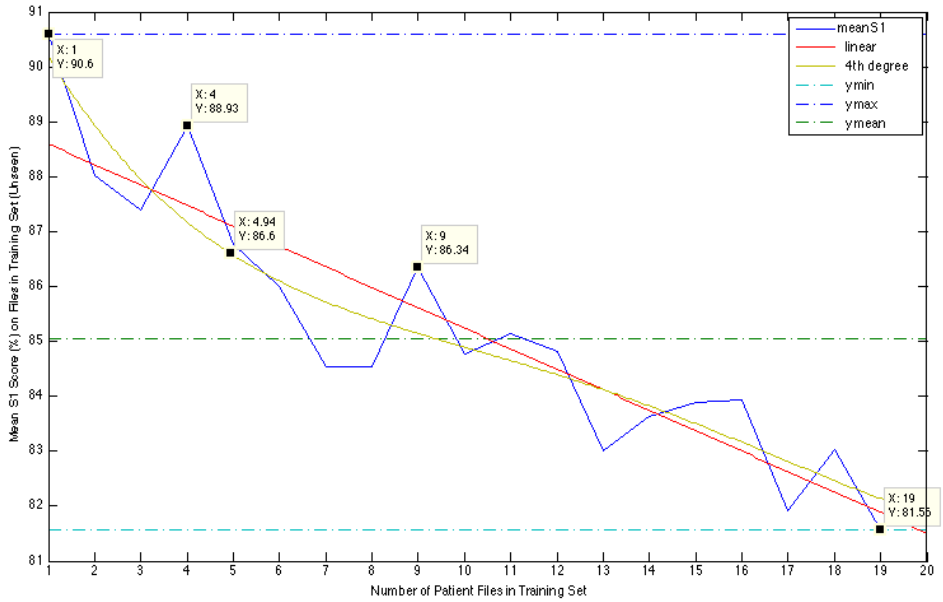


Figure 8.3 - The mean S1-Score of the Multi-Patient Analysis on 30% 'unseen-trained' Patient-Files: Across the X-axis, the number of Patient-Files in the training-set is displayed. The Blue line is the mean S1-Score of classification of each group size on unseen data. The red line displays the linear fitting and the green curve is the 4th degree polynomial fitted to the main plot. Maximum, mean and minimum values of the mean S1-Score across group sizes are displayed respectively in dashed blue, green and cyan.

The diagrams in Figures 8.1 - 8.4 suggest that in general, as the number of patients in a training-set grows, the performance on the unseen parts of the training-set decreases while generalisation on 'zero-training' patients improves. With respect to both results, we conclude that for the population of 21 patients, g = 4, 5, and 6 yield a relatively high performance on 'zero-training' Patient-Files as well as 'unseen-trained' Patient-Files.

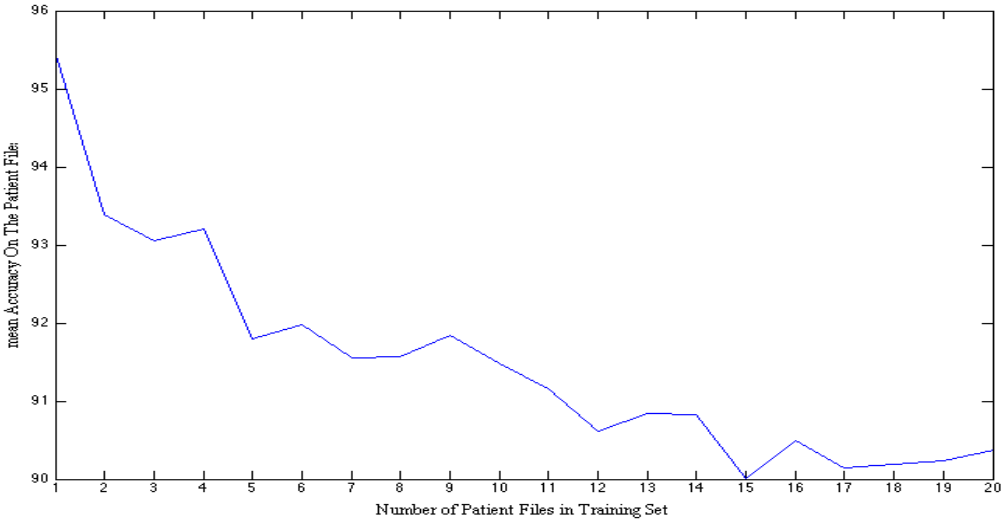


Figure 8.4. The mean Accuracy of the Multi-Patient Analysis on 30% 'unseen-trained' Patient-Files: The X-axis displays the number of Patient-Files in the training-set and the Y-axis displays the Accuracy averaged over all classifiers in the respective group.

7.9.3 Discussion of Results of Multi-Patient Classification

In this section we implemented a comprehensive set of Multiple Patient classifiers on various groups of training-sets. We evaluated the results in terms of S1-Score and Accuracy of the classifiers trained on various numbers of patients when tested on the 'zero-training' Patient-Files and on unseen 30% data from 'trained' Patient-Files (i.e. 'unseen-trained'). The variability among different values of g (number of patients in training-set) for both measures of Accuracy and S1-Score was relatively low. The results revealed that both S1-Score and Accuracy of 'zero-training' test-set increased as the number of Patient-Files in the training-set increased. This was while the same classifiers tested on 'unseen-trained' data decreased as the number of patients in the training-set increased. However, an optimum range for which both 'zero-training' and 'unseen-trained' results were above the respective mean value was identified as patient groups 4, 5, 6 and 9.

The insignificant variability between different values of *g* from both test-sets indicates that using the current settings on Multi-class SVM, multi-patient analysis on 'zero-training' files were not significantly improved in any single case, even at the highest value of $g = 20$. In general, the performance outcome of multi-patient analysis on 'unseen-trained' Patient-Files was significantly higher than 'zero-training' files. We conclude that under the current settings, though changes are observed at different levels of *g,* 'unseen-trained' and 'zero-training' Patient-Files do not vastly benefit form multi-patient analysis.

7.10 Patient-Specific Performance

In section 8.2 we presented a number of experiments where classifiers were trained on multiple patients and tested on unseen data from trained and untrained Patient-Files. The number of Patient-Files used in training classifiers ranged from 1 to *g*-1, where *g* is the number of Patient-Files in our dataset. In our study where $g = 21$, 1-Patient-File classifier denotes the default single-patient classification presented in previous chapters and 20-Patient-File classifier denotes a Leave-One-Out experiment, where in each round *g*-1 patients are used in constructing the model and 1 patient is held out for test data at each round. The exhaustive multi-patient experiments resulted in 1102 Multi-class SVM classification models. The computational time of building a model increases as the number of patients in the model increase. Additionally, all possible combinations of files (l) increases as the number of *g* Patient-Files used in constructing the classification model increase, the calculation of which was impossible in finite time for $g>2$. Therefore, for most groups with *g* number of Patient-Files, where $g \neq 1, 2 \,\&\, 20$, 50 random non-repeat combinations were sampled and used for building the classifiers. Information from the outcome of these experiments gave us an invaluable insight into the problem of multi-patient generalisation; however, questions about the Patient-Files and their effect on generalisation of the model have been neglected thus far. In order to answer such questions, we need to have a closer look at the results of multi-patient generalisation. Since $g = 2$ and $g = 20$ experiments have been conducted on all possible combinations of Patient-Files, they are the most suitable for finding such patient-centric effects.

We start from g = 2, where all possible combinations of 2-Patient-Files were used, resulting in a total of 210 classifiers. Each constructed classification model was tested on 'zero-training' Patient-Files and results were recorded in terms of Accuracy, Sensitivity and Specificity. Accuracy and Specificity values were relatively high for tests on all 'zero-training' data; Sensitivity was however highly variable. We use the S1-Score, which is the harmonic mean of Sensitivity and Specificity as it justly reflects changes in both values. The bar chart in Figure 8.5 illustrates the performance of Patient-Files when used to train classification models. In the X-axis we see the Patient-File index p, which ranges from 1 to 21. For patient p, we average the S1-Score of the classifiers which patient p has been used to train (in total 20 classifiers). The mean S1-Score is measured over the performance of the classifier on 'zero-training' Patient-Files. The Patient-Files that were used as training data are excluded from these results in order to have an unbiased and realistic view of the outcome of the model on 'zero-training' patients. The minimum performance is on models trained on patient 14, with S1-Score 1.82%. The mean S1-Score is 10.22% and the standard deviation is 4.75% showing relatively low variation between Patient-File generalisations. The median is 10.56%. The maximum is 19.46% for p =12. Patients 12, 5, 13, 2, 18, 7, 1, 9, 8, 11, 16 and 10 are above the mean while patients 14, 3, 4, 15, 20, 6, 17,19 and 21 are below the mean.

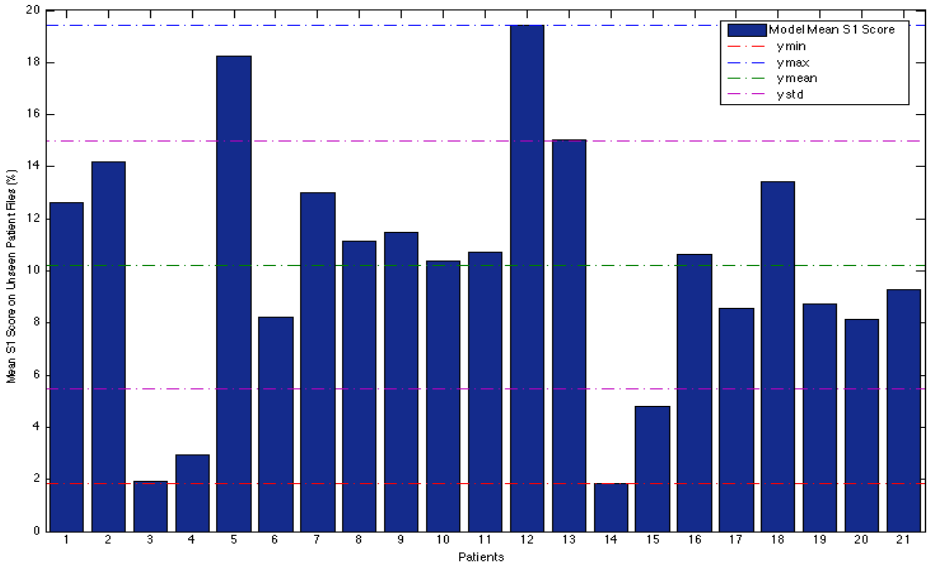
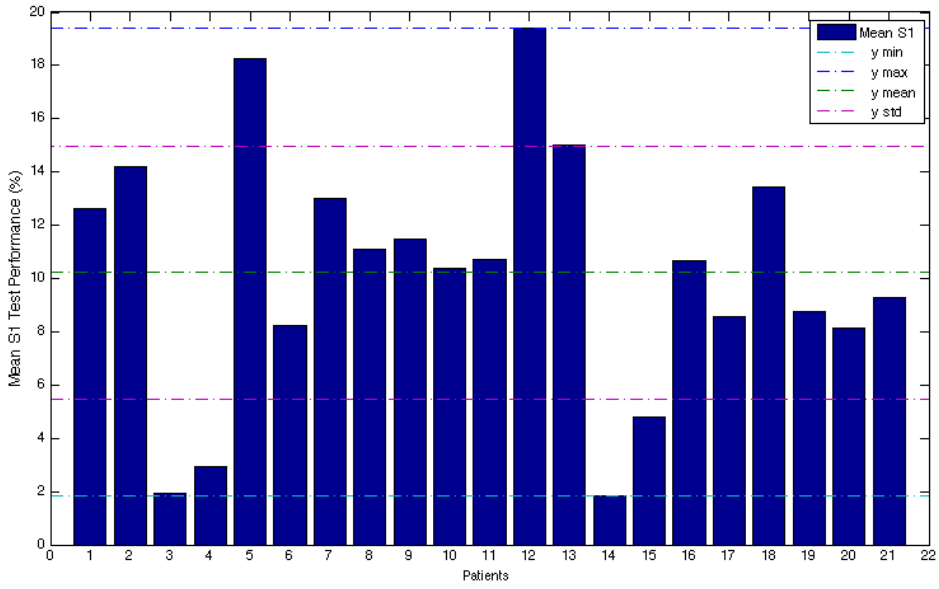


Figure 8.5 The Generalisation Ability of Patient-Files when used as Training-set: These results are reported for the g = 2 Multi-Patient Analysis where the X-axis indicates the Patient name and the bars display average S1-Score of classifiers the respective patient was used in, tested on 'zero-training' patients.

In Figure 8.6 we illustrate the mean S1-Score of models tested on patient *p*, where *p* is on the X-axis and is within the range [1, 21]. The bar for patient *p* represents the mean outcome of test S1-Score when models (excluding those which were built on the data) classified the corresponding data.

It is useful to see if there are any correlations between the generalisation capacity of a patient and its ability to be generalised well by other models. It is also interesting to see if there are remarkable clusters of patients that are easier to classify correctly. It is worth noting, that the average Accuracy and Specificity of the tests on each Patient-File *p* is relatively high and were therefore not considered in this analysis.



8.6 The mean S1-Score of Patient-files when tested as 'zero-training' patients: The results are from *g* = 2 multi-patient classifiers where mean S1-Score is reported on the average cases where the Patient-File was tested as 'zero-training' test-set by classifiers trained on other patients. The X-axis displays the patient name and the bars are average S1-Score for the respective patient.

From the chart in Figure 8.6, it is evident that there is high variability among different patients, with some values as low as 0% (*p* = 11) and others as high as 21.13% (*p* = 3). The mean is 10.22% with standard deviation of 5.69%. The median value of S1-Score is 8.82%, which is below the mean, indicating that there is higher density of Patient-Files with less than average performance in the population. The Patient-Files with measures above the mean are *p* =3, 4, 20, 9, 14, 13, 12, 6, 15, 17 and the Patient-Files with performance measure below the mean are *p* = 11, 1, 19, 2, 16, 8, 7, 10, 18, 5. By

comparing Figures 8.5 and 8.6, we can see that Patient-Files whom are not generalised correctly by classifiers of other Patient-Files, when used in constructing 2-Patient-File classifiers, have a high capability to generalise other 'zero-training' Patient-Files.

7.10.1 Impact of Feature-set on Multi-Patient Analysis of $g = 2$

We repeated the multi-patient training and classification for all combinations of 21 patients with group-size $g = 2$ on three different settings of 'Best' mRMR, 'Best' ReliefF and 'Extended' feature-set for each Patient-File (See Appendix B and C). Further information about the amended feature-sets can be found in chapters 6 and 7. The results of these experiments are listed in Table 8.1. The results are reported for the mean S1-Score and Accuracy on patients that were not in the training-set (i.e. 'zero-training') and on the 'trained' data which comprise 30% unseen data from the Patient-Files in the training-set (i.e. 'unseen-trained'). The results reveal that while the extended feature-set produced the highest outcome on the 'unseen-trained' Patient-Files, it yielded the worst outcome, on the unseen 'zero-training' Patient-Files, deeming the feature-set most suitable for generalising over 'unseen-trained' patients and highly unsuitable for generalising over 'zero-training' patients. The best performance on 'zero-training' patients is yielded by the Best ReliefF feature-set, where each Patient-File comprises 14 features of the channel ranked 'best' by the ReliefF feature selection algorithm. The same feature-set however yields a relatively lower outcome on the 'unseen-trained' instances. Overall, the feature-sets which generalise well to 'zero-training' patients, perform relatively poor on the 'unseen-trained' patients and vice versa, suggesting that depending on the objective of learning, certain feature-sets perform better for particular classification objectives.

	Single-Channel		**Best mRMR**		**Best ReliefF**		**Extended Channel**	
	ZT	UT	ZT	UT	ZT	UT	ZT	UT
S1-Score	10.22	88.02	11.11	62.84	19.07	24.39	6.29	88.80
Accuracy	63.01	93.38	61.60	83.80	64.20	67.86	80.02	93.94

Table 8.1. The Multi-Patient Analysis for $g = 2$ on various feature-sets: 'Zero-training' (ZT) and 'unseen-trained' (UT) Patient-Files were tested by Multi-Patient Classifiers of training-group 2,

constructed on single-channel, 'best' mRMR channel, 'best' ReliefF channel and extended feature-sets. Results are reported as Mean S1-Score and mean Accuracy for each experimental setup.

7.10.2 Multiple-Patient Classification on Training-set with 20 patients

The multi-patient classifiers tested on 20 Patient-Files are an important case of multiple patient analysis known as Leave-One-Out. In this case, 21 classifiers were built, each on a unique combination of 20 Patient-Files (from a total of 21 available) and tested on 30% unseen data of these 'trained' Patient-Files as well as on the left out 'zero-training' Patient-File. The results are listed in Table 8.2 where S1-Scores are color-coded based on a predefined range. The range [80% - 100%] is coded in pink, [60% - 79.99%] is coded in green, [40% - 59.99%] is coded in blue and S1-Scores below 40% are indicated as grey. The single Leave-One-Out test result is along the diameter, distinguished from other results with thick borders.

The results reveal that except for Leave-One-Out values 18 and 13, which are ~40%, other 'zero-training' S1-Scores are very low in most cases. The results also reveal a variable pattern of S1-Scores for the 'unseen-trained' test results for each patient. Patient-file 1 displays the poorest performance with 0 values in the range [80%, 100%], 17 within the range [60%, 79.99%] and 3 within the range [40%, 59.99%] when tested as the 'unseen-trained' test-set. The best 'unseen-trained' performance is for patients 2, 4, 17, 8 & 21 where all results are in the range [80%, 100%].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	75.88	97.10	80.04	92.41	90.08	83.48	91.86	89.18	87.73	80.99	84.22	93.17	81.73	78.04	88.21	90.84	94.63	81.95	84.40	81.08	0.00
2	74.93	96.44	75.30	91.32	91.25	77.03	86.92	87.63	90.12	81.80	82.98	91.04	83.46	75.37	88.60	83.16	92.39	83.87	86.44	0.00	87.19
3	66.64	90.71	69.96	91.07	82.41	75.80	89.13	91.56	77.63	81.12	45.71	90.62	63.83	75.40	85.89	81.57	91.66	83.39	9.36	76.30	81.60
4	65.78	94.49	68.13	92.39	86.24	74.97	90.43	87.23	83.96	86.08	56.29	87.96	78.75	79.97	78.46	71.14	92.98	47.98	87.27	80.43	81.52
5	61.76	95.22	71.36	90.49	83.94	62.37	86.46	83.88	85.91	85.48	61.24	92.72	73.31	78.52	88.78	87.59	0.75	83.76	78.25	85.38	85.02
6	65.73	95.99	61.08	94.94	81.18	73.61	91.40	95.33	88.04	86.81	64.25	89.61	65.85	81.37	86.47	36.63	91.05	81.27	91.39	79.58	82.94
7	69.28	96.79	75.81	93.47	84.89	73.12	87.54	96.52	87.24	80.63	63.79	94.21	67.93	81.39	7.58	79.44	88.14	73.56	79.92	83.73	84.78
8	70.89	93.29	66.26	92.94	80.60	69.30	85.63	94.81	76.25	79.50	69.10	88.93	81.59	0.00	87.94	75.51	89.45	79.38	85.32	76.07	86.24
9	60.36	96.00	74.16	95.98	78.06	68.69	76.89	84.58	79.39	81.48	57.01	91.47	55.97	72.42	81.28	83.05	92.36	77.46	76.22	77.52	82.89
10	77.58	93.32	74.93	94.31	82.46	77.21	87.50	89.18	80.62	78.76	65.18	21.79	78.55	76.32	83.17	81.38	93.00	82.79	82.45	81.09	84.08
11	55.56	96.64	64.70	90.43	83.19	72.58	81.43	90.83	80.45	81.51	0.00	88.10	75.42	79.89	81.18	82.47	89.64	74.79	79.52	74.71	82.06
12	69.17	97.88	72.15	88.83	90.29	67.15	88.24	97.73	84.56	1.44	73.68	87.09	68.92	77.32	85.33	82.55	94.45	84.95	89.61	82.23	85.94
13	52.17	87.99	74.49	90.11	87.14	75.09	85.96	93.67	16.73	82.39	59.91	84.13	81.40	77.88	84.28	78.85	89.50	74.64	80.62	81.68	87.50
14	72.50	96.27	74.04	94.41	87.97	76.90	87.78	3.28	84.02	80.16	74.38	89.70	83.92	80.28	85.54	83.01	91.79	83.94	91.04	82.15	88.06
15	68.70	96.13	77.57	96.34	89.89	74.90	2.87	80.23	88.30	83.45	76.01	73.05	89.10	84.66	85.07	84.60	95.27	86.82	88.52	83.18	84.86
16	60.28	97.10	75.56	91.41	84.47	0.00	86.18	90.68	77.51	85.09	57.39	85.10	79.72	76.78	83.80	66.51	90.81	76.36	77.98	85.28	81.05
17	62.66	94.59	75.39	92.80	17.15	79.91	84.31	94.48	82.33	81.03	61.76	86.27	71.83	77.24	74.80	75.42	91.33	69.47	86.79	82.83	82.59
18	59.61	98.03	72.70	0.00	86.06	80.50	84.75	89.81	88.73	82.23	72.89	89.52	84.07	76.02	90.75	84.82	90.76	80.09	86.06	85.04	86.84
19	76.68	98.58	15.41	89.48	91.52	79.45	88.04	94.01	83.95	75.24	71.61	93.62	86.54	78.54	82.86	86.20	92.50	82.01	85.43	85.51	82.51
20	60.28	20.09	70.84	92.33	83.28	75.74	86.00	90.18	79.62	73.70	62.86	87.09	70.15	79.54	84.75	82.20	89.79	80.05	82.57	86.98	81.94
21	0.00	93.05	73.86	87.41	75.76	69.61	85.62	90.47	82.49	83.60	58.43	94.92	69.66	78.07	88.40	78.78	88.33	67.73	80.66	79.13	88.43
P	0	20	1	20	18	2	19	20	15	16	2	19	7	4	18	13	20	11	15	14	20
G	17	0	19	0	2	18	1	0	5	4	12	1	12	16	2	7	0	8	5	6	0
B	3	0	0	0	0	0	0	0	0	0	6	0	1	0	0	0	0	1	0	0	0

Table 8.2 The S1-Score of the leave-one-out multi-patient experiment (g = 20): The pink cells are S1-Scores within the range [80,100], green cells are in the [60, 79.9999] range; blue cells are in range [40, 59.9999] and grey cells are in the range [0-39.999]. The three bottom rows indicate the number of pink, green and blue cells for each patient.

7.10.3 Discussion on Patient-Specific Performance

The multi-patient analysis were conducted on all Patient-File combinations for the two special cases of $g = 2$ and $g = 20$. By closely observing results from $g = 2$, we concluded that patients vary in how well they can be generalised by other Patient-Files and how well classifiers trained on them can generalise other patients. We also saw that in both generalisation powers (as classifier and as test-set), patients were generally within the same range of S1-Score indicating that the two are proportional. However, there is no observable correlation between the ability of a single patient to generalise and be generalised. This is particularly important when building multi-patient analysis tools. Those patients that are not generalised well are best implemented as an individualised seizure detection method, and Patient-Files whose classifiers generalise well are particularly suitable for multi-patient analysis. This suggests that prior to multi-patient analysis, an initial 'screening' may be useful in order to determine whether the tested patients can be generalised and whether the candidate training files can build generic classifiers. The difference in generalisation capability amongst patients can be linked to the inherent characteristics of the seizures of that patient. In $g = 20$, we see those files that do not generalise well in 'unseen-trained' classification, also perform poorly in the 'zero-training' instances, suggesting that generalisation of a patient is tractable in trained and zero-training scenarios.

The multi-patient seizure detection for $g = 2$ were further examined on three experimental settings: single-channel feature-set derived from 1) ReliefF 2) mRMR and 3) extended feature-set derived from all channels. The results were reported for 'zero-training' and 'unseen-trained' test-sets and were compared against the single-channel original feature-set. The results revealed that for the zero-training test-sets, 'best' ReliefF channel yielded the highest S1-Score and Accuracy, while for the 'unseen-trained' test-set, the extended feature-set comprising 204 features across all channel recording, produced a slightly higher performance to that of single-channel. This suggests that for 'zero-training' classification of multiple patients, a smaller, well-selected, highly discriminative feature-set can produce the best outcome while for 'unseen-trained' Patient-Files, a large number of features have a better ability of

building a highly generic model across patients. This is while the full-extended feature-set benchmark yielded a lower S1-Score and Accuracy on the single-patient case.

7.11 Improving Multi-Patient Analysis

The results from the Multiclass SVM on various combinations of Patient-Files (section 8.2) revealed that while generalisation of 'zero-training' patients improve as we add more patients to the training-set, the generalisation of the 'unseen-trained' data of the trained patients decreases, indicating a trade-off between 'zero-training' and 'unseen-trained' classification. For both scenarios, the change in S1-Score was not significantly large, though notable. For 'zero-training' values ranged between [7.29%, 13.58%] and for 'unseen-trained' ranged between [81.59%, 90.60%]. These results indicate that although the S1-Score of the trained patients decreases as g grows, the change is relatively small. The same holds for the 'zero-training' patient generalisation.

The minimum S1-Score of the multi-patient classification on 'unseen-trained' data is relatively high, however, the maximum S1-Score on 'zero-training' Patient-File is significantly low for real time application. In this section, we experiment with a number of machine learning algorithms in order to improve the S1-Score on 'zero-training' Patient-Files.

7.11.1 Transforming the learning problem: Binary vs. Multi-class

In the experiments mentioned so far, each dataset comprises 4 labels: ictal, pre-ictal, post-ictal and inter-ictal. The four labels were penalised in the Multi-class SVM architecture as described in chapter 6, in order to reduce the effects of the unbalanced dataset, while introducing higher resolution seizure states through the assumption that pre-seizure and post-seizure activity are different from the inter-ictal seizure activity (Mormann et al. 2007), and therefore merit distinct class labels. For the remainder of this chapter, we simplify the seizure detection problem to a binary classification task, which our target machine learning algorithms are able to solve more efficiently. The ictal data are labeled as '1' and the inter-ictal data are labeled as '-1' or '0', depending on the underlying machine learning algorithm. The skewness in dataset is handled via algorithm-specific solution as well as the data up-sampling method SMOTE (Chawla et al. 2002).

7.11.2 Handling Skewness in the Binary Dataset

By changing the multi-class problem to a binary classification problem, we have introduced further skewness to the dataset: for 1 hour of EEG data, there are 3 minutes of ictal and 57 minutes of inter-ictal instances, posing problems of incorrect classification of instances due to high levels of skewness; the skewness in the data may cause a learning model that classifies all states as inter-ictal yield ~86% Accuracy, which is clearly an erroneous and misleading figure. One way round this problem is by introducing Sensitivity and Specificity and trading off between the two measures, with respect to the dataset, and changing the cost-function where possible to favor one measure over the other. Another way of handling skewness in the dataset is by introducing balance between the two classes by either i) reducing the number of instances with the 'majority' class-labels otherwise known as under-sampling ii) increasing the dataset with artificial copies of the under-represented class-labels referred to as up-sampling (Chawla et al. 2002). Due to the scarcity of our EEG data, we have used the up-sampling method in addition to modifying the cost function where possible, to amend the threshold for Sensitivity and Specificity.

The up-sampling algorithm used in our experiments is a Matlab package called SMOTE (Synthetic minority Over-sampling technique) (Chawla et al. 2002; Cs.cmu.edu 2002). It is based on the principle of over-sampling the minority class, by introducing synthetic instances derived from variations of the real minority instances with respect to the feature-space, a technique widely used in handwriting recognition. It uses a random selection of same-class nearest neighbors for each instance of the minority class in order to construct the synthetic data. The synthetic data are constructed by taking the difference between the sample feature-map of the nearest neighbors of the instance and multiplying it by a random number between 0 and 1. This allows for the decision boundary for the minority class to become more general with respect to current samples and neighboring samples in the dataset. The data are up-sampled according to the following algorithm:

```
more=1;
while more==1
      [fidata,filabel]=SMOTE(fidata, filabel);
      [r1,c1]=find(filabel==1);
      [rm1,cm1]=find(filabel==-1);
      if (length(rm1)<(length(r1)+100))
            more=0;
      end
end
```

The algorithm ensures that the difference between the two classes (i.e. -1 and 1) is ≤ 100, expanding the dataset to ~3-times the original size. The data are normalized according to the underlying machine learning algorithm (in this case between [-1, 1] prior to up-sampling.

7.11.3 Experimental Setup

The results from the Multi-class SVM multi-patient experiments revealed that for training group size g = 4, 5 & 6 a relatively high S1-Score can be achieved on the 'zero-training' Patient-Files in a considerably lower computational time compared to the slightly superior performance of the higher values of *g*, with minimum compromise on the performance on the 'unseen-trained' Patient-Files. The advance prediction and feature selection experiments of representatives of the three group sizes also revealed that patients 5 and 6 had the edge over patient 4. In order to compare and contrast the performance of our candidate multi-patient classifiers, we focus on $g = 5$ as it yielded some of the higher performances and also has less computation overhead compared to higher-performance settings, with greater values of *g*.

In the following sections, we present the most important results of our multi-patient Experiments. For the sake of consistency and ease of comparison, we implement a $g = 5$ learner on a fixed combination of Patient-Files 1, 2, 3, 4 and 5 and test the learners on the 'zero-training' unseen Patient-Files 20 and 21.

Multi-Task Learning

In chapter 2 we described Multi-Task learning as a form of transferable learning, with the objective of learning across multiple related tasks. We further introduced cASO, convex Alternating Structural Optimisation (Zhou et al. 2012). In this chapter, we have implemented cASO in a number of settings to evaluate its effect on enhancing multi-patient learning on untrained Patient-Files. The computation time for multi-patient analysis by cASO is substantially low since learning of several Patient-Files is parallelised rather than the combinatorial learning of large sets of Patient-File data presented in Multi-class SVM learning. For this reason, we have implemented the cASO multi-patient experiment for $d \times 2$ where models are tested on Patient-Files 20 and 21. For each implemented algorithm a fixed combination of files was used for each group size. For instance group 2 contains Patient-Files 1 and 2; group 3 contains

Patient-Files 1, 2 and 3 and so on. Each group size was implemented in 7 distinguished experimental settings:

1- Single-Channel Skewed: The training–set is up-sampled using SMOTE and is therefore balanced while test-set is untouched. Both training and test-sets are single-channels. Results of this experiment are displayed in Figure 8.7.a.

2- Single-Channel balanced: The dataset comprises up-sampled, single-channel data for both training-set and test-set. The results are depicted in Figure 8.7.b.

3- Multi-Channel Skewed: All Patient-Files used in this experiment have features computed across multiple-channel EEG recordings. Each Patient-File in the training-set is independently up-sampled and the test-set remains untouched in that respect. Results are displayed in Figure 8.7.c.

4- Multi-Channel Balanced: The Patient-Files in both the training-set and test-set are independently up-sampled using SMOTE. The files contain features from across all 6 recorded EEG channels for each patient. Results are in Figure 8.7.d.

5- Extended Feature-set Skewed: the Patient-Files in this experiment comprise of 204 features presented in chapter 6. The Patient-Files in the training-set are up-sampled while the Patient-Files in the test-set remain skewed (Figure 8.7.e).

6- Extended Feature-set Balanced: The Patient-Files comprise the original feature-set as well as an extended 20 features along all 6 channels, summing up to a total of 204 features per Patient-File. Patient-files in the training-set and test-set are up-sampled. The results are displayed in Figure 8.7.f.

7- All Skewed Single-Channel: The Patient-Files in this experiment are the single-channel 14 features. Both training-set and test-set comprise skewed Patient-Files. Results of this experiment are presented in Figure 8.8.

For all the experimental settings, the Logistic cASO (Zhou et al. 2012) is implemented

for $2 \le g \le 19$, using the MALSAR (Multi-tAsk Learning via StructurAl Regularisation) library for Matlab (Public.asu.edu 2012). Each Patient-File in the training-set is separately normalised to values within the range [-1, 1]. The construction of classifiers for each value of *g* involves the loading and normalisation of the training Patient-Files and the transformation of the input file as a Matlab cell-array of input instances and target values, where each cell array represents a 'Task' in the Multi-Task Learning model. The 3 parameters of ρ_1 (task relatedness controlling parameter, where 0 represents non-relatedness), ρ_2 (L2-norm regularisation parameter) and k (dimension of shared structure between tasks) are selected through 5-fold cross validation on the following values:

```
rho1_range = [0.001 0.01 0.1 1 10 100 1000 10000];
rho2_range = [0.001 0.01 0.1 1 10 100 1000 10000];
k_range = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21];
```

The evaluation measure of the cross-validation stage as well as the test-set evaluation on files 20 and 21 is the S1-Score, which is the harmonic mean of Sensitivity and Specificity, and given the high level of Accuracy in previous examples, acts as the deciding performance measure. The outcome of the cross validation parameter selection is 3 parameters of ρ_1, ρ_2 and *k*, which are then fed into the Logistic_cASO method. The Logictis_cASO returns model parameters *W* and *C*, which build up the model, and parameter *M* which is $\Theta\Theta^{\mathrm{T}}$, where Θ is the shared subspace between classes. The weight vector *W* comprises a vector of cell-arrays for each file in the 'Task' list. The models are then used in the evaluation method where every instance of each test file is classified by a 'Task' model. This results in $g \times 2$ cell-array struct, where each element corresponds to the classification of the test Patient-file, by the classifier built on a particular Patient-File from the training-set. In the case of the skewed test-data (2, 4 and 6) and skewed training-set and test-set (7) the criteria of the evaluation function (and cross-validation in the case of 7) is also based on S1-Score, though the threshold for Sensitivity is modified to -0.8 instead of a normal threshold of -0.5; if the threshold is passed, the instance is classified as class-label 1. This modification entails that classification is more sensitive to instances of 1 (ictal data). This serves as an effective way of managing skewed data by adjusting the classification threshold of the positive instance.

```
y_pred = sign((X{t} * W(:, w))+C(1,w));
y_pred = (((X{t} * W(:, w))+C(1,w))>=-0.8);
```

In Figures 8.25 and 8.26 the displayed measures are mean S1-Score, maximum S1-Score and mean Accuracy of cASO tested on Patient-Files 20 and 21 for various values of *g*. The cASO constructs a classifier for each 'Task' (in this case Patient-File) in the training-set, which is built on a shared low dimensional feature map as described in chapter 2. Various Pertinent-files may classify test Patient-Files depending on how well the constructed classifier generalises the test file. We therefore report both mean and maximum S1-Score for each *g* as the majority of Patient-Files may produce the same outcome, while a particular classifier could potentially generalise better for the particular Patient-File.

In Figure 8.7, we display the results of the Multi-Task Learning experiments, numbered 1-6. The graphs contain mean S1-Scores, max S1-Score and Accuracy of separate classification of patients 20 and 21. The values on the X-axis indicate the number of Patient-Files used in the learning algorithm and the Y-axis shows the performance outcome in percentage.

The results across the balanced experiments are more consistent than the experiments conducted on skewed test-sets. For skewed examples, the mean Accuracy across most values of *g* are high for both patients 20 and 21 and they both follow a similar pattern throughout various values of *g,* indicating that variability between different experiment steps is not coincidental. The mean and maximum S1-Scores are however low and unchanged for the single-channel and multi-channel multi-patient analysis on the skewed test-set. The extended feature-set displays variability in the S1-Score across different values of *g,* particularly for $g = 5$, but in general, the values of S1-Score for this experiment are lower than the single-channel and multi-channel experiments.

The results also suggest that in the case of the extended feature-set on skewed test-sets, though performance is generally low, the overall performance on patient 21 is superior to that of patient 20, suggesting that the suitability of the feature-set can vary between patients. Among the skewed experiments, that is experiments 1, 3 and 5, single-channel and multi-channel lead to a relatively better performance, compared to the extended feature-set, with the maximum Accuracy for single-channel at 23.95% and

maximum Accuracy for multi-channel at 28.87% and respectively maximum S1-Score at 6.41% and 6.02%.

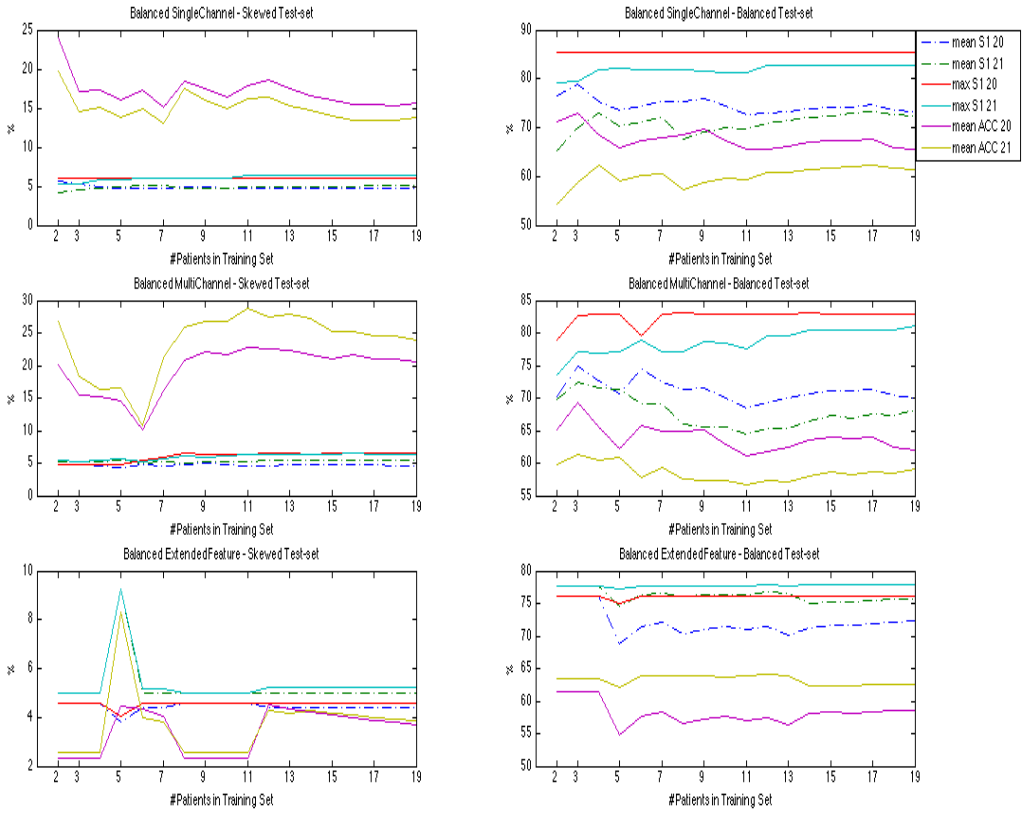


Figure 8.7: The Multi-Task Learning for values of *g* in the range [2, 19]. The classifiers at each step are constructed on a fixed combination of balanced Patient-Files. a) Displays result of MTL when trained on balanced single-channel and tested on skewed single-channel test-files b) MTL is trained and tested on balanced single-channel files. c) MTL is trained on multi-channel balanced and tested on single-channel skewed data. d) MTL is trained and tested on Multi-Channel skewed data e) MTL is trained on extended balanced and tested on extended skewed data f) MTL is trained and tested on extended balanced data.

The performance on the balanced dataset is a lot more coherent in comparison to the skewed files and generally within a higher range of values. The S1-Scores for the test patients are generally higher than the Accuracy, unlike what we have seen previously in the Multi-class SVM experiments. With the exception of the extended feature-set experiment, classification produces better results for patient 20 and once again, extended feature-set seemingly performs better on patient 21. The variation between

different values of g is consistent for both patients in each experiment and is generally highest at $g = 3$ with little notable variation for other values of g. The highest S1-Score in the balanced experiments is for patient 20 in the single-channel experiment.

The best overall performance is by patient 20 with single-channel balanced followed by patient 20 multi-channel MTL. Performance at $g = 5$ is relatively and consistently lower than the better performance observed for $g = 3$ & 4.

In Figure 8.8 the multi-patient cASO classifiers were constructed on skewed Patient-Files where the ictal label was under-represented in comparison to the inter-ictal labels. The classifiers for various values of g were tested on skewed test-files. The results reveal that Accuracy was high for both patients 20 and 21 for all values of g with respective mean Accuracy of 96.52% and 95.33%. The cASO classifiers however yield different S1-Score results for the two test-files. The results for patient 21 are considerably poor with mean S1-Score at 0% and maximum S1-Score predominantly at 0% with a slight rise to $0 < F1-Score \leq 6.05\%$ for $g>11$. The mean S1-Score for patient 20 is around 30% with little variability between steps and the maximum S1-Score has values in the range [50.32%, 66.18%]. The S1-Score particularly rises for $g>14$ and remains constant for other values of g, suggesting that the peak performance under these conditions is 66.18%. The difference in the range of the mean and maximum S1-Score for patient 20 suggests that cASO constructed on certain Patient-Files can perform superior to other (in this case the majority) of the Patient-Files. The difference between the outcome for Patient-Files 20 and 21 further suggests that optimum experimental settings are patient-specific.

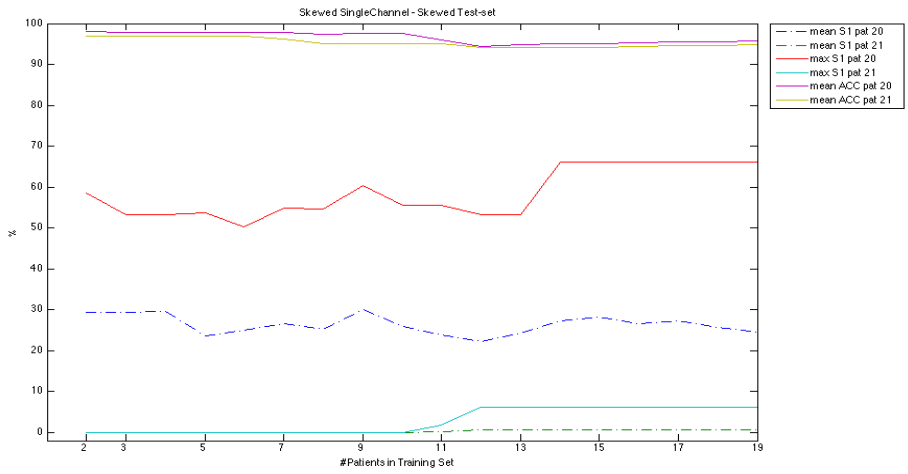


Figure 8.8 The Multi-Task Learning experiments for $g = 2$ through 19 and tested on Patient-Files 20 and 21: The training-set and test-set comprise skewed single-channel data.

Deep Belief Nets

Deep Belief Nets (DBN) are generics graphical representations of a learning problem (see chapter 2) (Arel et al. 2010; Bengio 2009). In this experiment, we use Deep Belief Nets to construct the multi-patient seizure detection task, with aid from the Matlab code from the MINIST handwriting recognition science papers by Ruslan Salakhutdinov and Geoff Hinton. The original program is available from their webpage (Cs.toronto.edu 2013).

Deep Belief Nets take long to train and have several hyper-parameters that need to be selected. There are various commonly used hyper-parameter selection methods in machine learning, one of which we have extensively used in this thesis is k-fold cross validation. Cross validation however is suggested not to be a suitable parameter selection algorithm for Deep Belief Nets, as the high number of hyper-parameters (some 11 hyper-parameters) coupled with the exhaustive cross-checking of every parameter combination over several folds, makes the computation impractically lengthy with little added benefit. With deep belief networks we also find that a few small changes do not cause convergence to optimum results. This is while in cross-validation, every combination of parameters is tested, with only a single parameter modified at each step. This entails that parameter selection is stuck in sub-optimal parameter space for a long time.

Parameter selection techniques for Restricted Boltzman Machines (RBM) and therefore, DBNs are not well-researched as the field has only recently attracted the attention of applied researchers, and the majority of reported experiments use heuristics to find parameters that best suit the task. This usually comes with ample experience of applying RBMs on various datasets. In the recipe book for parameter selection for RBMs (Hinton 2010), a number of heuristic steps have been mentioned which suggest a range of values for the hyper-parameters and signs of how and when they should be modified. The following are sets of most commonly tested range of values for the 11 hyper-parameters:

```
unsumaxepoch=[400 200 100 50]
sumaxepoch=[400 200 100 50]
unsulearning_rate=[0.1 0.01 0.001 0.0001];
sulearning_rate=[0.1 0.01 0.001 0.0001];
hid_layer_unit=[100 250 500 1000 1500 2000]
last_layer_unit=[500 1000 2000 4000 ]
```

```
tot_num_Layers=[3 4 5]
unsuinitial_momentum=[0.5 0.1 0.9 0.01]
weight_learning_rate=[0.5 0.1 0.9 0.01]
batch_size=[100 50 10]
mini_batch_size = [100 50 10]
```

Due to the nature of the task at hand, (building a multi-patient classifier comprising EEG records of 5 patients) it would not be practical to observe and fine-tune each hyper-parameter heuristically in the scope of this project. There are two principal recommended ways of finding an optimal solution (Bergstra et al. 2011); the first is with the aid of Gaussian Processes, which produce a probability distribution on each parameter, predicting which part of the distribution is most likely to produce the optimal solution. Necessary to this task is an initial archive of DBN training and testing logs in order to build the prior for the Gaussian Processes. This requires extensive computation time (still less than cross validation) but will produce a solution, which is closer to optimal. The computation cost of this process as well as the need for access to and understanding of numerous libraries for implementing the Gaussian Processes (which in itself requires parameter selection) proved to be out of the scope of this project.

The other solution that works far better than cross-validation is to randomly generate *n* number of parameter-vectors from a diverse range of values. The random selection of all parameters at once allows the solution to 'jump' between parameter regions and possibly converge to an optimal solution, unlike cross-validation, where the solution would iterate in a sub-optimal region for a nonessential amount of time.

We created 30 parameter-vectors for the 11 hyper-parameters, out of which, one vector could not be implemented due to computational restriction. The remaining 29 parameter-vectors are listed in Appendix A.

Deep belief nets, which are the layered representation of the raw dataset, are built on stacks of RBMs. The first layer (visible layer) is an RBM of the raw input, which is used to construct a representation of the input. This representation is chosen as the sample of $p(h^{(1)}|h^{(0)})$, which is the conditional distribution for the second layer units on the visible unit. This representation is used as input to the second unit. The second layer is an RBM based on the transformed data. The second layer and consecutive layers can be similarly transformed and represented for the total number of layers, propagating the data upwards. After constructing all layers (the layer-wise pre-training which is unsupervised), the parameters of the deep network are fine-tuned with

respect to the underlying classifier, so that the last layer is used to classify an input. The fine-tuning is carried out using a supervised gradient descent. This is known as the supervised or fine-tuning stage.

For each parameter-vector in our experiment, firstly the architecture of the network was specified using the `tot_num_layers`, a variable that determines the total number of layers in the architecture of the network including the visible layer. The number of units in a hidden layer and units in a visible layer were respectively determined by `hid_layer_unit` and `last_layer_unit`. The raw input data, that is, Patient-Files 1, 2, 3, 4 & 5 were loaded into the application and normalised within the range 0 and 1. The 5 training files were merged and their labels were stored in a separate matrix with $d \times 2$ elements: [0, 1] corresponds to ictal and [1, 0] corresponds to inter-ictal instances. The training matrix was then divided into random batches of length `batch_size`. The network was pre-trained on (d/`batch_size`) batches for `unsumaxepoch` times. Each epoch is a full pass through the network. The RBM was trained with `unsulearning_rate` as the learning rate and `weight_learning_rate` and `unsuinitial_momentum` parameters. The momentum method is a simple way of increasing the learning rate α by a factor of $1/(1-\alpha)$ without immediate effect on the gradient estimate. The initial momentum is randomly selected from the range of `unsuinitial_momentum` and increases to a predetermined final value of 0.9. After the pre-training stage, the fine-tuning, using gradient descent is conducted on the training-set and each of the two test files for `sumaxepoch` number of times. The gradient descent for each epoch uses a smaller number of batched data determined by `mini_batch_size`. The result is a matrix of measurements per epoch for each set of training-data, Patient-File 20 and Patient-File 21.

In Figures 8.9, 8.10 and 8.11 the outcome of the multi-patient DBN on the 29 parameter-vectors is displayed. Each Figure displays the maximum S1-Score and corresponding Accuracy of the test epochs. In Figure 8.9, we illustrate the training performance, where the classification was tested on a validation-set from the training data. The results reveal similar trends for both S1-Score and Accuracy across all parameter-groups with little variation between the ranges of the two measures. For most parameter-groups, performance measures are in a generally high range with the exception of parameter-group 4 where value for both measures is 0%. The second

lowest performance is at parameter-group 15 with Accuracy as low as 76.88%. Among the higher values are parameter groups 3 and 13 with respective S1-Score of 95.14% and 93.17%. The values of a number of these parameter-groups are listed in Table 8.3.

Parameter Group	3	4	15	16
unsumaxepoch	50	100	50	100
sumaxepoch	400	200	50	50
unsulearning_rate	0.01	0.001	0.0001	0.001
sulearning_rate	0.01	0.0001	0.001	0.001
hid_layer_unit	250	100	1500	1500
last_layer_unit	4000	1000	2000	500
tot_num_Layers	3	3	5	3
unsuinitial_momentum	0.1	0.1	0.1	0.1
weight_learning_rate	0.5	0.9	0.5	0.01
batch_size	10	50	50	50
mini_batch_size	10	100	100	10

Table 8.3 – The values of randomised hyper parameters for some of the important Deep Belief Nets, namely, DBN 3, 4, 15, and 16. For full list of hyper-parameters.

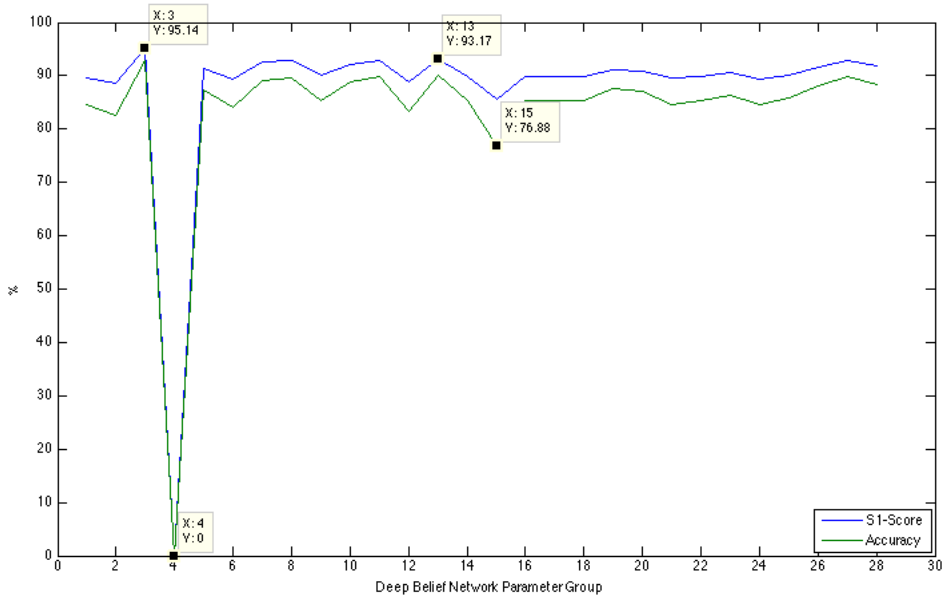


Figure 8.9 The Training Accuracy and S1-Score for the 29 Deep Belief Nets with Randomly selected hyper-parameters

The diagram in Figures 8.10 and 8.11 display classification results of the DBNs on test-set patients 20 and 21. Both classification outcomes display a similar general trend for S1-Score and Accuracy. The values of performance measures for patient 20 are in a generally higher range, with S1-Score holding higher values than Accuracy, and with little difference between the ranges of the two measures. The maximum S1-Score of 83.40% at parameter-group 16 is ~10% lower than the maximum S1-Score for the training-set results. For patient 21, the difference between Accuracy and S1-Score values is much more significant at ~13% but still both values resemble a similar trend for the various parameter-groups. For patient 21, the maximum outcome is at group 15, which is amongst the lower performance parameter-groups for patient 20 and the training-set performance. Though with the exception of groups 4, 12 and 15, the performance along different parameter groups is constant for patient 21.

Performance measures are 0% in parameter-group 4 for both Patient-Files, similar to that observed in the training-set performance. For other parameters, most values are generally high, and in the more extreme performance points, there are no common parameter-groups amongst the three different results, indicating that though performance is generally high, the optimum hyper-parameters may differ amongst patients, but not by large.

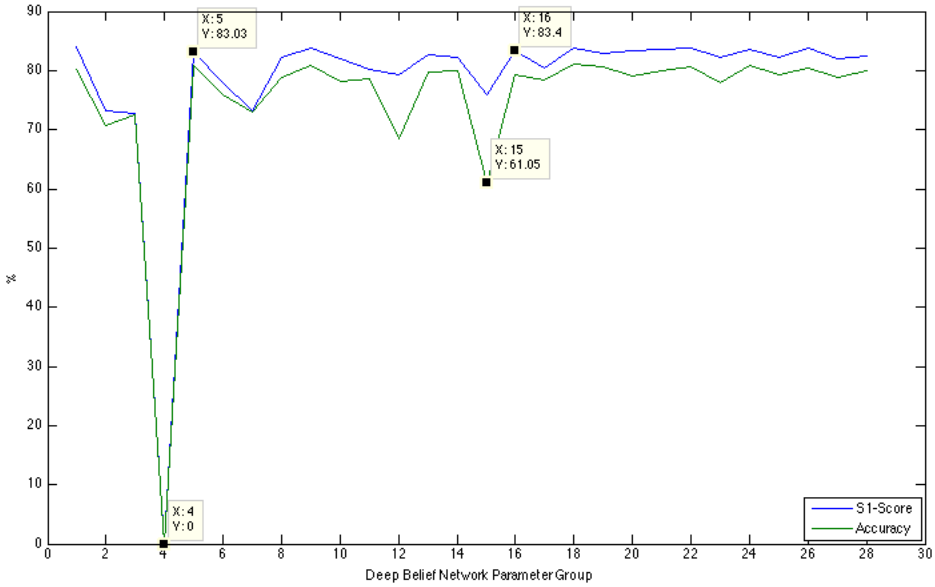


Figure 8.10 The Test-set Accuracy and S1-Score for the 29 Deep Belief Nets with Randomly selected hyper-parameters Tested on Patient 20.

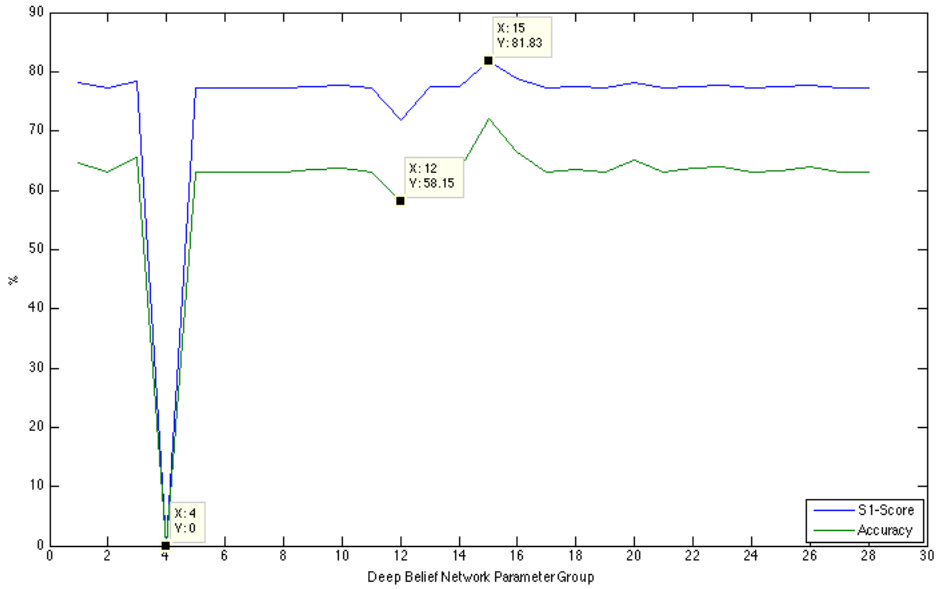


Figure 8.11 The Test-set Accuracy and S1-Score for the 29 Deep Belief Nets with Randomly selected hyper-parameters Tested on Patient 21

The parameter group with the highest outcome of both S1-Score and Accuracy among the two test results was selected and further tested on a skewed version of patients 20 and 21. This parameter group was group 16 which yielded the highest performance on the test data. The results, which are listed in Table 8.4, revealed that Accuracy was relatively high in the skewed test-sets with values 97.64% and 97.44% respectively for Patient-Files 20 and 21, while S1-Score was considerably low with values 4.18% and 6.75% respectively for patients 20 and 21.

7.11.4 Discussion on Improving Multi-Patient Classification

In this section we presented a number of machine learning algorithms that were particularly suitable for multi-source learning, namely Multi-Task Learning and Deep Belief Nets, in order to verify if multi-patient algorithms can be improved via an improved feature selection algorithm. We also transformed the learning task into a binary classification problem where the seizure state was the positive instance and inter-ictal, pre-ictal and post-ictal states were the negative instances.

The multi-task experiments were conducted on a fixed combination of $g = 5$ training-set and tested on two distinct Patient-Files. The experiments were repeated for Binary-SVM with altered classification and cross-validation function, where S1-Score and

Accuracy were the main measures to be maximised. A summary of results is presented in Table 8.4.

		Pat. 20 ACC	Pat. 20 S1	Pat. 21 ACC	Pat. 21 S1
MC-SVM		61.37	17.6	70.6	1.97
Binary SVM	Skewed	38.92	0	37.07	0
	Balanced	97.65	0	97.45	0
Deep Belief Networks	Skewed	97.64	4.18	97.44	6.75
	Balanced	83.84	81.06	81.83	72.06
Multi-Task Learning	Skewed (MC)	31.47	4.72	36.25	5.22
	Balanced (SC)	77.47	82.7	64	77.01
	Skewed Training	98.12	53.62	97.15	0

Table 8.4. The summary of Multi-Patient Analysis on a fixed combination of patients for g = 5. The results are reported in terms of S1-Score and Accuracy for the two test-files. The experiments were carried out on skewed and balanced test-sets where possible.

The Results revealed that Multi-class SVM on the same training and test-set yielded higher output compared to the cost-modified binary SVM on skewed and balanced data. We cannot make conclusive comments about the superiority of multi-task learning on skewed data in comparison to Multi-class SVM, as S1-Score is high while Accuracy is low. Deep Belief Nets produced the highest performance outcome on both skewed test-sets compared to all other classifiers. In the case of performance on balanced data, Deep Belief Nets have yielded the highest levels of S1-Score and Accuracy, shortly followed by Multi-Task Learning on balanced data. One particular interesting outcome is that the Multi-Task Learning tested and trained on skewed data produced the highest outcome on skewed data of Patient-File 20, despite performing poorly on Patient-File 21. This is consistent with our previous claim in the results section 8.4 that Models constructed from some patients can better classify certain similar patients. A closer observation of results from other multi-patient analysis further supports this finding. In the case of Multi-Task Learning particularly we see that the mean performance on the test-sets varies from the maximum performance. This further indicates that while the majority of the Patient-Files in the training-set produce similar outcome for a patient in test, at least one patient classifier yields significantly high results on the respective Patient-File in comparison to the others. This further suggests some inherent similarity among the high-performing training files and test files, which may be due to profile characteristics or seizure signal characteristics.

7.12 General Discussion

The work presented in this chapter was particularly focused on epileptic seizure detection on multiple patients. We presented a comprehensive analysis of multi-patient seizure detection for all possible training group sizes of patients in the Freiburg database. For each group size, we included a reasonable number of combinations of patients in the training-set and test-set. The results were reported in terms of classification of 'zero-training' and 'unseen-trained' patients. The zero-training patients in each experimental module were those patients whose files were not in the respective training-set while the 'unseen-trained' patients comprise 30% unseen data of the patients on which the respective classifier was trained. The results revealed that multi-patient classifiers for every possible combinatorial group tested on 'zero-training' Patient-Files yielded relatively poor S1-Score, while producing high levels of Accuracy. This outcome was expected and is consistent with the multi-patient seizure detection results presented in the literature. However, the multi-patient seizure classifiers tested on 'unseen-trained' Patient-Files produced high Accuracy and S1-Score values. These sets of results were generally lower than that reported in Costa et al. although this was expected as their results were reported for single algorithm runs of only 2 patients.

By observing both S1-Score and Accuracy as a function of g, where g is the number of Patient-Files in a training-set and is within the range [1, 20], we revealed that performance of 'zero-training' and 'unseen-trained' are negatively correlated: As g increases, the performance on 'zero-training' Patient-Files improves while that of 'unseen-training' declines. As g is decreased, the S1-Score of 'zero-training' is decreased while the 'unseen-trained' instances yield higher performance outcomes. This suggests a potential trade-off for the choice of g between improved 'zero-training' or 'unseen-trained' classification and will have to be decided with special consideration given to the application objective. With real-time applicability in mind, 'unseen-trained' does not particularly benefit from higher values of g, and since it is derived from the 70% of the relevant Patient-File, an individualised classification is favoured when such data are available. For cases when there is little or no seizure information available for a particular patient, multi-patient analysis with a high value of g can achieve high Accuracy and improved S1-Score (albeit relatively low).

We had a closer look at results of multi-patient analysis for $g = 2$ & 20 where patient specific trends are easily observable, as all possible combinations of training-set

files were studied. The results revealed a clear variability amongst patients in terms of 1) how well they could be generalised as 'zero-training' files 2) how well classifiers made on these files can generalise other 'zero-training' patients. The variability among results of patients suggests that multi-patient analysis may not bear a substantial effect on certain Patient-Files. This implies that particular consideration should be made about the success or failure of multi-patient analysis with respect to the number of patients the claim has been verified on. A poor outcome on a single patient may lead to the false claim that high-performance multi-patient analysis is implausible, while it may be the case that the characteristics of a particular patient makes it an outlier. This suggests a possible screening step prior to multi-patient analysis can be useful, to eliminate the files with low generalisation impact from the training-set and excluding files with little ability to be generalised from the test-set.

The multi-patient analysis for $g = 2$ was also evaluated on variations of feature-sets in order to evaluate the effect of using several different types of features. The outcome revealed that the features derived from the 'best' ReliefF EEG channel produced the highest outcome on 'zero-training' patients while the extended feature-set yielded the highest outcome on 'unseen-trained' files. This indicates that fewer highly discriminative features are particularly useful for zero-training generalisation, while a large number of features extracted from multiple channels for multiple patients (in this case 2) yield best generalisation for 'unseen-trained' Patient-Files.

We further improved the performance of multi-patient seizure detection by introducing two machine learning algorithms suitable for multi-source or transfer learning problems; these are Deep Belief Nets and Multi-Task Learning. The dataset was also transformed so that ictal data were labelled as positive instances while inter-ictal, pre-ictal and post-ictal data were labelled as negative instance, hence, posing a binary classification problem rather than a multi-classification problem. For MTL we used convex Alternative Structural Optimisation (cASO) for $g = 2$ though $g = 19$, on a fixed combination of files for each value of g (rather than the various combinations in the comprehensive multi-patient analysis). The cASO for each combination in g was trained on skewed and balanced training-files. The skewed classifier was tested on skewed data and the balanced classifier was tested on skewed and balanced data. The test-set in all cases were skewed/balanced variations of Patient-Files 20 and 21. The results revealed that balanced test-set yielded the highest outcome in single-channel

feature-set and skewed test-set performed the best in the multi-channel setup. The results also revealed that Patient-File 20 had the edge over Patient-File 21 in most cases. The skewed test-files yielded lower results than the respective balanced test-sets.

A fixed combination of 5 files, which also yielded high outcomes in the Multi-class SVM, was also used to train various DBNs with randomly selected hyper-parameters. The 29 Deep Belief Nets were pre-trained and fine-tuned on balanced training-sets. They were then tested on skewed and balanced training files. The results revealed that while for a number of the Deep Belief Nets performance was higher than others, this difference was not greatly significant with the majority of Deep belief nets yielding constant outcomes, with the exception of one network, which performed poorly on the training-set and test-sets. Moreover, these 'high-performance' cases were inconsistent across networks in the test-set and training-set, leading to inconclusive results with respect to the best performance set of hyper-parameters. For the purposes of comparison with other algorithms, the highest hyper-parameter-set from combined results of test-set was selected.

Comparison to other multi-patient algorithms on a fixed combination of 5 patients and a fixed test-set, where results were averaged over 10 runs of each learning algorithm, revealed that DBNs yielded the highest performance on both skewed and balanced data when training on balanced data. The Multi-Task Learning cASO trained and tested on skewed data also performed well on one patient while led to a poorer performance on the other Patient-File. This and other results from MTL and DBNs suggest that not only generalisation ability varies among patients, more so, certain training/test combinations of patients yields higher performance outcome in comparison to other combinations. This suggests that some inherent similarity exists between certain patients that can lead to a better training/test combination of those patients.

7.13 Conclusion

In this chapter, we conducted a comprehensive analysis of multi-patient seizure detection systems, as a study to this extent had not been conducted. In our findings we revealed that the number of patients in a machine learning training-set could directly impact the generalisation ability of 'zero-training' and 'unseen-trained' Patient-Files. We further observed that there is a trade-off between generalising for a completely unseen patient and generalising for a patient whose data was involved in training: as the number of patients in a training-set grows, the performance of 'zero-training' Patient-Files improves while the performance of 'unseen-trained' instances decreases.

We also observed variability among patients in terms of ability to generalise and be generalised. This entails that while multi-patient seizure detection may perform well for some patients, it may yield poor results for others, leaving individualised seizure detection more suitable for these patients.

The chapter also presented two machine learning algorithms which were used to improve the performance outcome of the multi-patient seizure detectors. We revealed that using machine learning algorithms suitable for a multi-task problem, generalisation of 'zero-training' Patient-Files could be improved.

Chapter 9
Conclusions and Future Work

This thesis has investigated the characteristics of epileptic seizure detection from EEG records as a machine learning problem, under several experimental conditions, in order to better understand the potential applications in the management of epilepsy.
In the introduction to this thesis (chapter 1) we presented four central contributions of this thesis:

The first contribution is the implementation of extensive, many-patient experiments on feature selection and machine learning in seizure prediction, showing that robust and statistically validated performance can be achieved with appropriate feature selection strategies.

The second contribution is that, by using feature selection methods and drawing from previous studies and empirical results, we produced a new set of features that led to a better performance than the previous set. The statistical significance of our findings was verified across all patients from our dataset.

The third contribution is that, epileptic seizures can be predicted up to 25 minutes prior the physiological onset with high levels of Accuracy, Sensitivity and Specificity. The advance prediction of seizures can yield higher performance than the onset detection in special cases.

The fourth contribution in this thesis is that, by using machine learning algorithms suitable for multi-source learning that are trained on the invasive EEG of multiple patients with epilepsy, the generalisation of the epileptic state for zero-training patients can improve with acceptable Sensitivity and Specificity.

This chapter presents a summary of the main contributions of this thesis in section 9.1 and points out direction of further research in section 9.2.

8.1 Main Contributions

In the first few chapters of the thesis, we saw that successful prediction of epileptic seizures from EEG records can lead to better management of seizures. We also saw how machine learning tools can be used in order to improve such systems. The experimental part of this thesis was divided into three main categories: seizure detection under reduced dimensionality; the prediction of seizures in advance of the onset and multi-patient seizure prediction. The following, highlights the main contributions made across these experiments:

- **Evaluating the effects of feature-set dimensionality across all patients:**
  We used two filter feature selection methods in a stepwise dimensionality reduction experiment on the feature-set derived from Costa et al. (Costa et al. 2008) of all patients in the Freiburg EEG Database. We were able to evaluate the effects of reducing and increasing the size of the feature-set across all patients. We discovered that a feature-set of size 4 was the lowest possible size for epileptic seizure prediction studies, without significantly compromising the performance outcome of the classifier. This is while a large number of studies in the literature have reported their results on a single feature. Our results justify the recommendation that research in this field should move towards routinely considering multiple EEG features since these will improve outcomes (see sections 6.2).

- **Identification of best EEG features:**
  We used filter feature selection tools to identify the ranking of the features according to the relevant feature selection criteria. We discovered that using good feature selection algorithms and stepwise feature-reduction, we were able to find groups of features that contributed the most to the individualised seizure classification across all patients (see section 6.2-6.3).

- **Extending the original Feature-set:**
  Drawing from the seizure prediction and dimensionality reduction experiments on single-channel and multi-channel feature-sets (sections 6.2, 6.3), we heuristically expanded the feature-set to include an additional 20 features per channel. The results revealed that performance of several subsets of the

extended feature-set, is higher than all other experimental results, including the benchmark (sections 6.4). The outcome was reported on an unseen test-set, where Accuracy, Sensitivity and Specificity were high.

- **Identifying the high performance experimental setup for dimensionality reduction**

  We implemented our stepwise dimensionality reduction on a number of settings including: single-channel, multi-channel, extended feature-set. We discovered that out of all experiments, a subset of the extended multi-channel feature-set, yielded the highest performance on the held-out test-set. The results were validated in terms of S1-Score and Accuracy (sections 6.2-6.4).

- **Prediction of seizures in advance of the onset**

  We used an advance prediction algorithm in conjunction with a suitable machine learning algorithm, in an effort to identify seizures in advance of their occurrence. The results revealed that advance prediction could be achieved with a relatively high performance, up to 25 minutes in advance of the seizure onset. The highest predictive performance was produced on a 14 dimensional subset of our extended feature-set (presented in chapter 6), determined by ReliefF feature selection method, using the Delete prediction algorithm. High values of S1-Score were obtained at $t = 1$ (96.30%) and $t = 8$ (96.13), the former being higher than the performance at seizure onset (96.18%), and where t is the minutes in advance of the seizure onset. We also identified intervals where advance prediction yielded a higher performance than the benchmark i.e. seizure onset detection. We revealed that for t minutes in advance of the seizure on-set, where $t = 5$ and 18, for all experiments, averaged over all Patient-Files, performance noticeably dropped. We also established that moments $t = 1$ and 2 constantly yielded high performance outcome across all experiments, and moments $t = 8$ and 16 were within a high-performance range for some of the best performing experiments. These findings suggest that these moments bear considerable predictive impact for a large population of patients (see section 7.2-7.5).

- **Identifying the high performance experimental setup for advance seizure prediction**

  The advance seizure prediction was conducted on a number of experimental feature-sets: single-channel, multi-channel, extended multi-channel feature-set and a subset of 14 features from the extended feature-set. The subset of extended feature-set produced the highest outcome and yielded a relatively constant performance throughout all prediction time-frames. This was followed by the two cases of 'best' EEG channel advance prediction. In general, we recommend the construction of a large, heuristically constructed feature-set, from which, by using a good feature selection method such as ReliefF, a highly discriminatory set of features can be extracted according to the characteristics of the individual patients (see section 6.4, 7.4-7.5).

- **Validation of results:**

  In all experiments, classification results were averaged over 10 runs of training and testing. The classifiers were validated on a validation set and tested on a held-out test-set. The results were reported in terms of Accuracy and S1-Score (the harmonic mean of Sensitivity and Specificity). These characteristics of our experiments support the soundness of the presented contributions made by this thesis (Chapters 6, 7 & 8).

- **Comprehensive evaluation of multi-patient seizure prediction:**

  The literature presents little evidence of extensive multi-patient seizure prediction analysis. The majority of seizure-prediction studies, has either dismissed the potential power of this tool, or has reported poor outcomes on a small set of patients. In this thesis, we conducted a full and exhaustive multi-patient seizure prediction analysis using Multi-class SVM, which has been used throughout this study. A comprehensive study entails the analysis of all possible training-set group-sizes with a reasonable number of combinations of patients in each group. These results were reported over 10 runs per classification. A total of 1102 classifiers were built for this experiment and results were reported in terms of 'zero-training' files, where patients on whom the classifier was not

trained were tested, and 'Unseen trained' files where 30% held-out data of the patients in the training-set were tested by the classifier.

The results were consistent with those in the literature: using a generally good machine learning algorithm, the classification of zero-training Patient-Files is significantly lower than the individualised seizure prediction algorithm. But in addition, we also revealed that group-size has a principal influence (albeit little) on both variations of test-set performance. Generally speaking, by adding more Patient-Files to the training-set, the generalisation of 'zero-training' files improves while the generalisation over 'unseen-trained' files decreases (see section 8.2).

- **Using suitable machine learning algorithms, multiple-patient seizure prediction can be improved**

  We implemented a number of advanced machine learning algorithms, which are more suitable for multi-source training and classification. From which, we were able to demonstrate that, by using a better machine learning algorithm (such as Deep belief Nets and Multi-Task Learning) and by handling skewness in the dataset, we are able to better generalise seizure detection for zero-training patients (see section 8.5).

- **Ability to generalise and be generalised varies amongst patient**

  By closely observing results from the multi-patient analysis, where the sizes of training groups were 2 and 20, we revealed that the ability to generalise, and the tendency to be generalised as zero-training data, varies from patient to patient. More so, these two are not correlated for any single patient; a patient whose classifier does not generalise zero-training files well, may or may not be well-generalised by other classifiers. This was further verified in the enhanced multi-patient analysis using Multi-task learning and Deep Belief Nets, where the mean performance was less than the maximum performance produced by a single-patients' generalisation (see section 8.4).

8.2 Directions for further Research

The results obtained in this thesis open directions to further research and improvement. The following lists points to improvement and future work:

8.2.1 Mining the Results from Multiple patient seizure detection

The volume of results produced from the comprehensive multi-patient seizure-prediction is significant. Visualisation of these results is particularly difficult as characteristics such as combination of files, characteristics of patients, characteristics of patients in the feature-set and four performance measures of Accuracy, Sensitivity, Specificity and S1-score, demand multi-dimensional representation which, even if done, would be very difficult to interpret. In this thesis, we have mainly looked at the two special cases where data could easily be visualised. Using association rules and other data-mining tools, useful results could potentially be extracted, which could reveal more about how the characteristics of Patient-Files come in to play with respect to multi-patient analysis. We suspect that there are correlations between characteristics of patients, whether in their profiles, or seizure attributes, which makes them more predictive in generalising some patients rather than others.

8.2.2 Accounting for patient similarity in Multi-Patient Seizure Prediction

In the multiple-patient seizure-prediction analysis, we discovered that a patient is generalised differently by other patients; in most cases, the majority of patients produce a similar generalisation outcome on the specific training-set, whilst a small number of patients may produce exceptionally high performance outcomes, compared to other patients. This suggests the inherent similarity among groups of patients. It would be significantly useful to understand what makes these patients similar. By using, and automatically detecting these similarities, we are able to build more powerful multi-patient analysis predictors, where patients of similar characteristics are used in training/test combinations. A simple example of such methods, is a weighted average of classification results, where patients that are similar to the tested patient are given a higher weight compared to other patients.

8.2.3 Expand the Deep Belief Nets:

Amongst the multi-patient seizure detection methods, Deep Belief Nets yielded the highest test-set outcome on both skewed and balanced datasets. The results were however reported for 29 variations of hyper-parameter sets. Since we have used random jumps to find these parameters, it is likely that we are yet to find the global optima. With further fine-tuning these parameters by either introducing more random jumps or using Gaussian Processes (chapter 8), we are able to yield potentially better outcomes (Bergstra et al. 2011).

8.2.4 On-line Seizure Detection:

In this thesis, we have only looked at off-line seizure detection, where training and test data are at hand and can be pre-processed and analyzed in batches. Now that optimal, experimental settings have been identified, it is useful to evaluate them in variations of on-line learning which is closer to the real-time application (Anderson 2008). This is significantly useful for the case of advance prediction, as prediction windows have merely been simulated in the work presented in this thesis.

8.2.5 Other machine learning algorithms

In most experiments presented in this thesis we used a single, generally good, machine learning algorithm, namely Multi-Class SVM, which was suitable for the seizure classification task. We have successfully identified the optimal experimental settings for improved, individualised epileptic seizure detection and prediction. It would be of high advantage to apply other suitable machine learning algorithms such as Bayesian methods on the best experimental setups, in order to evaluate how each algorithm measures up against our benchmark, and whether further improvements are obtained using other machine learning tools.

8.2.6 Further Exploration of Features

This thesis presented a new set of features that yielded a higher performance than that of the benchmark results produced by a previous set of features. These results were averaged across all patients, with low variation amongst patients in various stages of the experiment. However, some feature-sets led to variation in the performance among

patients. In addition to this, the feature-rankings were observed as highly patient-specific. By having a closer look at each patient, and its feature rankings and respective classification-performance, and by looking at recurring patterns across patients, we may find patient-specific or patient-group specific characteristics, which led to the particular ranking order of features.

In addition to studying individual and cluster performance-factors of the patients on the derived features, we can also look towards further expanding the feature-set to incorporate additional features, particularly multivariate ones which involve all channels of the patient in order to verify the effect of such features in a multi-feature experimental setup.

8.2.7 Multi-Modal Training Set

In chapter 8 of this thesis, we introduced two machine learning algorithms, which yielded a high performance for multi-patient classification of seizures. These methods, namely Multi-Task Learning and Deep Belief Nets, are forms of Transfer Learning, where information from solving one problem is stored and used for solving a different, but related problem. In the multi-patient classification problem, we regarded each patient's invasive EEG data as a specific task to the Multi-Task Learning algorithm, where the ictal and non-ictal states of the invasive EEG were a commonality between the related tasks. Transfer Learning methods are also particularly suitable for Multi-Modal training-sets, where the information about a task is obtained from different modes of data retrieval. Drawing from the success of the multi-patient analysis using Transfer Learning presented in this thesis, and other research conducted on multi-modal training using this method of learning (Yuan et al. 2012; Charuvaka & Rangwala 2012), we can formulate a new problem where a patient-specific learner is trained on multiple modes of epilepsy data, such as Invasive EEG, Scalp EEG, MRI scans etc. where all modes are related in the representation of the seizure and non-seizure characteristics of a specific patient.

8.2.8 Predicting Seizures Further in Advance

In chapter 7 of this thesis, we implemented prediction algorithms with alternating prediction window lengths, on all patients from the Freiburg EEG Database. We successfully found predictive markers up to 20 minutes in advance with high

performance measures (S1-Score 96.30%). We also used continuous 24-hour EEG recordings for two patients in order to evaluate the effect of this extended dataset on detection and prediction of epileptic seizures. However, the experiments reported results for up to 20 minutes in advance. This is while some studies in the literature have reported ictal activity of further minutes in advance (7 hours in one reported statistical approach) (Litt et al. 2001; Iasemidis et al. 2005; Chaovalitwongse et al. 2005), albeit with low performance measures. Some of these optimistic results were not reproduced in other similar studies. It would clearly be useful, to widen the predictive window to hours prior the seizure onset, and evaluate predictability using the optimal experimental settings established in the work presented in this thesis.

8.2.9 Real-Life Application

Finally, this research can be tailored towards a realised neurological pacemaker. NeuroPace (Vachtsevanos 2003) is the first FDA approved pacemaker, particularly useful for patients with drug-resistant partial or complex generalised epilepsy, where surgery is not deemed useful. It is essentially a deep brain stimulation implant, which detects a seizure and stops it by resetting the state of the brain. The product has been tested in clinical trials for the past 15 years, and was just recently approved as a medical intervention device in February 2013. By tailoring this research and similar work to a technology with practical application, we are able to further fine-tune the performance with specific real-life requirements. This could potentially lead to improved and far more accurate seizure-prevention systems.

Appendix A

	ACC	f	SP	f	SS	f	S1	f
min	83.94	2	95.29	2	68.24	2	78.48	2
max	95.46	14	98.83	14	87.02	12	92.42	12
f full	95.46	14	98.83	14	86.84	14	92.31	14
f = 2	83.94	2	95.29	2	68.24	2	78.48	2
mean	93.28		98.19		83.31		89.77	
median	94.70		98.65		85.21		91.29	
mode	83.94		95.29		68.24		78.48	
std	4.16		1.29		6.72		5.03	
range	11.53		3.54		18.78		13.94	

Table A.1 Summary of important data statistics from the stepwise **dimensionality reduction** on 18 **single-channel** patients using **mRMR**.

Feature Type	*Features*	*mRMR (Avg. Rank)*	*ReliefF (Avg. Rank)*
Signal Energy	Accumulated energy	**1.00**	**2.95**
	Energy level	13.19	8.48
	Energy variation (STE)	13.43	8.62
	Energy variation (LTE)	11.86	**4.90**
Wavelet Transform	Energy STE 1	**5.10**	**8.00**
	Energy STE 2	**6.81**	**6.76**
	Energy STE 3	**7.48**	**7.67**
	Energy STE 4	9.90	**7.86**
	Energy LTE 1	**4.57**	**5.57**
	Energy LTE 2	**5.81**	**7.24**
	Energy LTE 3	**7.19**	**7.33**
	Energy LTE 4	9.38	**6.24**
Nonlinear system dynamics	Correlation dimension	**6.76**	10.95
	Max Lyapunov Exponent	**2.52**	12.43

Table A.2 Rankings of the 14 features averaged over all single-channel patients from feature selection methods mRMR and ReliefF. The table is organised based on the category of the features. The rankings in bold are those of values ≤8 which will further be used in the feature extension experiments.

	ACC	f	SP	f	SS	f	S1	f
min	83.50	2	94.70	2	69.64	2	79.63	2
max	97.73	84	99.35	84	92.17	78	95.60	78
f full	97.73	84	99.35	84	92.04	84	95.53	84
f = 2	83.50	2	94.70	2	69.64	2	79.63	2
mean	96.12		98.95		88.02		92.65	
median	96.73		99.11		89.29		93.83	
mode	83.50		94.70		69.64		79.63	
std	2.32		0.72		4.38		3.44	
range	14.23		4.65		22.54		15.96	

Table A.3 Summary of important data statistics from the stepwise **dimensionality reduction** on 18 **Multi-Channel** patients using **mRMR**.

Feature No.	Feature Name	Channel	Avg. Rank	Standard deviation
5	Lyapunov Exp.	1	3.81	±8.09
19	Lyapunov Exp.	2	13.38	±11.19
33	Lyapunov Exp.	3	17.14	±12.17
11	LTE 1	1	18.43	±14.61
47	Lyapunov Exp.	4	20.52	±12.97
25	LTE 1	2	23.05	±13.57
39	LTE 1	3	23.67	±14.20
61	Lyapunov Exp.	5	23.86	±14.22
12	LTE 2	1	25.24	±20.25
35	STE 1	3	26.19	±14.00
63	STE 1	5	27.05	±14.74
75	Lyapunov Exp.	6	27.81	±14.83
40	LTE 2	3	28.10	±18.74
67	LTE 1	5	28.71	±14.48

Table A.4 Properties of 14 features from the 84 Multi-Channel feature-set that hold the highest rankings according to mRMR feature selection method. The rankings were averaged over all 21 patients from the Freiburg EEG database. Listed are the feature number in the dataset, the feature name irrespective of channel, the channel the feature was obtained from, the average ranking and the standard deviation of the average rank.

	Best mRMR Channel	Feature Contribution	Best ReliefF Channel	Features Contribution
Patient 1	4	28.57%	**1**	35.71%
Patient 2	**1**	35.71%	**3**	50.00%
Patient 3	**1**	28.57%	4	35.71%
Patient 4	6	35.71%	**1**	28.57%

Patient 5	**1**	42.86%	**1**	35.71%
Patient 6	6	35.71%	**2**	35.71%
Patient 7	**2**	35.71%	**3**	35.71%
Patient 8	**1**	35.71%	**1**	21.43%
Patient 9	**1**	21.43%	**1**	28.57%
Patient 10	**1**	21.43%	**2**	21.43%
Patient 11	**3**	35.71%	**2**	50.00%
Patient 12	**1**	35.71%	**2**	42.86%
Patient 13	**3**	28.57%	6	28.57%
Patient 14	**1**	28.57%	**3**	28.57%
Patient 15	5	35.71%	5	50.00%
Patient 16	5	35.71%	**2**	42.86%
Patient 17	**2**	28.57%	**1**	21.43%
Patient 18	4	35.71%	4	35.71%
Patient 19	**1**	28.57%	**1**	42.86%
Patient 20	**1**	28.57%	**2**	28.57%
Patient 21	**1**	21.43%	**1**	50.00%

Table A.5 The Best EEG Channel for each of the 21 Patient-Files from the Freiburg EEG database – Best Channels are identified according to mRMR and ReliefF Feature selection methods. The selection was based on the highest amount of contributions of each channel to the top 14 feature-ranks, listed in respective feature-contribution columns. The underlined elements in the channel columns indicate a match between the channels deemed 'Best' by both feature selection methods. The bold listings indicate focal channels.

	ACC	**f**	**SP**	**f**	**SS**	**f**	**S1**	**f**
min	86.58	2	95.27	2	76.94	2	84.34	2
max	98.25	44	99.64	116	94.62	42	96.93	42
f full	94.30	204	99.62	204	82.07	204	89.05	216
f = 2	86.58	2	95.27	2	76.94	2	84.34	2
mean	97.17		99.51		91.13		94.86	
median	97.89		99.58		93.09		96.14	
mode	86.58		95.27		76.94		84.34	
std	1.52		0.44		3.76		2.40	
range	11.67		4.38		17.69		12.59	

Table A.6 Summary of important data statistics from the stepwise **dimensionality reduction** on the 18 patients with extended **204-dimentional** feature-set using **mRMR**.

Feature No.	Feature Name	Channel	Avg. Rank	Standard deviation
5	Lyapunov Exp.	1	27.19	±41.12
30	SBP LTE Delta	1	58.48	±47.25
96	SBP STE Beta	3	59.19	±40.48
62	SBP STE Beta	2	59.38	±34.29
28	SBP STE Beta	1	60.05	±35.90
31	SBP LTE Theta	1	60.62	±55.45
49	SEF STE	2	61.10	±46.32
168	SBP LTE Alpha	5	61.10	±50.30
32	SBP LTE Alpha	1	62.57	±49.10
26	SBP STET beta	1	62.86	±36.59
35	Accum. Energy	2	63.33	±78.93
94	SBP STET beta	3	64.29	±29.91
169	SBP LTE Beta	5	64.29	±43.46
27	SBP STE Alpha	1	64.38	±32.15

Table A.7 Properties of 14 features from the 204 extended feature-set that hold the highest rankings according to mRMR feature selection method. The rankings were averaged across all 21 patients from the Freiburg EEG database. Listed are the feature number, the feature name irrespective of channel, the channel the feature was obtained from, the average ranking and the standard deviation of the rank.

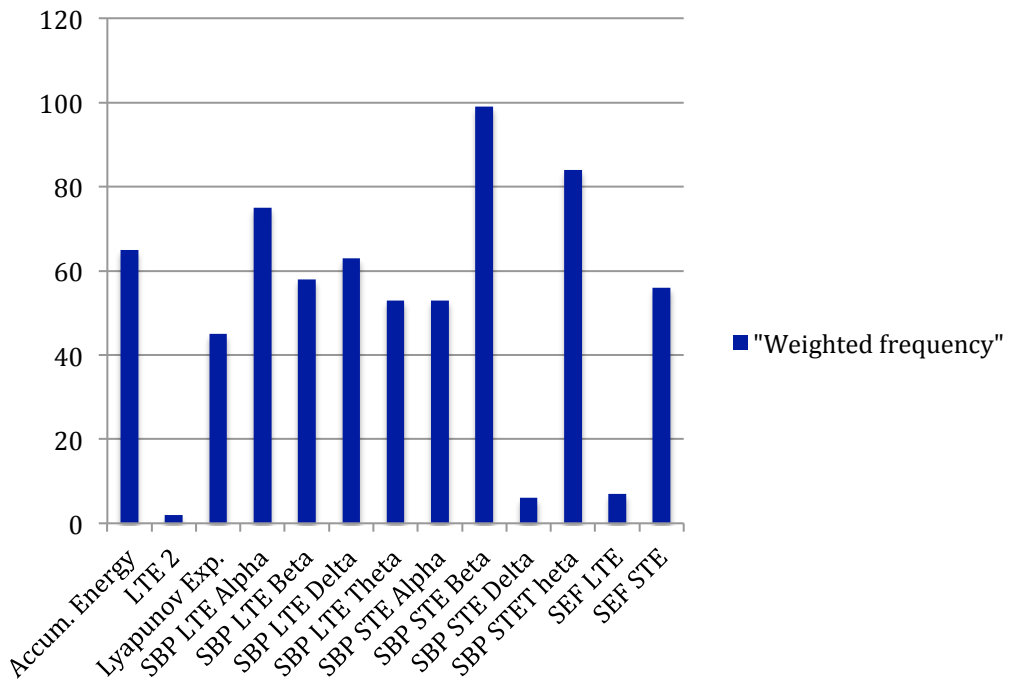
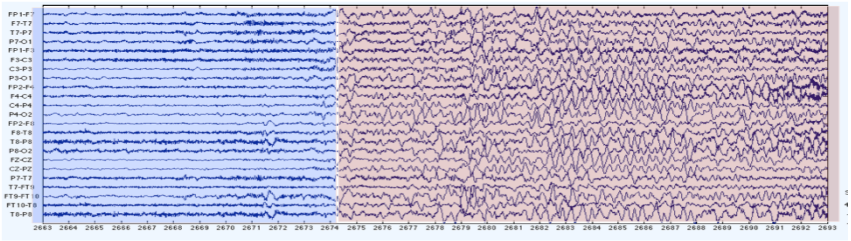


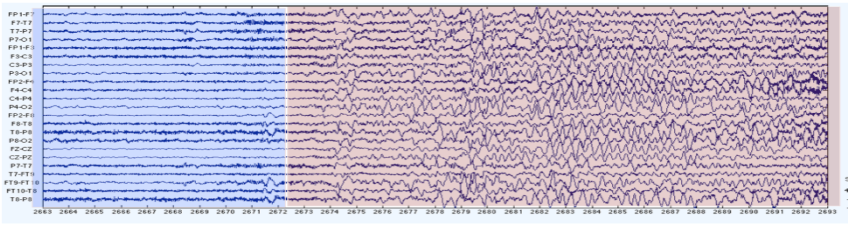
Figure A.8 Weighted Frequency of top 84 **mRMR** features – The weighted frequency of features within the top 84 range of the ranking table produced by ReliefF feature selection were calculated based on their ranking and frequency independent of the channel they came from

Appendix B

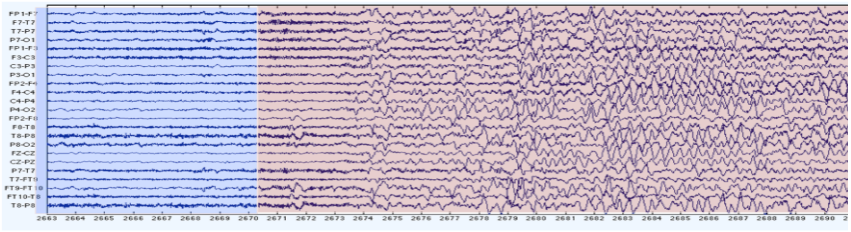
$t = 0$
$D = m \times n$
$Ictal = i \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = x \times n$



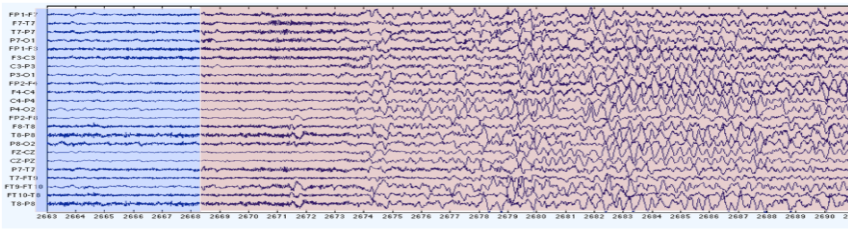
$t = 1$
$D = m \times n$
$Ictal = (i + t) \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



$t = 2$
$D = m \times n$
$Ictal = (i + t) \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



$t = 3$
$D = m \times n$
$Ictal = (i + t) \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$



$t = 4$
$D = m \times n$
$Ictal = (i + t) \times n$
$pre\_ictal = p \times n$
$Inter\_ictal = (x - t) \times n$

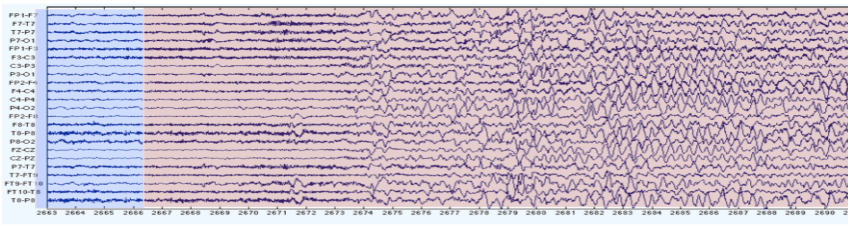


Figure B.1 Prediction Simulation by Relabeling Data – The top image displays the default EEG status where the ictal data is preceded by Pre-ictal data. In this illustration we display up to 4 steps of data manipulation. At each step *t,* the ictal window is pushed back for a fixed interval by renaming the preceding data as ictal. For each stage of the Rename process, updates to time-step t, ictal length, pre-ictal length, inter-ictal length and overall dimensionality of the dataset is indicated.

	ACC	t	SP	t	SS	t	S1	t
min	94.90	1	98.62	1	71.11	4	82.31	4
max	95.78	0	99.25	15	87.88	0	92.94	0
t = 0	95.78	0	98.84	0	87.88	0	92.94	0
t = 20	95.56	20	98.90	20	74.92	20	85.04	20
mean	95.36		98.99		75.23		85.15	
median	95.44		99.00		74.84		85.04	
mode	94.90		98.62		71.11		82.31	
std	0.29		0.15		3.70		2.35	
range	0.88		0.63		16.77		10.63	

Table B.2 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** performed on 18 **single-channel** patients.

	ACC	t	SP	t	SS	t	S1	t
min	96.78	12	99.00	1	67.75	19	80.10	19
max	97.74	0	99.37	0	91.96	0	95.49	0
t = 0	97.74	0	99.37	0	91.96	0	95.49	0
t = 20	97.00	20	99.16	20	73.17	20	83.76	20
mean	97.00		99.18		73.40		83.87	
median	96.98		99.18		71.72		82.78	
mode	96.78		99.00		67.75		80.10	
std	0.19		0.07		5.49		3.53	
range	0.97		0.37		24.21		15.39	

Table B.3 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** performed on 18 **Multi-Channel** patients.

	ACC	t	SP	t	SS	t	S1	t
min	95.05	6	98.66	1	71.65	4	82.66	4
max	95.93	0	99.17	15	88.45	0	93.28	0
t = 0	95.93	0	98.84	0	88.45	0	93.28	0
t = 20	95.62	20	99.00	20	74.14	20	84.29	20
mean	95.40		98.94		75.45		85.25	
median	95.30		98.96		74.54		84.78	
mode	95.05		98.66		71.65		82.66	
std	0.25		0.12		3.59		2.25	
range	0.88		0.51		16.81		10.62	

Table B.4 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** performed on 18 patients with feature-set comprising the **Best Channel** determined by **mRMR** feature selection method.

	ACC	t	SP	t	SS	t	S1	t
min	93.37	6	99.37	1	58.21	18	71.69	18
max	94.35	0	99.66	15	81.98	0	89.01	0
t = 0	94.35	0	99.61	0	81.98	0	89.01	0
t = 20	94.12	20	99.57	20	63.41	20	76.40	20
mean	93.78		99.58		65.74		77.78	
median	93.70		99.59		64.91		77.41	
mode	93.37		99.37		58.21		71.69	
std	0.25		0.07		4.84		3.51	
range	0.98		0.29		23.77		17.32	

Table B.5 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** on 18 **Multi-Channel Extended Feature-Set** patients.

	ACC	t	SP	t	SS	t	S1	t
min	97.14	10	99.25	20	81.56	18	89.20	18
max	97.73	0	99.45	4	93.82	0	96.43	0
t = 0	97.73	0	99.37	0	93.82	0	96.43	0
t = 20	97.40	20	99.25	20	85.66	20	91.65	20
mean	97.33		99.34		85.41		91.52	
median	97.28		99.32		84.75		91.30	
mode	97.14		99.25		81.56		89.20	
std	0.14		0.05		2.43		1.43	
range	0.59		0.20		12.26		7.23	

Table B.6 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** on 18 patients – The dataset of each patient comprises a **subset of 14 Multi-Channel Extended Feature-Set** determined by **mRMR** feature selection method.

	ACC	T	SP	t	SS	t	S1	t
min	97.47	19	99.27	20	89.24	9	93.78	9
max	97.94	8	99.53	7	93.88	0	96.47	0
t = 0	97.74	0	99.37	0	93.88	0	96.47	0
t = 20	97.52	20	99.27	20	91.15	20	94.77	20
mean	97.74		99.43		90.97		94.84	
median	97.74		99.43		90.53		94.64	
mode	97.47		99.27		89.24		93.78	
std	0.13		0.08		1.18		0.68	
range	0.48		0.26		4.64		2.69	

Table B.7 Summary of important data statistics from the stepwise advance seizure prediction by **Delete** on 18 patients – The dataset of each patient comprises a **subset of 14 Multi-Channel Extended Feature-Set** determined by **mRMR** feature selection method.

	ACC	t	SP	t	SS	t	S1	t
min	97.22	11	99.37	1	82.75	18	89.70	18
max	97.57	16	99.58	16	93.16	0	96.06	0
t = 0	97.53	0	99.40	0	93.16	0	96.06	0
t = 20	97.30	20	99.45	20	83.23	20	90.13	20
mean	97.39		99.47		85.42		91.51	
median	97.40		99.46		84.79		91.02	
mode	97.22		99.37		82.75		89.70	
std	0.09		0.06		2.46		1.50	
range	0.35		0.21		10.41		6.37	

Table B.8 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** on 18 patients – The dataset of each patient comprises a **subset of 14 Multi-Channel Extended Feature-Set** determined by **ReliefF** feature selection method.

Appendix C

	ACC	t	SP	t	SS	t	S1	t
min	95.84	20	98.84	1	80.29	3	88.39	3
max	96.36	13	99.30	15	88.93	0	93.56	0
t = 0	96.07	0	98.87	0	88.93	0	93.56	0
t = 20	95.84	20	98.98	20	83.64	20	90.49	20
mean	96.12		99.07		83.86		90.62	
median	96.16		99.10		83.31		90.26	
mode	95.84		98.84		80.29		88.39	
std	0.16		0.12		1.86		1.14	
range	0.52		0.47		8.63		5.17	

Table C.1 Summary of important data statistics from the stepwise advance seizure prediction by **Delete** performed on 18 patients with feature-set comprising the **Best Channel** determined by **mRMR** feature selection method.

	ACC	t	SP	t	SS	t	S1	t
min	95.05	6	98.66	1	71.65	4	82.66	4
max	95.93	0	99.17	15	88.45	0	93.28	0
t = 0	95.93	0	98.84	0	88.45	0	93.28	0
t = 20	95.62	20	99.00	20	74.14	20	84.29	20
mean	95.40		98.94		75.45		85.25	
median	95.30		98.96		74.54		84.78	
mode	95.05		98.66		71.65		82.66	
std	0.25		0.12		3.59		2.25	
range	0.88		0.51		16.81		10.62	

Table C.2 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** performed on 18 patients with feature-set comprising the **Best Channel** determined by **mRMR** feature selection method.

	ACC	t	SP	t	SS	t	S1	t
min	96.16	20	98.99	1	81.44	3	89.23	3
max	96.62	7	99.27	14	89.03	0	93.69	0
t = 0	96.38	0	99.01	0	89.03	0	93.69	0
t = 20	96.16	20	99.10	20	83.52	20	90.49	20
mean	96.42		99.14		84.89		91.32	
median	96.39		99.13		85.11		91.41	
mode	96.16		98.99		81.44		89.23	
std	0.13		0.09		1.86		1.12	
range	0.46		0.28		7.59		4.47	

Table C.3 Summary of important data statistics from the stepwise advance seizure prediction by **Delete** performed on 18 patients with feature-set comprising the **Best Channel** determined by **ReliefF** feature selection method.

	ACC	t	SP	t	SS	t	S1	t
min	95.40	4	98.77	2	71.41	4	82.67	4
max	96.23	0	99.21	15	88.79	0	93.52	0
t = 0	96.23	0	98.91	0	88.79	0	93.52	0
t = 20	95.82	20	99.01	20	73.20	20	83.88	20
mean	95.78		99.01		75.83		85.62	
median	95.82		99.03		75.39		85.41	
mode	95.40		98.77		71.41		82.67	
std	0.22		0.12		3.75		2.36	
range	0.83		0.44		17.38		10.85	

Table C.4 Summary of important data statistics from the stepwise advance seizure prediction by **Rename** performed on 18 patients with feature-set comprising the **Best Channel** determined by **ReliefF** feature selection method.

Bibliography

Abou-Khalil, B. & Misulis, K.E., 2006. *Atlas of EEG and seizure semiology*, Butterworth-Heinemann.

Adelson, P.D. et al., 1999. Noninvasive continuous monitoring of cerebral oxygenation periictally using near-infrared spectroscopy: a preliminary report. *Epilepsia*, 40(11), pp.1484–1489.

Aldenkamp, A.P., Krom, M.D. & Reijs, R., 2003. Newer antiepileptic drugs and cognitive issues. *Epilepsia*, 44(s1), pp.21-24.

Anderson, T., 2008. *The Theory and Practice of Online Learning*, Athabasca University Press.

Ando, R.K. & Zhang, T., 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6, pp.1817–1853.

Arel, I., Rose, D.C. & Karnowski, T.P., Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]. *IEEE Computational Intelligence Magazine*, 5(4), pp.13–18.

Arnhold, J., Grassberger, P. & Lehnertz, K., 1999. A robust method for detecting interdependences: application to intracranially recorded EEG. Physica D: *Nonlinear Phenomena*, 134(4), pp.419-430.

Bai-Lin, H., 1989. *Elementary symbolic dynamics and chaos in dissipative systems*.

Barber, D., 2012. *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

Baumgartner, C. et al., 2005. Long-term Prognosis of Analgesic Withdrawal in Patients with Drug-Induced Headaches. *Headache: The Journal of Head and Face Pain*, 29(8), pp.510–514.

Baumgartner, C., Serles, W. & Leutmezer, F., 1998. Preictal SPECT in temporal lobe epilepsy: regional cerebral blood flow is increased prior to electroencephalography-seizure onset. *Journal of nuclear medicine*, 39(6), pp.978-82.

Baxter, J., 2000. A model of inductive bias learning. *Journal of Artificial Intelligence Research,* 12(1).

Bengio, Y., 2009. Learning Deep Architectures for AI, *Foundations and Trends® in Machine Learning*, 2(1), pp.1–127.

Bengio, Y., Courville, A. & Vincent, P., 2013. Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence*. 35(8), pp. 1798 - 1828

Berg, A.T. et al., 2001. Early development of intractable epilepsy in children: a prospective study. *Neurology*, 56(11), pp.1445–1452.

Berg, A.T. et al., 2010. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. *Epilepsia*, 51(4), pp.676–685.

Bergstra, J. et al., 2011. Algorithms for hyper-parameter optimization. *25th Annual Conference on Neural Information Processing System*.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*, Springer.

Blankertz, B. et al., 2004. The BCI Competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *IEEE Transactions on Biomedical Engineering*, 51(6), pp.1044–1051.

Caruana, R., 1997. Multitask Learning. *Machine learning*, 28(1), pp.41–75.

Cascino, G. & Sirven, J., 2011. *Adult Epilepsy*, John Wiley & Sons.

Cs.toronto.edu. 2013. *Home Page of Geoffrey Hinton*. [online] Available at: http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html [Accessed: 14 Aug 2013].

Csie.ntu.edu.tw. 2013. *LIBSVM -- A Library for Support Vector Machines*. [online] Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [Accessed: 14 Aug 2013].

Celka, P. & Colditz, P., 2002. A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison. Biomedical Engineering, *IEEE Transactions on Biomedical Engineering*, 49(5), pp.455–462.

Chang, C.-C. & Lin, C.-J., 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), p.27.

Chaovalitwongse, W. et al., 2005. Performance of a seizure warning algorithm based on the dynamics of intracranial EEG. *Epilepsy Research*, 64(3), pp.93–113.

Chapelle, O., 2007. Training a support vector machine in the primal. *Neural computation*, 19(5), pp.1155–1178.

Charuvaka, A. & Rangwala, H., 2012. *Multi-task Learning for Classifying Proteins using Dual Hierarchies*. pp.834–839.

Chawla, N.V. et al., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research,* 16(1).

Chávez, M., Le Van Quyen, M. & Navarro, V., 2003. Spatio-temporal dynamics prior to neocortical seizures: amplitude versus phase couplings. *IEEE Transactions on Biomedical Engineering*, 50(5), pp. 571- 583

Chen, J. et al., 2009. *A convex formulation for learning shared structures from multiple tasks*. In the 26th Annual International Conference. New York, New York, USA: ACM Press, pp.1–8.

Ciresan, D., Meier, U. & Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. In Computer Vision and Pattern Recognition (CVPR), *2012 IEEE Conference on*. pp.3642–3649.

Cockerell, O.C. et al., 1997. Prognosis of epilepsy: a review and further analysis of the first nine years of the British National General Practice Study of Epilepsy, a prospective population-based study. *Epilepsia*, 38(1), pp.31–46.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine learning,* 20(3), pp. 273- 297.

Costa, R.P. et al., 2008. Epileptic seizure classification using neural networks with 14 features. *Knowledge-Based Intelligent Information and Engineering Systems*, 5178, pp.281–288.

Cs.cmu.edu. 2002. *SMOTE: Synthetic Minority Over-sampling Technique*. [online] Available at: http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html [Accessed: 14 Aug 2013].

Da Silva, F.L. et al., 2003. Epilepsies as dynamical diseases of brain systems: basic models of the transition between normal and epileptic activity. *Epilepsia*, 44(s12), pp.72-83.

Daubechies, I. & Sweldens, W., 1998. Factoring wavelet transforms into lifting steps. *Journal of Fourier analysis and applications*. 4(3), pp.247-269

Dayan, P. & Abbott, L.F., 2005. *Theoretical Neuroscience*, MIT Press (MA).

De Clercq, W., Lemmerling, P., Van Huffel, S. & Van Paesschen, W., 2003a. Anticipation of epileptic seizures from standard EEG recordings. The Lancet, 361(9361), pp.970–author reply 970–1.

De Clercq, W., Lemmerling, P., Van Huffel, S. & Van Paesschen, W., 2003b. Anticipation of epileptic seizures from standard EEG recordings. *The Lancet*, 361(9361), pp.971–author reply 971.

Ding, C. & Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(2), pp.185–205.

Duda, R.O., Hart, P.E. & Stork, D.G., 2012. *Pattern Classification,* John Wiley & Sons.

Elger, C.E. & Lehnertz, K., 1998. Seizure prediction by non-linear time series analysis of brain electrical activity. *The European journal of neuroscience*, 10(2), pp.786–789.

Elger, C.E. & Schmidt, D., 2008. Modern management of epilepsy: A practical approach. *Epilepsy & Behavior*, 12(4), pp.501–539.

Epilepsiae. n.d.. *EPILAB Software - Epilepsiae*. [online] Available at: http://www.epilepsiae.eu/project_outputs/epilab_software [Accessed: 14 Aug 2013].

Epilepsyresearch.org.uk. 2013. *Our Research Portfolio | Epilepsy Research UK*. [online] Available at: http://www.epilepsyresearch.org.uk/research/research-portfolio-2/ [Accessed: 15 Aug 2013].

Epilepsy.com. 2010. *Neuropace | epilepsy.com*. [online] Available at: http://www.epilepsy.com/node/971263 [Accessed: 14 Aug 2013].

Epilepsy.uni-freiburg.de. 2007. *EEG Database — Seizure Prediction Project Freiburg*. [online] Available at: https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database [Accessed: 14 Aug 2013].

Fabiani, G.E. et al., 2004. Conversion of EEG activity into cursor movement by a brain-computer interface (BCI). *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 12(3), pp.331–338.

Faul, S. et al., 2005. An evaluation of automated neonatal seizure detection methods. *Clinical Neurophysiology*, 116(7), pp.1533–1541.

Fazli, S. et al., 2009. Subject-independent mental state classification in single trials. *Neural Networks*, 22(9), pp.1305–1312.

Federico, P. et al., 2005. Functional MRI of the pre-ictal state. *Brain*, 128(8), pp.1811-1817.

French, J.A., 2007. Refractory Epilepsy: Clinical Overview. *Epilepsia*, 48(s1), pp.3–7.

Gigola, S. et al., 2004. Prediction of epileptic seizures using accumulated energy in a multiresolution framework. *Journal of Neuroscience Methods*, 138(1-2), pp.107–111.

Grassberger, P. & Procaccia, I., 1983. Characterization of strange attractors. *Physical review letters*, 50(5), pp.346-349.

Grassberger, P. & Schreiber, T., 1991. Nonlinear time sequence analysis . *International Journal of Bifurcation and Chaos*, 1(3).

Guyon, I. & Elisseeff, A., 2006. *An introduction to feature extraction*. Springer, pp.1–25.

Hand, D., Mannila, H. and Smyth, P. 2001. *Principles of data mining*. Cambridge, Mass.: MIT Press.

Harrison, M. et al., 2005a. Correlation dimension and integral do not predict epileptic seizures. *Chaos*, 15 (3).

Harrison, M.A.F., Frei, M.G. & Osorio, I., 2005b. Accumulated energy revisited. *Clinical Neurophysiology*, 116(3), pp.527–531.

Hastie, T., Tibshirani, R. & Friedman, J.H., 2009. *The Elements of Statistical Learning*, Springer.

Hinton, G., 2010. A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1).

Hjorth, B., 1970. EEG analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3), pp.306-310.

Hsu, C.W., Chang, C.C. & Lin, C.J., 2003. *A practical guide to support vector classification*. University of National Taiwan.

Iasemidis, L.D. et al., 2005. Long-term prospective on-line real-time seizure prediction. *Clinical Neurophysiology*, 116(3), pp.532–544.

Iasemidis, L.D. et al., 1990. Phase space topography and the Lyapunov exponent of electrocorticograms in partial seizures. *Brain topography*, 2(3), pp.187–201.

Kalitzin, S. et al., 2002. Enhancement of phase clustering in the EEG/MEG gamma frequency band anticipates transitions to paroxysmal epileptiform activity in epileptic patients with known visual Sensitivity. *IEEE Transactions on Biomedical Engineering*, 49(11), pp.1279–1286.

Kantz, H., 1994. A robust method to estimate the maximal Lyapunov exponent of a time series. *Physics letters A*, 185(1), pp.77-87.

Kantz, H. & Schreiber, T., 2004. *Nonlinear Time Series Analysis*. Cambridge University Press.

Kaplan, D. & Glass, L., 1992. Direct test for determinism in a time series. *Physical review letters*, 68(4), pp.427–430.

Kerem, D.H. & Geva, A.B., 2005. Forecasting epilepsy from the heart rate signal. *Medical and Biological Engineering and Computing*, 43(2), pp.230–239.

Khan, Y.U. & Gotman, J., 2003. Wavelet based automatic seizure detection in intracerebral electroencephalogram. *Clinical Neurophysiology*, 114(5), pp.898–908.

Knerr, S., Personnaz, L. & Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network. *Neurocomputing*, 68, pp.41-50.

Kononenko, I., 1994. Estimating attributes: analysis and extensions of RELIEF. In ECML-94: *Proceedings of the European conference on machine learning on Machine Learning*. Springer-Verlag New York, Inc, 784, pp.171-182.

Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1), pp.89–109.

Kotsiantis, S.B., Zaharakis, I.D. & Pintelas, P.E., 2007. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), pp.159–190.

Kreßel, U., 1999. Pairwise classification and support vector machines. *Advances in kernel methods*, pp.255-268.

Kwan, P. & Brodie, M.J., 2000. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5), pp.314–319.

Lai, Y.-C. et al., 2004. Controlled test for predictive power of Lyapunov exponents: their inability to predict epileptic seizures. *Chaos (Woodbury, N.Y.)*, 14(3), pp.630–642.

Le Van Quyen, M. et al., 1999. Anticipating epileptic seizures in real time by a non-linear analysis of similarity between EEG recordings. *Neuroreport*, 10(10), pp.2149–2155.

Le Van Quyen, M. et al., 2001. Characterizing neurodynamic changes before seizures. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 18(3), pp.191–208.

Lecun, Y. et al., 1998. *Gradient-based learning applied to document recognition*. In Proceedings of the IEEE. pp.2278–2324.

Lehnertz, K. & Litt, B., 2005. The First International Collaborative Workshop on Seizure Prediction: summary and data description. *Clinical Neurophysiology*, 116(3), pp.493–505.

Litt, B. et al., 2001. Epileptic seizures may begin hours in advance of clinical onset: a report of five patients. *Neuron*, 30(1), pp.51–64.

Liu, A. et al., 1992. Detection of neonatal seizures through computerized EEG analysis. *Electroencephalography and clinical neurophysiology*, 82(1), pp.30–37.

Martinerie, J. et al., 1998. Epileptic seizures can be anticipated by non-linear analysis. *Nature medicine*, 4(10), pp.1173–1176.

Mathworks.co.uk. 1994. *MATLAB - The Language of Technical Computing - MathWorks United Kingdom*. [online] Available at: http://www.mathworks.co.uk/products/matlab/ [Accessed: 14 Aug 2013].

Mirowski, P. et al., 2009. Classification of patterns of EEG synchronization for seizure prediction. *Clinical Neurophysiology*, 120(11), pp.1927–1940.

Moghim, N. & Corne, D., 2011. *Evaluating bio-inspired approaches for advance prediction of epileptic seizures*. In 2011 Third World Congress on Nature and Biologically Inspired Computing (NaBIC). IEEE, pp.540–545.

Moore, J.H. & White, B.C., 2007. *Tuning ReliefF for genome-wide genetic analysis*. Springer.

Mormann, F. et al., 2000. Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D: Nonlinear Phenomena*, 144 (3–4), pp.358-369.

Mormann, F. et al., 2005. On the predictability of epileptic seizures. *Clinical Neurophysiology*, 116(3), pp.569–587.

Mormann, F. et al., 2007. Seizure prediction: the long and winding road. *Brain*, 130(2), pp.314–333.

Mormann, F., Elger, C.E. & Lehnertz, K., 2006. Seizure anticipation: from algorithms to clinical practice. *Current Opinion in Neurology*, 19(2), pp.187–193.

Morrell, M., 2006. Brain stimulation for epilepsy: can scheduled or responsive neurostimulation stop seizures? *Current Opinion in Neurology*, 19(2), pp.164-168.

Niedermeyer, E. & da Silva, F.H.L., 2005. *Electroencephalography, Lippincott*. Williams & Wilkins.

Noachtar, S. & Rémi, J., 2009. The role of EEG in epilepsy: a critical review. *Epilepsy & behavior* , 15(1), pp.22–33.

Ott, E., 2002. *Chaos in Dynamical Systems*, Cambridge University Press.

Park, Y. et al., 2011. Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. *Epilepsia*, 52(10), pp.1761–1770.

Peng, H., Long, F. & Ding, C., 2005a. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, *IEEE Transactions on Biomedical Engineering*, 27(8), pp.1226–1238.

Peng, H., Long, F. & Ding, C., 2005b. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, *IEEE Transactions on Biomedical Engineering*, 27(8), pp.1226–1238.

People.kyb.tuebingen.mpg.de. n.d.. *Spider*. [online] Available at: http://people.kyb.tuebingen.mpg.de/spider/ [Accessed: 14 Aug 2013].

Perucca, E. et al., 1998. Antiepileptic drugs as a cause of worsening seizures. *Epilepsia*, 39(1), pp.5–17.

Prandoni, P. & Vetterli, M., 2008. *Signal Processing for Communications,* CRC Press.

Public.asu.edu. 2012. *MALSAR: Multi-Task Learning via Structural Regularization*. [online] Available at: http://www.public.asu.edu/~jye02/Software/MALSAR/ [Accessed: 14 Aug 2013].

Quyen, M. et al., 2003. Toward a neurodynamical understanding of ictogenesis. *Epilepsia*, 44 pp.30-43

Reynolds, E.H., Elwes, R. & Shorvon, S.D., 1983. Why does epilepsy become intractable?: prevention of chronic epilepsy. *The Lancet*, 322(8356), pp.952–954.

Rieke, C. et al., 2002. Measuring nonstationarity by analyzing the loss of recurrence in dynamical systems. *Physical review letters*, 88(24), p.244102.

Rogowski, Z., Gath, I. & Bental, E., 1981. On the prediction of epileptic seizures. *Biological cybernetics*, 42(1), pp.9-15.

Rosenstein, M.T., Collins, J.J. & De Luca, C.J., 1993. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D: *Nonlinear Phenomena,* 65(1-2), pp.117-1134.

Saeys, Y., Inza, I. & Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), pp.2507–2517.

Salant, Y., Gath, I. & Henriksen, O., 1998. Prediction of epileptic seizures from two-channel EEG . *Medical and Biological Engineering and Computing*, 36 (5), pp.549-556.

Santaniello, S. et al., 2011. Quickest detection of drug-resistant seizures: An optimal control approach. *Epilepsy & Behavior*, 22(S1), pp.S49–S60.

Schiller, Y., 2009. Seizure relapse and development of drug resistance following long-term seizure remission. *Archives of neurology*, 66(10), p.1233.

Schuster, H.G. & Just, W., 2006. *Deterministic Chaos*, John Wiley & Sons.

Sccn.ucsd.edu. 2011. *EEGLAB - Open Source Matlab Toolbox for Electrophysiological Research*. [online] Available at: http://sccn.ucsd.edu/eeglab/ [Accessed: 14 Aug 2013].

Shoeb, A. et al., 2004. Patient-specific seizure onset detection. *Epilepsy & Behavior*, 5(4), pp.483–498.

Siegel, A., Grady, C.L. & Mirsky, A.F., 1982. Prediction of spike-wave bursts in absence epilepsy by EEG power-spectrum signals. *Epilepsia*, 23(1), pp.47–60.

Stanski, D.R. et al., 1984. Pharmacodynamic modeling of thiopental anesthesia. *Journal of pharmacokinetics and biopharmaceutics*, 12(2), pp.223–240.

Stein, A.G. et al., 2000. An automated drug delivery system for focal epilepsy. *Epilepsy Research*, 39(2), pp.103–114.

Suffczynski, P. et al., 2006. Dynamics of epileptic phenomena determined from statistics of ictal transitions. Biomedical Engineering, *IEEE Transactions on Biomedical Engineering*, 53(3), pp.524–532.

Teixeira, C.A. et al., 2011. EPILAB: A software package for studies on the prediction of epileptic seizures. *Journal of Neuroscience Methods*, 200(2), pp.257–271.

Theodore, W.H. & Fisher, R.S., 2004. Brain stimulation for epilepsy. *The Lancet Neurology*, 3(2), pp.111-118.

Vachtsevanos, G.J., 2003. *Neuropace, Inc. Research Agreement*, Georgia Institute of Technology.

Viglione, S.S. & Walsh, G.O., 1975. Proceedings: Epileptic seizure prediction. *Electroencephalography and clinical neurophysiology*, 39(4), pp.435–436.

Wendling, F. et al., 2003. Epileptic fast intracerebral EEG activity: evidence for spatial decorrelation at seizure onset. *Brain*, 126(6), pp.1449-1459.

Williamson, J.R. et al., 2012. Seizure prediction using EEG spatiotemporal correlation structure. *Epilepsy & Behavior*, 25(2), pp.230–238.

Yuan, L. et al., 2012. *Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data*. The Eighteenth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, pp.1149–1157.

Zhou, J., Chen, J. & Ye, J., 2012. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State Univ.