Use of Support Vector Machines to Automatically Detect Epileptic Activity in EEG Data

Miguel Fernandes

A dissertation submitted to the Faculty of Engineering, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Masters of Science in Engineering.

Johannesburg, September 2011

Abstract

This dissertation evaluates the effectiveness of using Support Vector Machines (SVM) to identify Inter-ictal epileptic activity in Electroencephalogram (EEG) data. There are existing systems that already do this but identifying the best solution requires comparative studies. A sample of data was randomly selected from 20 patients. A number of features were extracted and a PBIL algorithm was then used to identify the set of features that provided the best separability within the dataset. These were found to be related to the chaoticity, frequency and variability of the signal. The features were then used to train an SVM model. This method resulted in 88.7% accuracy but left 31.1% of inputs unclassifiable. This performance is comparable with existing solutions.

The strengths of the designed system were computational efficiency and system accuracy. The limitations were that the high degree of Artifacts masked indicative patterns and therefore decreased the classification accuracy in some cases. This system could be used as a screening tool for detecting patients that possibly have epilepsy. The system cannot be used to rule out epilepsy in a patient due to the high rate of unclassifiable data.

Some areas were identified where future work could be done, namely a spike detection method, multi-channel analysis, and improved Artifact extraction classification. These areas could improve the general performance of the classification system in terms of more clearly separating epileptic from non-epileptic activity.

Use of Support Vector Machines to Automatically Detect Epileptic Activity in EEG Data

Miguel Fernandes

A dissertation submitted to the Faculty of Engineering, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Masters of Science in Engineering.

Johannesburg, October 2011

Declaration

I declare that this dissertation is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this 24 day of OCTOBER 20 11

M Fernandes

Miguel Fernandes.

Abstract

This dissertation evaluates the effectiveness of using Support Vector Machines (SVM) to identify Inter-ictal epileptic activity in Electroencephalogram (EEG) data. There are existing systems that already do this but identifying the best solution requires comparative studies. A sample of data was randomly selected from 20 patients. A number of features were extracted and a PBIL algorithm was then used to identify the set of features that provided the best separability within the dataset. These were found to be related to the chaoticity, frequency and variability of the signal. The features were then used to train an SVM model. This method resulted in 88.7% accuracy but left 31.1% of inputs unclassifiable. This performance is comparable with existing solutions.

The strengths of the designed system were computational efficiency and system accuracy. The limitations were that the high degree of Artifacts masked indicative patterns and therefore decreased the classification accuracy in some cases. This system could be used as a screening tool for detecting patients that possibly have epilepsy. The system cannot be used to rule out epilepsy in a patient due to the high rate of unclassifiable data.

Some areas were identified where future work could be done, namely a spike detection method, multi-channel analysis, and improved Artifact extraction classification. These areas could improve the general performance of the classification system in terms of more clearly separating epileptic from non-epileptic activity.

Acknowledgements

I would like to thank my supervisors, Professor T. Marwala, Professor D. Rubin and M. Perez, for their invaluable assistance and guidance.

I would also like to thank the Department of Electrical, Electronic and Computer Engineering at the University of Pretoria that provided the data used in this research.

Table of Contents

Chapter 1 Introduction

1.1 Motivation

Epilepsy is the term used to describe a serious neural disorder where there is abnormal electrical activity in the brain caused by a specific group of brain cells misfiring resulting in a seziure. A seizure is a temporary alteration in muscular control or consciousness which can take various forms which are not always easy to recognise [1]. The technical definition of Epilepsy is when a person experiences recurrent seizures; a single seizure does not automatically mean that a person has epilepsy [1].

Although this disorder is fairly uncommon in terms of affected people (approximately 5% of the world's population) in comparison to other health problems impacting Africa (for example malnutrition and epidemic or infectious diseases) the effect on people's lives is still significant [2]. This is because the disorder can render a person dysfunctional as sufferers have problems participating in some activities, and in some cases severe seizures can cause permanent or extensive (sometimes fatal) damage.

Epilepsy is a serious disorder but it can be treated in some cases and otherwise controlled. In order to do this it is necessary to have accurate identification of the disorder in individual patients. This identification requires a method of finding a distinctive sign or pattern indicating the presence of this disorder. Epilepsy is a disorder that affects the brain, it is therefore logical to monitor brain activity in some way to identify related abnormal neural activity. The current method of monitoring brain activity is using an Encephalogram (EEG) which measures brain signals detected on the person's scalp by surface electrodes. A characteristic signal pattern has been identified that accompanies epileptic activity and can be measured with the EEG. This signal occurs both during and between seizures, which means that the patient does not have to be undergoing a seizure at the time of the measurement in order for epileptic activity to be detected. It is important to note, however, that identifying this pattern in a patient is not a definitive test for epilepsy but requires clinical diagnosis in order to do so. The usefulness of such a test is that it helps doctors screen out those patients that definitely do not have epilepsy from those who might [3].

There are some problems related to using this characteristic signal: the signal is often masked by underlying brain activity and the signal occurs infrequently. These underlying brain signals are known as Artifacts and are characterised by abnormal non-epileptiform activity. The epileptic signal does not occur continuously over time and so the EEG data would have to be analysed over extended periods of time [3].

The EEG works by printing out large quantities of pages of readouts which display the signals obtained from the scalp electrodes which are scanned for characteristic patterns. This causes two problems: the system is open to human interpretation error and it is time consuming to analyse all the data. In order to combat this there have been many attempts at creating an automated system that analyses the data and recognises the presence of epilepsy. The keyword here is 'recognise' as the underlying mechanics of brain activity are unknown so any detection system has to recognise the epileptic pattern from previous examples. There already exist solutions which use Neural Networks for this task but it is important to investigate all possible methods so that the optimal solution (in terms of the trade-off between accurate prediction and computational speed) is found

[4]. A further benefit of investigating new approaches is to understand why certain methods are better than others for this application.

This dissertation explores the feasibility of creating a system based on Support Vector Machines (SVM) to automatically identify the epileptic pattern. In order to do this various feature extraction, data processing, Artifact extraction and classification methods are analysed.

1.2 Background

In this section the fundamental aspects of the research are introduced. These are: what is Epilepsy and why is it such a problem; the EEG apparatus that is used to measure the related neural activity plus the brain wave patterns that are indicative of epilepsy; and artificial intelligence and how it applies to the automatic detection of epileptic activity in EEG data.

1.2.1 What is Epilepsy

Epilepsy is the term used for a medical condition where a person experiences recurring seizures and is caused by excessive or unusual electrical activity in the brain. This disorder can be present from birth or at a later stage in life due to trauma or disease; anyone can be affected but it is more prevalent in children and the aged. The seizures experienced can take various forms which are not always easy to recognise. [1,5]

There are two types of epilepsy depending on the cause or aetiology. The first is Idiopathic or Constitutional; this is when the cause is unknown and is usually found in people within the 5-20 year old age group. The second type is Symptomatic where the cause is known: metabolic disorders; head trauma; tumours; vascular diseases; degenerative disorders and infectious diseases. The frequency of each can be seen in **Figure1**. The figure shows that Idiopathic Epilepsy is the most common type which suggests that there is a lack of knowledge on the exact causes of this neurological disorder [6].

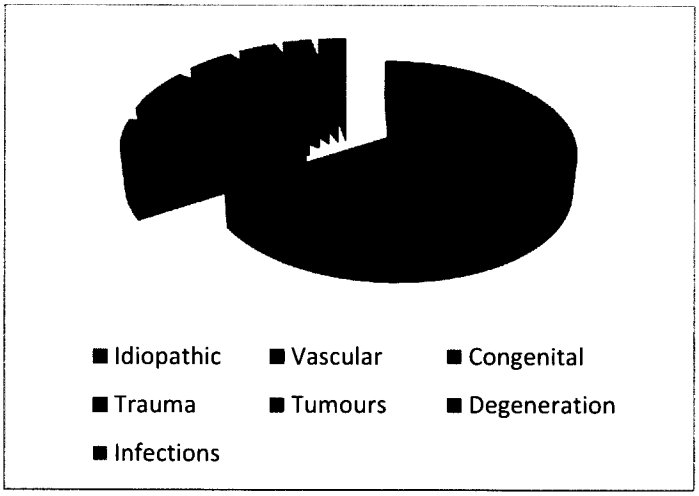

Figure 1: Pie chart of cause of epilepsy (Data source: Epilepsy South Africa) [6]

Seizures are the visible effects of epilepsy and it is therefore important to distinguish between the different types of seizure as well as describe their relative impacts. The two most common types are [6]:

- Partial Seizures. These occur when part of the brain is affected (one cerebral hemisphere). This can be further broken down into:
    - Simple Partial seizures which occur when the person's consciousness is unaffected, and
    - Complex Partial seizures which occur when the person's consciousness is impaired
- Generalized seizures. These occur when both hemispheres of the brain are affected. This can be further broken down into:
    - Absence (petit mal) seizures which occur when a person appears to be staring into space with possible uncontrolled eye movements and are generally short in duration,
    - Myoclonic seizures which occur when the person experiences an extremely brief muscle contraction and can result in jerky muscle movements, and
    - Tonic-clonic (grand mal) seizures which occur when the person experiences a sudden loss of consciousness and are the most serious type. A severe case of this is when the patient has another seizure while still unconscious from the first attack (Status Epilepticus) leading to possible brain damage or even death depending on the severity of the seizure.

If the type of Epilepsy is Symptomatic and the cause of the seizures can be identified then there is a possibility for alleviating or curing the disorder through treatment. On the other hand, if the cause cannot be found (Idiopathic) then the person can take long-term anti-convulsant medication such as Bromide in order to reduce the possible damage incurred while undergoing a seizure. A problem exists where some people do not respond to treatment and it is therefore sometimes required to remove the brain cells that are misfiring through an operation [6].

Apart from the large range of manifestations or types of seizures, it is also important to note that diagnosis is further hampered by other conditions that can mimic epilepsy. Some examples of such conditions are: transient ischemic attacks (TIAs), rage or panic attacks, and other disorders that cause loss of consciousness [2].

There are two main ways, other than witnessing recurrent seizures, to detect the presence of epilepsy in a patient. These are: analysing the EEG data on brain electrical activity and taking blood tests to identify other possible causes for the seizures [6]. This dissertation will look at analysing the EEG data.

1.2.2 The Encephalogram

The EEG is a device used to measure the brain's electrical activity. The two different types are: Scalp and Intra-Cranial which operate on the scalp and inside the head respectively. Only the Scalp EEG will be discussed as it provides sufficient information for diagnosis and is non-invasive. The EEG measures brain activity by measuring potential differences (voltages) between electrodes placed in specific configurations on the scalp. These voltages are caused by small current flows that are generated by synaptic activity. The signal produced is very low in power and so must be amplified

prior to converting it into a digital form or displaying graphically. The EEG data obtained from the electrodes is divided into multiple channels, each of which displays brain waves in a particular section of the brain (based on the placement of the electrodes) [7].

The EEG displays brain waves consisting of a mixture of both normal and abnormal activity. The Epileptic signal is a specific type of abnormal activity but it may be obscured by the presence of the normal activity and so can be hard to detect. The normal brain waves can be categorised into the following basic groups based on frequencies [7]:

- Beta (>13 Hz) waves occur during the normal state of wakefulness
- Alpha (8-13 Hz) waves occur during relaxation or drowsiness
- Theta (4-8 Hz) waves occur during lighter sleep
- Delta (0.5-4 Hz) waves occur during deeper sleep

The brain activity during seizures (Ictal) is characterized by abnormal activity which is defined as Epileptiform and consists of spikes, sharp waves or spike-wave complexes. The problem with waiting for Ictal data is that seizures occur infrequently and so the diagnosis would be too slow. Studies show that approximately 90% of people that have epilepsy also have Epileptiform activity in the intervals between attacks (Inter-ictal). The spike-wave complexes may be isolated or repetitive but usually occur for a shorter period of time than the Ictal (during seizure) signals. The exclusion of the 10% of people who do not experience this activity but have epilepsy is deemed acceptable as this is not a definitive test for epilepsy [3].

1.2.3 Artificial Intelligence

Artificial Intelligence (AI) is a branch of computer science that attempts to understand and build intelligent systems through the use of computational methods. Intelligence is a very broad term whose definition is not clear as it has different interpretations depending on the context. In AI, it is currently taken to mean attributes that require mechanical and / or cognitive processes in order to be realisable. The underlying assumption that forms the basis of AI is the fact that these processes can be modeled using computational means. Most AI problems can be reduced to the State Space Search form whereby the current situation and the desired outcomes are modeled as states and the AI algorithm attempts to search for the optimal path between the two. The dimension of these state spaces can be quite substantial depending on the complexity of the problem being looked at and so a primary concern of any AI algorithm is to represent the space with minimal detail as well as reduce the searchable dimension [8].

AI can be generally separated into two fields: Conventional AI and Computational Intelligence (CI). Conventional AI (also known as: machine learning, symbolic or logical AI) mostly involves the use of formalised methods and statistical analysis. Methods include [9]:

- Expert systems that apply rule-based reasoning in order to reach a conclusion. They can process large amounts of known information and provide conclusions.
- Case based reasoning that attempts to solve current problems based on past solutions to similar problems.
- Bayesian networks that are probabilistic models that measure the relationships between variables and can be used to calculate the probability of unknown variables based on the state of known ones.

- Support Vector Machines that attempt to create a decision hyperplane in the state space in order to separate different classes of data.

Computational Intelligence uses an iterative learning approach (e.g. parameter tuning) based on input from empirical data. Methods include [9]:

- Artificial Neural Networks that are systems made up of interconnected artificial neurons or nodes.
- Fuzzy systems that use approximation and multi-value logic techniques to deal with uncertainty.
- Evolutionary computation that uses an iterative approach to select the desired result from a given population. This selection process is randomised and the improved solution is selected using concepts such as mutation and survival of the fittest. This type of method most notably divides into evolutionary algorithms and swarm intelligence.

One example of each of these fields is discussed and will form the basis of our comparison in future chapters. These two are Artificial Neural Networks and Support Vector Machines respectively.

1.2.3.1 Artificial Neural Networks

ANNs attempt a simplified imitation of the known functionality of the biological neural system. This is done by having an interconnected network of artificial neurons or perceptrons. Each perceptron is a node that takes as input a combination of outputs from other perceptrons and weights them to produce an output. These networks are also similar to biological networks in that functions are performed collectively and in parallel by the nodes, rather than there being a linear or sequential delegation of sub-tasks. The benefits of ANNs are: they can cope with non-linear relationships between inputs and outputs by adding nodes between the input and output, they are generally insensitive to noise, and the parallel nature of the network allows for fast processing speeds. While a neural network does not have to be adaptive, its practical use comes with algorithms designed to alter the strength (weights) of the connections in the network to produce a desired signal flow. The network learns the rules through an iterative process from given training samples [10].

1.2.3.2 Machine Learning

An important concept in AI is learning. Learning is the process in which a model develops a mapping between the inputs and outputs of a given process. In other words learning algorithms attempt to find a function that identifies the optimal solution for a given task based on a set of observations of this task. This requires the definition of a cost function which is a measure of how far away each possibility is from an optimal solution to the problem under observation. The main objective of all learning algorithms is to find a solution that minimises the cost function making the model's mapping closer to the real mapping. A common implementation of machine learning is Support Vector Machines (SVM) [11].

The field of Machine learning consists of the development of algorithms and techniques that allow computers to "learn" the underlying characteristics of empirical data. The learning methods used by these algorithms can be divided into two separate categories: inductive and deductive. Inductive learning tries to learn new information that is not already defined from an existing dataset. Deductive learning, on the other hand, tries to find the most relevant information with an existing dataset [9].

Support Vector Machines

SVMs are a tool that use statistical machine learning for classification or regression tasks. Classification is achieved by differentiating between two different classes based on a set or range of properties that are distinct to each. The development of SVMs was different from that of ANNs in that they were developed from sound statistical theory instead of from experimentation followed by theory [12].

SVMs function by mapping data into a new feature space so that a linear classifier can be used as shown in **Figure2**. The mapping is done with the use of kernel functions which transform the data and is necessary because it may be impossible to distinctly classify the data in linear space. The classifier is a multi-dimensional hyperplane that divides the two classes as they often have more than one dimension. The SVM algorithms attempt to find a classifier that maximises the margin between the data as shown in **Figure3**. The margin of an SVM is defined as the distance from the classifying hyperplane to its closest data point. Maximising the margin provides the system with better generalisation capabilities so that it can perform well with unseen data. The support vectors are those data points that are the closest to the classifying hyperplane and are the only significant points when creating the system as they define the boundaries between the classes. The two main focuses of SVM algorithms are therefore to find a kernel function that optimally represents the differences in the data and to find the maximal margin hyperplane [13].

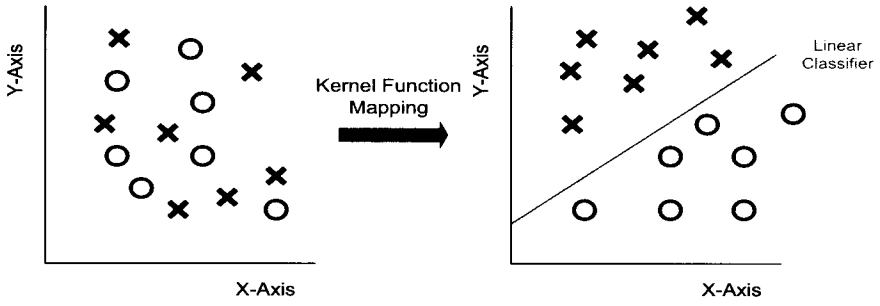

Figure 2: Example of mapping data into a different feature space (taken from results)

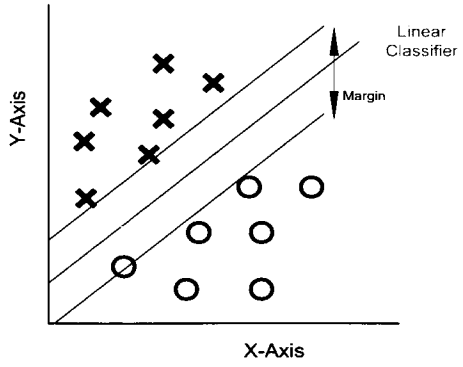

Figure 3: Example showing what a margin is (taken from results)

In real world problems these classes are not always linearly separable, regardless of what kernel function is used and so this simple classification is insufficient. Reasons for this inseparability could be due to the effect of noise or outlying points that do not conform to the rest of the data. This problem can be handled by performing soft margin optimisation which adds a misclassification tolerance decreasing the algorithm's sensitivity to noise and outliers. Other applications of SVMs are: non-linear classifiers and regression problems. The SVM is therefore a versatile tool that rivals and, in some cases, outperforms ANNs [13].

1.2.3.3 Comparison of two methods

ANNs using back-propagation training suffer from slow convergence to an optimal solution. They may also have larger testing errors as compared to SVMs due to the Empirical Risk Minimization (ERM) approach employed by the latter. The Radial Basis Function (RBF) model ANN converges quickly but if the training data is sparse there is no way to reduce the number of hidden neurons required. SVMs are good in that they have good generalisation properties. Some of the problems with SVMs are that they have difficulty in creating a generalised model when the inputs are too sparse and that the algorithm does not scale well to large training sets [14].

1.3 Research Objectives

1. Find a feature set that sufficiently represents Epileptiform activity such that it can be differentiated from normal activity

2. Train a system to automatically detect Epileptiform activity within EEG Data using SVM and the identified feature set

3. Test the feasibility, accuracy and generalization of this system in identifying Epileptiform Activity within the sample data set.

4. Compare the classification accuracy of the system against existing automated systems which use both Artificial Neural Networks and SVM based techniques. All of these systems are based on analysis of EEG data from different population sets.

1.4 Overview of Dissertation

The rest of this work is broken down into the sequential steps taken in order to create an SVM based Epileptiform classification system. **Chapter 2** reviews the existing literature that was relevant to the research, including more detailed descriptions of EEGs and SVMs. It also looks at the basic requirements for a generic detection system and some of the existing solutions to the problem.

Chapter 3 describes the methodology as implemented and describes the different methods and components thereof. **Chapter 4** discusses the various data processing techniques used to transform the data into a usable format as well as decrease the effect of noise on the Epileptiform signal.

Chapter 5 describes the different feature extraction methods used to extract the relevant information about the Epileptiform signal prior to input to the SVM. **Chapter 6** describes the SVM model and strategies used in training and testing the classifier.

Chapter 7 presents the results of the different methods and analyses their performance in relation to each other and compares this to other existing methods. **Chapter 8** concludes the study by identifying whether the objectives were met or not and looks at possible areas of improvement.

Chapter 2 Literature Review

2.1 Overview

This chapter provides an in-depth description and review of the following topics: EEG, SVM and existing epilepsy detection systems. These topics are central to the research performed as they form the building blocks for the methodologies proposed in later chapters.

2.2 The Electroencephalogram

There are many different imaging devices or techniques in the field of medicine that are used to monitor various biological signals. The EEG is one such device and uses metal electrodes to measure electrical impulses in the brain. This information is indicative of various types of neural activity and is useful in the fields of neurology and neurophysiology (the study of the human nervous system and its physiology). There are two types of EEG: scalp (electrocortiogram) and intra-cranial (electrogram). The scalp EEG is measured from the surface of the brain whereas the intra-cranial EEG uses depth electrodes to measure signals within the brain. Only the scalp will be reviewed as it is non-invasive and also sufficient for the purpose of this research [7].

The presence of electrical activity in the brain was identified in 1875 by an English physician called Richard Caton who identified this phenomenon in animals. This concept was then expanded on in 1924 when the German neurologist, Hans Berger, used standard radio equipment to amplify the brain's electrical signals and recorded the results. One of his observations was that the waveforms obtained were different depending on the patient's state of consciousness. He came up with the term EEG and is sometimes credited with inventing the modern EEG [7].

Experts state that human brain activity begins between the 17^{th} to the 23^{rd} week of prenatal development while the child is still in the mother's womb. It is assumed that by this stage the full number of cells (approximately 10^{11} neurons) is already developed. These neurons are then connected in networks through synapses; adults have about 500 trillion synapses on average. As a person grows older the number of synapses per neuron increases as more complex relationships or associations are developed. Although new synapses are constantly being created, the total number of neurons and synapses decrease due to general cell decay. This means that brain activity will initially be very weak and widespread in the brain but will get more concentrated with stronger signals as a person ages [7].

According to the literature the human brain can be divided into three distinct sections: cerebrum, cerebellum and brain stem. The cerebrum is made up of two hemispheres (left and right) whose outermost layer is the cerebral cortex. The cerebral cortex is the area of the brain that handles cognitive functions such as reasoning. The cerebellum is located at the back of the head between the cerebrum and the brain stem and is responsible for muscle control and balance. The last section is the brain stem which is the closest to the spinal cord and is responsible for the involuntary functions such as heart beat and breathing. The scalp EEG mainly measures activity in the cerebral cortex as this area is nearest to the scalp. This is deemed sufficient for the purposes of measuring epileptic activity as seizures can affect consciousness [7].

The EEG records electrical activity in different areas of the brain by measuring the potential differences caused by small current flows between electrodes placed in predefined configurations

on the scalp. These currents are generated via synaptic activity caused by communication between different neurons in the brain. The currents consist mostly of Sodium (Na+), Potassium (K+), Calcium (Ca++), and Chlorine (Cl-) ions and are caused by excitations of the synapses (neuron activation). In order for the signal to be measurable the number of active neurons has to be large as the individual electric currents are very small and so the EEG can only measure a summation of this activity around each electrode. Even if there are a large numbers of active neurons the currents are still small, meaning that the signals must be amplified and filtered in order for it to be meaningful. The filtering is performed as there is a large amount of amplification required in order to make the signal usable which can lead to the addition of a lot of noise. After amplification the signal is then converted to a digital format or printed onto a graph for analysis. The EEG data signals are divided into multiple channels, each of which represents brain waves or activity in a particular section of the brain [7].

2.2.1 EEG System

In order to understand the process involved when recording the EEG brain signals it is important to look at the EEG system: its component parts and the placement patterns of the electrodes. An in depth understanding of the details of the EEG is essential for selection of data processing methods.

2.2.1.1 Components

The EEG recording system consists of [7]:

- Electrodes to measure the currents.
- Amplifiers to increase the signal to a usable level.
- Filters to reduce the presence of noise and Artifacts from the base signal.
- Recording unit to store the results.

The key information obtained by the EEG system is electrical in nature and very small in magnitude, this means that the selection of electrodes is very important as their conductive properties dictate the quality of the data obtained. Higher conductivity leads to greater sensitivity to the small signals emitted by the brain [7].

There are many different electrode types that have suitable conductivity. One of the most common consists of Silver/ Silver Chloride (Ag/AgCl) disks, 1-3 mm in diameter, with long flexible leads that can be plugged into an amplifier. These electrodes are preferred as they can accurately record very slow changes in signal. Another type of electrode is the Needle which, as the name implies, is inserted under the scalp and is used for long recordings [7].

2.2.1.2 Electrode placements

In order to obtain meaningful and reproducible results from an EEG measurement system there is a standardised method of placing the electrodes on the scalp positioning them for maximum results. Manually placing the electrodes on scalps also reduces the reproducibility as exact placements cannot always be achieved. This problem is solved through placing the electrodes on predefined positions on a cap and then placing this cap on the scalp [7].

In 1958, the International Federation in Electroencephalography and Clinical Neurophysiology (IFECN) adopted a standardised system for the physical placement of electrodes on the scalp called 10-20. The 10-20 refers to the distance between electrodes as a percentage of the total front-back

and right-left distance of the scalp. This system divides the head into proportional distances in relation to prominent skull characteristics such as [7]:

- The point where the frontal and nasal bones meet (Nasion).
- The points just above the ears (Pre-auricular).
- The highest point on the back of the head (Inion).

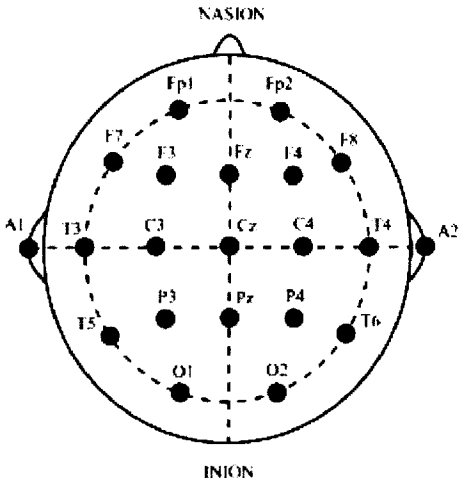

Figure 4: EEG Electrode Placements on Scalp (taken from Fundamentals of EEG Measurement) [7]

Figure 4 shows the electrode placement on a human scalp. The labelling of the electrode positions is done according to the respective brain area and is as follows: F (frontal), C (central), T (temporal), P (posterior), and O (occipital). The letters are followed by odd numbers on the left side of the head and with even numbers on the right side, the orientation of which is determined from the patient's perspective [7].

The EEG is used to monitor brain activity and so the electrodes need to be placed near the centres of such activity. These positions do not accurately reflect the centres described due to the fact that the shape of the skull is non-homogeneous as well as being different between people. They are, however, sufficiently accurate in order to monitor the activity. The electrodes are near the following centres [7]:

- F7 is near the part of the brain responsible for rational activity
- Fz is near the part of the brain responsible for intention and motivation.
- F8 is near the part of the brain responsible for emotional impulses.
- C3, C4, and Cz are near the part of the brain responsible for sensory and motor functions.
- P3, P4, and Pz are near the part of the brain responsible for perception and differentiation.
- T3 and T4 are near the part of the brain responsible for emotional processors.
- T5, T6 are near the part of the brain responsible for certain memory functions.
- O1 and O2 are above the part of the brain primarily responsible for vision.

Consideration of Impedance characteristics play a very important role in deciding where to place electrodes. This is because very high source impedance can lead to distortions in the brain signal. The signals under consideration are small in nature, so any such distortion would be hard to identify and separate from the underlying data. The literature states that in order to prevent these distortions the impedances at each electrode should all be below 5 K Ohms, and within 1 K Ohm of each other [7].

When measuring the signals produced in the brain it is necessary to have a reference point for comparison purposes. The two most common reference points are the ears and the Cz vertex. Both ears are used as a linked reference in order to treat the activity from each hemisphere equally. The problem with using this linked reference is that it may drift from the median if the electrical resistance at each electrode differs. Using the Cz vertex as a reference can be a better option as it is located in the middle among the active electrodes, however the resolution for close points is very poor. It is also possible to use Reference-free techniques which are represented using a common/weighted average reference. These are often preferred as such techniques do not suffer from the problems associated with using a physical reference [7].

2.2.1.3 Amplifiers and filters

In order to be able to measure or display the signals picked up by the electrodes it is necessary to amplify them as they are very small. The input tolerances or sensitivities of most digital devices are also not high enough to pick up these signals. The amplifier therefore needs to have a high gain and must be selective. This means that only the physiological signal must be augmented and all other components such as noise and interference must be rejected. The amplification, however, cannot be too great as this could potentially lead to damage to either the patient or the equipment [7].

An amplifier takes two signals as input and produces an augmented version of the difference between the two as its output. There are two important considerations when designing or selecting an amplifier: the gain and the common mode rejection ratio (CMRR). The gain of an amplifier is defined as the ratio between the input and output signals whereby a high gain denotes a large degree of amplification. In order to obtain a usable signal from the scalp the EEG must provide a gain of 100 to 100 000 and also needs to maintain a high signal to noise ratio (SNR). SNR is the ratio between the signal and system noise, a high SNR means that the effect of noise is minimal. CMRR is the ratio of the gain of the differential signal and the signal that is common to both (in this case noise). In order to further reduce the effect of noise on the output signal differential amplifiers are designed to have high CMRR [7].

There are other ways to minimise the effects of interference such as noise after amplification, the most common method being the use of filters to remove the unwanted signals. These filters try and differentiate between signal types based on their relative frequency compositions. Two examples of filters used prior to storing the data are: a high-pass filter and low-pass filter. The high-pass filter removes the low frequency signals caused by natural rhythmic activity such as breathing whose cut-off frequency is usually within the 0.1-0.7 Hz range. The low-pass filter cuts off all the frequencies that are more than half the digital sampling rate (also known as the Nyquist frequency), ensuring that there is no interference in the sampling process (when converting the analogue signal to a digital format) due to aliasing [7].

2.2.2 Uses for EEG

Although there are many different imaging techniques for monitoring brain activity the benefit from using the EEG is its speed. This is because the effect of stimuli is detectable in the output signal within fractions of a second of occurrence. One of the drawbacks of EEG when compared to methods such as MRI and PET is that it does not have as good spatial resolution as it primarily focuses on the upper levels of the brain (this is particularly true for the scalp EEG). These limitations are deemed acceptable as the resolution is sufficient to detect the presence of Eplieptiform activity and EEG offers a method of directly measuring this [7].

2.2.3 Signal Components

2.2.3.1 Standard Brain Waves

In order to identify Eplieptiform activity it is necessary to have a method of differentiation from other brain activity. In order to do this it is important to understand the different types of signals that occur in the brain and are picked up by the EEG. Normal brain activity forms wave patterns that are commonly sinusoidal in nature and have a peak-to-peak amplitude of 0.5 to 100 μV. The dominant brain pattern is quite often indicative of the mental state of the individual. The key indicator of these patterns is the frequency range which can be obtained from the signal's power spectrum using a Fourier Transform. The most common brain wave types can be categorized into four groups (Examples of the EEG outputs are shown in **Figure 5** below) [7]:

- Beta (These waves have any frequency greater than 13 Hz),
- Alpha (These wave have a frequency between 8-13 Hz),
- Theta (These waves have a frequency between 4-8 Hz), and
- Delta (These waves have a frequency between 0.5-4 Hz).

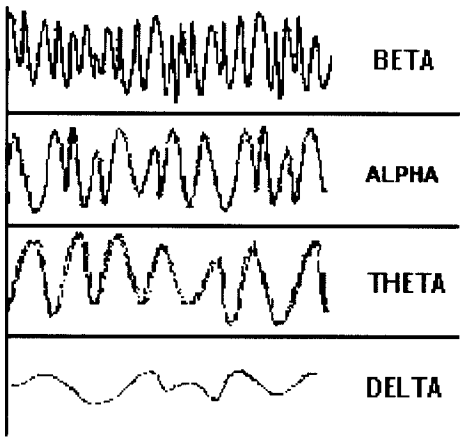

Figure 5: Wave Pattern of the four most common brain wave signals (taken from Fundamentals of EEG Measurement) [7]

The frequency of these waves can be equated to the level of activity in the brain with the highest wakeful state being Beta waves. The next level of brain activity is Alpha waves and these occur when the individual is in a state of relaxation or drowsiness. The last two levels occur while the individual is in a sleep state and represent different depths of sleep. Human sleep can be divided into two main types: non rapid eye movement (NREM) and rapid eye movement (REM). These types occur in alternating cycles with NREM being the deeper state and REM dominated by Theta waves. NREM

can be further divided into stages I - IV with the last two stages corresponding to deeper sleep with predominantly slow delta activity [7].

Alpha Waves are the most understood of the four and occur in the posterior and occipital regions of the brain. This wave pattern is characterised by an amplitude of approximately 50 μV from peak to peak and typically occurs when an individual closes their eyes or is in a relaxed state. This activity can start immediately after most individuals close their eyes as the change in brain activity is rapidly noticeable. The reverse is also true in that this wave pattern can stop immediately as soon as the eyes are reopened or some form of mental processing occurs. The precise of the origin of this wave are still unknown but are thought to be caused by the sum of dendrite potentials. These four wave patterns form the basic composition of standard brain activity and some of these will usually be present in any EEG reading [7].

The different regions of the brain do not experience the same brain wave frequency simultaneously and therefore an EEG signal between electrodes placed on the scalp can consist of multiple waves with different characteristics. This means that there is a large amount of data that is received from even one single EEG recording, making analysis very difficult and time intensive. Another problem is that every individual's brain wave patterns are unique and so an exact comparison cannot be achieved. The patterns are so unique that, in some cases, it is possible to distinguish persons only according to their typical brain activity. An example of this is that individuals who are more rational experience higher activity in the frontal left hemisphere whereas more holistic / intuitive individuals experience frontal left hemisphere. This means that signal patterns cannot be exactly defined and need to be relative to the individual [7].

2.2.3.2 Epileptiform Activity

In order to detect the potential presence of Epilepsy in the EEG readout of an individual there needs to be a method of detecting Epileptiform Activity. There are two observable types: Ictal (during seizures) and Inter-ictal (between seizures). Epileptic seizures usually occur infrequently and unpredictably so their occurrence cannot be determined with any certainty. As a result of this, the detection of Ictal activity can only be performed using long-term EEG recordings of up to a week. Inter-ictal activity, on the other hand, can occur more frequently and so shorter term readings can be performed prior to occurrence. Although not all patients experience this activity, the 90% that do make its a presence a good indicator of Epilepsy. The reverse of this statement, however, is not true as the absence of Inter-ictal activity does not mean that an individual does not have epilepsy [11].

Inter-ictal Epileptiform activity on the EEG readout is characterised by sharp or sudden transient waves of higher amplitude than the underlying activity. There are two variations of this waveform: Spike and Sharp Waves (SSWs); Spike and Wave Complex (SWC). Spikes and waves can be differentiated using their durations: spikes are more sudden (20 – 70 ms in duration) whereas waves are more gradual (70 – 200 ms). SSWs (Epileptiform Discharges, Spikes, Epileptiform Transients, sharp transients) consist of sharp, transient waveforms with a pointed peak value. SWCs occur when SSWs are followed by a slow delta or theta wave pattern. This activity is a good indicator as it is considered abnormal for adults to exhibit delta activity while in a state of wakefulness [11].

Measurement of Epileptiform activity is further complicated by the fact that it can occur simultaneously on more than one channel (in some cases almost all channels) of the EEG. This means that Epileptiform activity cannot be isolated to just one channel and all channels need to be

analysed simultaneously. Spikes and Sharp Waves are similar in nature to an impulse function which contains all frequencies. This causes an increase in spectral power over a wide range of frequencies, thereby affecting other wave patterns. The literature states that individual SSW and Spike and Wave Complex waveforms contain very little diagnosis information and are mostly used for the confirming the existence of epileptic activity [11].

2.2.3.3 Artifacts

A common problem in EEG analysis is that the desired signal is masked or distorted by activity other than the normal brain wave patterns. This activity comes in the form of Artifacts which can usually be distinguished from normal brain activity by their higher amplitudes and morphological differences. There are two basic types of Artifacts that can be identified in EEG signals: patient-induced and technically-induced. Patient-related Artifacts are caused by interference from other physiological activity such as eye or muscle movement. Technical Artifacts are caused by interference from external sources such as noise from the AC power line or some other background noise. The effect of these Technical Artifacts can be reduced by minimizing the impedances of the electrodes and by decreasing the length of the electrical connections (wires). The most common types of EEG Artifact sources are displayed in the table below [7].

Artifact Type	**Artifact Source**
Patient related	Minor body movements
Patient related	Muscle activity (EMG)
Patient related	ECG (pulse, pace-maker)
Patient related	Eye movements (EOG)
Patient related	Sweating
Technical	50/60 Hz noise
Technical	Impedance fluctuation
Technical	Cable movements
Technical	Broken wire contacts
Technical	Too much electrode paste/jelly or dried pieces
Technical	Low battery

Table 1: Artifact Types and Reasons

The EEG segments that are contaminated can be excluded from the recording by either a trained expert or an automatic process. In order to more easily identify these segments it is useful to have electrodes that monitor the patient's other physiological signals. The problem with this method of identification and exclusion is that a lot of potentially important data is removed along with the Artifact signals. The best way would therefore be to somehow remove the effect of these Artifacts without affecting the underlying signals [7].

2.3 Support Vector Machines

SVMs, unlike ANNs, were developed from statistical theory into a model that implements this theory. The initial statistical work was done by Vapnik and Chervonenkis but was not widely accepted due to the belief that it was not practically useful. This belief changed with the achievement of good results in practical learning applications such as text categorisation. SVMs have been worked on a great deal so that they currently display similar (in some cases better) results to ANNs and other statistical models [13].

The main aim of an SVM (as with most AI) is to 'learn' the underlying function f(x) that maps a given input vector **x** to an observed or desired output y. There is usually no information on the joint probability density functions of the input vectors. The main objective of the model is therefore to use a supervised learning approach to find a best fit function *f(x)* that maps the training data pairs (x_i, y_i) where x_i is the input that produces the output y_i [13].

This section looks at how SVM classification of Epileptiform Activity can be modelled as an optimization problem of the fitness function and then solved. The first section discusses the creation of a cost function to minimise the error while still keeping good generalisation performance. The second section discusses the processes and methods used for classification. The last section discusses the optimization of the solution [13].

2.3.1 Risk Minimization and Generalization

Risk minimization is the reduction of errors in a classification system through the minimization of a relevant cost function. This error is defined as the difference between the predicted classification and the actual classification. Generalization is the ability of a classification to perform well on unseen data whose patterns may not be similar to those used in the training cycle. These two concepts can be in conflict as a higher degree of accuracy on the training set could overfit the mapping and not be a true reflection of the globally optimal classifying function. A trade-off must therefore be achieved between them in order to obtain a solution that fits the training data well but will also achieve similar performance with previously unseen data [13].

This trade-off or balance is achieved in SVMs using a concept called Structural Risk Minimization (SRM). SRM is an inductive principle that improves learning over finite (sparse) data sets. It attempts to balance risk minimization with model capacity where capacity is equivalent to the complexity of the model. A lower capacity generally means that the model is a more generic fit to the underlying function, signifying that the error rate for the training set will be relatively high but will be more uniform over unseen data. SVMs use a parameter called the Vapnik-Chervonenkis (VC) dimension in order to measure this capacity [13].

The VC dimension is a measure of the capacity of a statistical classification algorithm. The VC dimension of a model *f* is the maximum number of points *h* such that some data point set of

cardinality h can be shattered by f. Shattering is when the function fits the points exactly (meaning the model makes no errors when evaluating the input set). This concept is useful for learning applications as it can predict a probabilistic upper bound on the test error of a classification model. The generalization error of the SVM is bounded with probability 1 - η, by the following function [15]:

$$R(w,b) \leq R_{emp}(w,b) + \emptyset \tag{1}$$

Where $R_{emp}(w,b)$ is the classification error and $\emptyset$ is the VC confidence interval for the unknown data and denotes the difference between the actual and estimated errors [16]:

$$\phi = \sqrt{\frac{h\left(\log\left(2N/h\right)+1\right)-\log\left(\eta/4\right)}{N}} \tag{2}$$

where h is the VC dimension of the classification model, and N is the size of the training set. The error bound can be estimated using this parameter as follows:

$$\text{Error Bound} = \text{Empirical Error} + \phi \tag{3}$$

For a hard-margin SVM the VC dimension, h, of the hyperplanes with margin $\|w\|^{-1}$ is bounded by [14]:

$$h \leq \min(D^2\|w\|^2, l) + 1 \tag{4}$$

where D is the diameter of the smallest hypersphere that includes all the training data.

A learning machine with large capacity (hence large VC dimension) will produce a low empirical risk but the VC confidence interval would also be large, indicating poor generalization performance. The measurement of this bound facilitates the selection of a function that has a low expected risk together with good generalization performance. When the training data are linearly separable in the feature space the empirical risk $R_{emp}(w,b)$ is zero and thus minimizing h will minimize the VC dimension. This is achieved by maximizing the margin $\|w\|^{-1}$. If the training data are not linearly separable then the empirical risk is not zero and the generalization ability is determined by the trade-off between $R_{emp}(w,b)$ and $\emptyset$ via the model parameter C [13,16].

2.3.2 Classification and Problem Formulation

SVMs attempt to find a separating hyperplane that maximizes the margin to the closest data points (support vectors) and separates the classes under investigation. The linear case of this problem is the simplest and occurs when the classes can be exactly separated by a linear hyperplane (line). The linear case forms the basis of the SVM process but most real world applications cannot be classified this simply or exactly. This is catered for with modifications to the linear solution when the data is non-separable or the separating hyperplane is non-linear [13].

2.3.3 Maximal Margin Classification

The simplest implementation of SVMs is the maximal margin classifier and is applicable only to the linearly separable case. The first step is to place bounds on the generalization error of the linear machine in terms of the margin $m_S(f)$ of a classifying function f with respect to the training set S as follows [13]:

$$err(f) \leq \varepsilon(l, L, \delta, \gamma) = \frac{2}{l}\left(\frac{64R^2}{\gamma^2}\log\frac{el\gamma}{4R}\log\frac{128lR^2}{\gamma^2} + \log\frac{4}{\delta}\right) \tag{5}$$

Provided:

$$l > 2/\varepsilon \text{ and } 64R^2/\gamma^2 < l \tag{6}$$

where l is the number of random samples, R is the radius around the origin of a sphere around the origin with probability [13]:

$$1-\delta,\ m_S(f) \geq \gamma \tag{7}$$

This bound does not contain any restrictions on the dimensionality of the separation feature space and can therefore be extended to higher dimensional spaces. The Maximal margin classifier optimizes this bound by separating the data through the selection of a hyperplane with maximal margin [13].

The next step in the process is to reduce the problem to that of convex optimization which involves the minimization of a quadratic function under linear constraints. It is useful to work with Convex optimization problems as any local minimum of the unconstrained optimization problem with convex objective function f is also a global minimum. This means that finding the local minimum is the same as finding the global minimum [13].

A function is said to be convex for:

$W \in N$ if, $\forall\ w, u \in N$, and for any $\theta \in (0,1)$

$$f(\theta\omega + (1-\theta)\mu) \leq \theta f(\omega) + (1-\theta) f(\mu) \tag{8}$$

This formula ensures that the shape of the function is convex in nature. The SVM problem meets the conditions for convexness as shown when discussing kernel functions [13].

The optimal classifying function should be chosen in order to minimize a certain functional, in the case of SVMs this is in the form of a cost function that is subject to constraints. The Geometric margin of a hyperplane is the functional margin of a normalized weight vector. This means that we can optimize the geometric margin by fixing the functional margin to 1 (canonical hyperplane) and minimizing the weight vector. A Functional margin of 1 implies the following [13]:

$$\begin{aligned} \langle \omega \cdot x^+ \rangle + b &= +1 \\ \langle \omega \cdot x^- \rangle + b &= -1 \end{aligned} \tag{9}$$

in order to compute the Geometric margin we must normalize ω. The geometric margin γ is the functional margin of the resulting classifier

$$\begin{aligned} \gamma &= \frac{1}{2}\left(\left\langle \frac{\omega}{\|\omega\|_2}\cdot x^{+}\right\rangle - \left\langle \frac{\omega}{\|\omega\|_2}\cdot x^{-}\right\rangle\right) \\ &= \frac{1}{2\|\omega\|_2}\left(\langle \omega\cdot x^{+}\rangle - \langle \omega\cdot x^{-}\rangle\right) \\ &= \frac{1}{\|\omega\|_2} \end{aligned} \tag{10}$$

The optimization problem can be stated in the following format

$$\begin{aligned} &minimise_{\omega,b} \quad \langle \omega\cdot\omega\rangle, \\ &\text{subject to} \\ &y_i\left(\langle \omega\cdot x_i\rangle + b\right)\geq 1, \\ &\quad i=1,\ldots,l \end{aligned} \tag{11}$$

which is in the form: objective function plus both the inequality and equality constraints.

The next step is to transform this optimization problem into its corresponding Lagrangian form. In constrained problems a function, known as the Lagrangian, has to be defined that incorporates information about both the objective function and its constraints. The stationarity of this function can be used to calculate solutions. A more precise description of a Lagrangian funtion is as an objective function plus a linear combination of its underlying constraints, where the coefficients of the combination are called the Lagrange multipliers [13].

Given an optimization problem with objective function $f(\omega)$, and equality constraints $h_i(\omega)=0,\ i=1,\ldots,m$

we define the Lagrangian function as

$$L(\omega,\beta)=f(\omega)+\sum_{i=1}^{m}\beta_i h_i(\omega) \tag{12}$$

where the coefficients β_i are called the Lagrange Multipliers.

In Lagrangian theory there exists a primal and a dual formulation of the problem where the dual format uses a different co-ordinate system for the data. This dual system takes into account the misclassification rate where an optimal solution is reached when the difference between the primal and dual (duality gap) is minimized [13,17].

The primal of the maximal margin classification problem is:

$$L(w,b,\alpha)=\frac{1}{2}\langle w\cdot w\rangle-\sum_{i=1}^{l}\alpha_i\left[y_i\left(\langle w\cdot x_i\rangle+b\right)-1\right] \tag{13}$$

where $\alpha_i \geq 0$ are the Lagrange Multipliers.

The dual form of this equation can now be found by differentiating the primal with respect to ω and b, thereby ensuring stationarity. This differentiation also removes any dependence on the primal variables from the dual formulation. This substitution corresponds to explicitly computing the function [13,16]:

$$\theta(\alpha,\beta)=\inf_{\omega\in\Omega} L(\omega,\alpha,\beta)$$

$$\begin{aligned}\frac{\partial L(\omega,b,\alpha)}{\partial\omega} &= \omega-\sum_{i=1}^{l} y_i\alpha_i x_i=\mathbf{0},\\ \frac{\partial L(\omega,b,\alpha)}{\partial b} &= \sum_{i=1}^{l} y_i\alpha_i=0\end{aligned} \tag{14}$$

obtaining the following relations:

$$\begin{aligned}\omega &= \sum_{i=1}^{l} y_i\alpha_i x_i,\\ 0 &= \sum_{i=1}^{l} y_i\alpha_i\end{aligned} \tag{15}$$

which we substitute back into the primal to get

$$\begin{aligned}L(\omega,b,\alpha) &= \frac{1}{2}\langle\omega\cdot\omega\rangle-\sum_{i=1}^{l}\alpha_i\left[y_i\left(\langle\omega\cdot x_i\rangle+b\right)-1\right]\\ &= \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j\alpha_i\alpha_j\langle x_i\cdot x_j\rangle-\sum_{i,j=1}^{l} y_i y_j\alpha_i\alpha_j\langle x_i\cdot x_j\rangle+\sum_{i=1}^{l}\alpha_i\\ &= \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l} y_i y_j\alpha_i\alpha_j\langle x_i\cdot x_j\rangle\end{aligned} \tag{16}$$

The value of b in the above formula cannot be calculated using the dual and so we use the following primal constraints:

$$b^*=-\frac{max_{y_i=-1}\left(\langle w^*\cdot x_i\rangle\right)+min_{y_i=1}\left(\langle w^*\cdot x_i\rangle\right)}{2} \tag{17}$$

The next concept to look at is the Karush-Kuhn-Tcuker (KKT) conditions which provide some constraints that ensure a solution's optimality:

$$\alpha_i^*\left[y_i\left(\langle w^*\cdot x_i\rangle+b^*\right)-1\right]=0, i=1,\ldots,l \tag{18}$$

This statement implies that only the inputs with unit functional variance (the points lying closest to the hyperplane) have non-zero α_i . These are points are therefore the only important points of the weight vector and so are called support vectors [13].

Another important consequence of the KKT conditions is that:

$$y_i f(x_j, \alpha^*, b^*) = y_j \left(\sum_{i \in SV} y_i \alpha_i^* \langle x_i \cdot x_j \rangle + b^* \right) = 1 for j \in SV \tag{19}$$

and therefore

$$\begin{aligned} \langle w^* \cdot w^* \rangle &= \sum_{i,j=1}^{l} y_i y_j \alpha_i^* \alpha_j^* \langle x_i \cdot x_j \rangle \\ &= \sum_{j \in SV} \alpha_j^* y_j \sum_{i \in SV} y_i \alpha_i^* \langle x_i \cdot x_j \rangle \\ &= \sum_{j \in SV} \alpha_j^* (1 - y_j b^*) \\ &= \sum_{i \in SV} \alpha_i^* \end{aligned} \tag{20}$$

demonstrating that the objective function can be completely defined by the sum of the Lagrangian multipliers.

Both the dual objective and the decision function only represent the data points inside an inner product, making it possible to find and use optimal hyperplanes in the feature space through the application of kernels as shown in the following section. This formulation only deals with exactly separable linear problems [13].

2.3.4 Kernels

In order to extend the previous formulation to deal with non-linear cases it is necessary to introduce the concept of kernels. Kernel representations or functions project the data into a higher dimensional feature space in which the classes are linearly separable thereby extending the usefulness of the standard learning machine. The use of the dual representation is fundamental to the application of these kernels as this makes the application of this transformation step implicit. This is due to the fact that the adjustable parameters in the dual representation do not depend on the number of attributes being used. Replacing the inner product in the dual representation with an appropriate kernel function enables the use of a non-linear mapping to a high dimensional feature space without increasing or affecting the number of parameters [13,16].

A kernel is a function K, such that for all $x, z \in X$

$$K(x, y) = \langle \varphi(x) \cdot \varphi(z) \rangle \tag{21}$$

The only information from the data points used by the Kernel function is their Gram matrix in the feature space which is defined as:

Given a set $S = \{x_1, \ldots, x_n\}$ of vectors from an inner product space X, the $n \times n$ matrix **G** with entries $G_{ij} = \langle x_i \cdot x_j \rangle$ is called the Gram Matrix of S. This matrix is also known as the kernel matrix, we now denote it with K [13].

In order for a function to be suitable as a kernel for some feature space, it must have the following properties: it must be symmetric and must satisfy the inequalities that follow from the Cauchy-

Schwarz inequality. These properties are, however, not sufficient to guarantee a feature space, in order to do so the kernel function must also meet Mercer's condition which is as follows [13]:

Let X be a finite input space with $K(x,z)$ a symmetric function on X. Then $K(x,z)$ is a kernel function in and only if the matrix

$$K=\left(K\left(x_i,x_j\right)\right)_{i,j=1}^{n} \tag{22}$$

is positive semi-definite (has non-negative eigenvalues). This condition also ensures that the function is convex is nature and therefore has no local minima [13].

There are two methods of making kernel functions [13]:

- Transformation of existing kernels
- Working out the inner product of the features

The latter method is easier to implement because it does not need to check for conformity to Mercer's condition as this is one of the properties of an inner product [13,16].

The maximal margin optimization problem can therefore be rewritten and substitute the Kernel function into the dual Lagrangian as follows:

$$\begin{aligned} maximise W(\alpha) &= \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K\left(x_i,x_j\right) \\ subject to &\sum_{i=1}^{l} y_i \alpha_i = 0, \\ &\alpha_i \geq 0, i = 1,\ldots,l \end{aligned} \tag{23}$$

Then the decision rule separating the classes given by sign(f(**x**)), where

$$f(x)=\sum_{i=1}^{l} y_i \alpha_i^* K\left(x_i,x\right)+b^* \tag{24}$$

is equivalent to the maximal margin hyperplane in the feature space defined by the chosen kernel function. The corresponding geometric margin of this hyperplane can be expressed as [13,16]:

$$\gamma=\left(\sum_{i\in SV}\alpha_i^*\right)^{-1/2} \tag{25}$$

The strong duality theorem states that there is no duality gap for an optimal solution, this difference can therefore be used as a measure of the solutions convergence to optimum and is also known as the feasibility gap. Let α be the current value dual variables. The weight vector is calculated from setting the derivative of the Lagrangian to zero, and so the current value of the weight vector ω is the one which minimizes $L(\omega, b, \hat{\alpha})$ for the given $\hat{\alpha}$ where α is the current value of the dual variables. This difference can now be calculated [13,16]:

$$W(\hat{\alpha}) - \frac{1}{2}\|\hat{\omega}\| = \inf_{\omega,b} L(\omega, b, \hat{\alpha}) - \frac{1}{2}\|\hat{\omega}\|^2$$

$$= L(\omega, b, \hat{\alpha}) - \frac{1}{2}\|\hat{\omega}\|^2$$
$$= -\sum_{i=1}^{l} \hat{\alpha}_i [y_i((\hat{\omega} \cdot x_i) + b) - 1]$$
$$= \sum_{i=1}^{l} \alpha_i - \sum_{i,j=1}^{l} \hat{\alpha}_i y_i y_j \hat{\alpha}_j \langle x_j \cdot x_i \rangle$$

(26)

which is minus the sum of the KKT conditions. This difference corresponds to the feasibility gap provided ω satisfies the primal constraints:

$$y_i((\hat{\omega} \cdot x_i) + b) \geq 1 \forall i \quad (27)$$

which is equivalent to:

$$y_i(\sum_{j=1}^{l} y_j \hat{\alpha}_j \langle x_j \cdot x_i \rangle + b) \geq 1 \quad (28)$$

There is, however, no guarantee that this will hold and so calculation of the feasibility gap for the maximal margin problem is not straightforward. The use of a soft margin, however, makes it is possible to estimate this property [13].

The fact that only a subset of the Lagrange multipliers is non-zero is referred to as sparseness, and means that the support vectors contain all the information necessary for defining the separating hyperplane. This means that the other points can effectively be removed from the calculation of the classifier meaning that the representation of the system is compressed [13].

Another important point is that fewer support vectors in the solution also leads to better generalization performance as shown in:

$$err(f) \leq \frac{1}{l-2}\left(d \log \frac{el}{d} + \log \frac{l}{\delta}\right) \quad (29)$$

where d = number of support vectors, l = number of random samples and the probability equals $1-\delta$ [13].

2.3.5 Soft Margin Optimization

The maximal margin classifier is an important concept and it forms the basis for SVM classification but it cannot be used in real world applications as it cannot handle non-separable data. If the data are noisy, there will in general be no linear separation in the feature space unless very powerful kernels are used which leads to overfitting and poor generalization. The main problem with the maximal margin problem is that it always attempts to find a classifier with no training error. This is due to the fact that the error bound depends on the margin which is negative unless the data are exactly separable [13,16].

This dependence on the margin makes the classification solution overly sensitive to the effect of outliers within the data. The sensitivity can be demonstrated with the fact that the primary and dual

Lagrangian formulations have the following problems with non-separable data: the primal has an empty feasible region and the dual an unbounded objective function. It is therefore necessary to introduce more robust measures such as margin distribution which can tolerate noise and outliers, and utilises more points than just the support vectors [13].

The functional margin extends the concept of the maximal margin by defining the margin of an example (x_i, y_i) with respect to a hyperplane (ω, b) to be:

$$\gamma_i = y_i\left(\langle \omega \cdot x_i \rangle + b\right) \tag{30}$$

or more simply

$$\gamma_i = y_i f(x_i) \tag{31}$$

where a result of $\gamma_i > 0$ signifies correct classification and $f(x_i)$ is the classifying hyperplane. The margin distribution is defined as the distribution of the margins for the given sample set [13,16].

Soft margin optimization uses slack variables to cater for discrepancies between the margin and classification of each point. The optimization problem needs to be changed by introducing slack variables to the optimisation constraints allowing for non-conformity to the calculated margin [13,16].

$$\begin{aligned} y_i\left(\langle \omega \cdot x \rangle + b\right) &\geq 1 - \xi_i, 1 = 1, \ldots, l \\ \xi_i &\geq 0, i = 1, \ldots, l \end{aligned} \tag{32}$$

where ξ_i are the slack variables for each data point.

There are two methods of calculating at the optimization of the slack vector: 1-norm and 2-norm where n-norm of a vector x is given by:

$$\|x_n\| = \left(\sum_{i=1} |x_i^n|\right)^{\frac{1}{n}} \tag{33}$$

This equation shows that the 1-norm of a vector is the rectangular distance between the point and the origin whereas the 2-norm is the Euclidean (or straight line) distance [13,16].

The optimization problem can then be rewritten as:

$$\begin{aligned} & minimise_{\xi,\omega,b} \langle \omega \cdot \omega \rangle + C\sum_{i=1}^{l} \xi_i \\ subject to \quad & y_i\left(\langle \omega \cdot x_i \rangle + b\right) \geq 1 - \xi_i, i = 1, \ldots, l \\ & \xi_i \geq 0, i = 1, \ldots, l \end{aligned} \tag{34}$$

The two methods of calculating the margin for the 2-norm and 1-norm soft margin problems are discussed respectively.

The primal Lagrangian of the 2-norm problem is:

$$L(\omega,b,\xi,\alpha)=\frac{1}{2}\langle\omega\cdot\omega\rangle+\frac{C}{2}\sum_{i=1}^{l}\xi_i^2-\sum_{i=1}^{l}\alpha_i\left[y_i\left(\langle\omega\cdot x_i\rangle+b\right)-1+\xi_i\right] \tag{35}$$

The dual Lagrangian of this problem resolves to:

$$\begin{aligned}L(\omega,b,\xi,\alpha)&=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\langle x_i\cdot x_j\rangle+\frac{1}{2C}\langle\alpha\cdot\alpha\rangle-\frac{1}{C}\langle\alpha\cdot\alpha\rangle\\&=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\langle x_i\cdot x_j\rangle-\frac{1}{2C}\langle\alpha\cdot\alpha\rangle\end{aligned} \tag{36}$$

Hence maximizing the above function over α is equivalent to maximizing the following equation:

$$W(\alpha)-\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\left(\langle x_i\cdot x_j\rangle+\frac{1}{C}\delta_{ij}\right) \tag{37}$$

where δ_{ij} is the Kronecker δ defined to be 1 if i = j = 0.

The geometric margin of the 2-norm solution to this problem is therefore:

$$\gamma=\left(\sum_{i\in SV}\alpha_i^*-\frac{1}{C}\langle\alpha^*\cdot\alpha^*\rangle\right)^{-1/2} \tag{38}$$

The primal Lagrangian of the 1-norm problem is:

$$L(\omega,b,\xi,\alpha,r)=\frac{1}{2}\langle\omega\cdot\omega\rangle+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i\left[y_i\left(\langle x_i\cdot\omega\rangle+b\right)-1+\xi_i\right]-\sum_{i=1}^{l}r_i\xi_i \tag{39}$$

The dual Lagrangian of this problem resolves to:

$$L(\omega,b,\xi,\alpha,r)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\langle x_i\cdot x_j\rangle \tag{40}$$

The geometric margin of the 2-norm solution to this problem is therefore:

$$\gamma=\left(\sum_{i,j\in SV}y_iy_j\alpha_i^*\alpha_j^*K(x_i,x_j)\right)^{-1/2} \tag{41}$$

which is equivalent to the maximal margin case with the additional constraints that all α_i are upper bounded by **C**.

One of the problems associated with using the 1-norm and 2-norm soft margin optimization methods is the selection of **C** as a range of values have to be tested prior to selection. The scale of **C** can also vary depending on the selected feature space [13,16].

The equations show that the soft margin methods are simply modifications of the maximal margin case [13].

2.4 Existing systems

There exist various types of systems to detect epilepsy in EEG data. The systems have been implemented with varying degrees of success depending on the application. There has been a lot of work on spike detection systems and not so much on overall epileptic activity identifiers (which is the scope of this work). There are some generic components that are common to all of these that are necessary in any detection system. This section describes some of these components then goes on to list some of the more relevant examples and their respective successes.

2.4.1 Standard components

Any detection system requires the following components:

- Data processing
- Feature extraction
- Classification

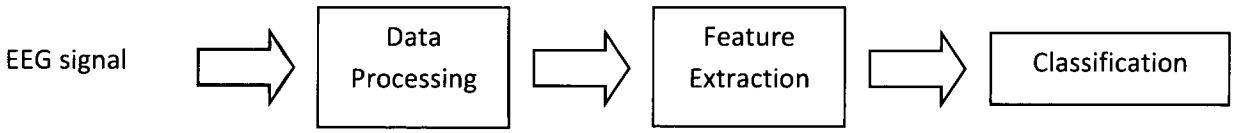

Figure 6: Standard Components of a Classification System

The data processing includes whitening, normalizing and filtering of the data. These processing techniques help clear the data of unwanted signal elements, simplifying the classification task. Another key component of data processing is segmenting the signal into smaller more manageable portions, each of which is separately processed.

Feature extraction is the process of representing a signal in terms of features that summarize the patterns well enough for Epileptiform Activity to be differentiable. The task of finding the best feature extraction method and feature set is very important as if the wrong features (or too many features) are selected then the classification task is complicated.

The final part of a detection system is the actual classification component which attempts to "learn" a mapping between an input and the probable presence of Epileptiform Activity. The classification accuracy as well as processing speed of the classifier is directly related to the choice of feature extraction method.

All of these components are explained fully (including a review of the relevant literature) in the methodology section due to their close and critical connection to the creation of the classification model.

2.4.2 Data processing

2.4.2.1 Artifact reduction

The presence of Artifacts in EEG data causes problems with classification models as this activity tends to obscure the presence of Epileptiform Activity. There are three typical ways of dealing with this: avoid the occurrence of Artifacts during measurement; remove all affected data; correction of affected data. As the data has already been captured there is no way to ensure or control the avoidance rate of artifacts during measurement and so this method is not discussed further. Artifact

removal is also not a feasible option as this leads to the discarding of potentially useful data. There are number of methods in the literature of correcting the effect of artifacts: spatial filters (Ille et al. [18]); blind source separation (Jung et al. [19], Joyce et al. [20], Zhou et al. [21]); regression based techinques (Schlogl et al. [22]). A study conducted by Wallstrom et al. [23] compares the effectiveness of PCA, ICA and regression based methods. This study showed that PCA and regression based methods work well with minimal distortion (particularly if used in conjunction) whereas ICA based methods tend to cause a power distortion in the 5-20 Hz spectral band. Regression based methods also have their own problems in that they reduce cortical activity within the data [20]. The ICA method was selected due to its wide, and previously successful, use as an Artifact reduction tool as the effect of the distortion is seen to be minimal.

2.4.3 Feature extraction

2.4.3.1 Raw data

Prior to attempting any further computation to extract relevant feature sets some researchers (Webber et al. [24], Ozdamar et al. [25]) used raw data to test its efficacy in multi-layer perceptron ANNs with adequate results. A more recent analysis performed by Ko and Chung [26] reinvestigated the use of Raw data as input for classification and found that the results from previous research to be incorrect as its performance was similar to that of random classification.

2.4.3.2 Temporal information

Epileptiform Activity has very distinctive characteristics that EEG experts can use for manual identification. Some attributes that can be used as measures of identification can therefore be found in the signal's morphology which include: sharpness, amplitude, duration and waveform convexity. An example of this is the work done by Gotman and Gloor [27] where they split the waveform into half-waves. Parameters are calculated for each of these half-waves to measure the shape, duration and sharpness of each. These parameters are used as to represent the waveform and thresholds are calibrated based on expert knowledge of the characteristics of Epiletiform waveforms. The results of this method have been shown to work well on the training set.

2.4.3.3 Fourier Transform

The Fourier Transform is a statistical method used to calculate the spectral distribution of a given waveform's power. This extraction method is potentially very good for identification of Epileptiform Activity as brain wave patterns are typically characterized by their frequency. A common and computationally efficient method of implementing this transform is the Fast Fourier Transform (FFT).

Polat et al. [28] utilised the Welch Method of FFT on the EEG data which essentially calculates the power spectrum over a windowed segment of data. This method achieved classification accuracies of between 96-99%. Jando et al. [29] also applied the FFT to segments of the EEG signal and separated the real and imaginary components to use as inputs for classification. This classifier was reported to have a classification accuracy of between 93%-99% in rats. The problem with these results is that only high voltage spike and wave patterns were detected, not all Epileptiform Activity was catered for. These results indicate that using the FFT on the signal can provide a highly effective feature set.

2.4.3.4 Wavelets

The wavelet transform is another statistical method used to calculate the spectral distribution of a waveform. The key properties of wavelets are translation and dilation which mean that the window is not fixed and multiple resolutions are calculated. Another advantage of this method is that it does not require stationarity from the input signal making it suitable for EEG applications.

The most common form of this technique is the Discrete Wavelet Transform (DWT) as it is far more computationally efficient than the Continuous Wavelet Transform (CWT). The DWT has been applied to the classification of Epileptiform Activity by Kalayci et al. [30] and Ubeyli [31]. These two studies used the wavelet coefficients as inputs for classification and achieved accuracies of 86%-92% and 94.83% respectively.

2.4.4 Classification methods

Once suitable features are derived from the EEG, they are presented to a classifier for classification as containing epileptic or non-epileptic activity. Different types of classifiers have been used in precious studies.

2.4.4.1 Classifier types

Expert based systems that require a set of rules in order to function are prone to high false positive rates. Defining these rules can also be complicated as EEG experts do not agree on what can be classified as Epileptiform Activity within the data.

AI methods have been extensively used to solve this problem and most of these report classification accuracies of over 90%. AI methods "learn" the underlying rules from the training data and so do not require any prior knowledge on what are the characteristics of Epileptiform Activity. Some examples of this are: ANN based detection [4,32], decision tree detection [28], SVM based detection [33]. A lot of these studies only focus on detecting epileptic spikes and ignore other types of Epileptiform activity that are less pronounced. Some problems with these solutions are that they were developed for single channel EEG systems and performance is limited by inter-patient waveform variability. This study attempts to find a classifier to differentiate all types of such activity from other brain wave patterns as well as analyse all channels of EEG data.

Chapter 3 Methodology Overview

This chapter describes the models and methods used in order to create the SVM based Epileptiform Activity Classifier that is the subject of this work. This chapter shows how the basic building blocks are used; the next few chapters will expand on each core component.

3.1 Components Required for a Detection System

As stated in the Literature review, there are three components necessary in order to create any automatic detection system. These are:

- Data processing in order to get the data into a uniform format;
- Extraction Methods in order to obtain a feature set that adequately defines the Epileptiform pattern;
- An automatic classifier (SVM in this work) in order to "learn" the Epileptiform Pattern.

3.2 Possible Interpretation of the Epileptiform Pattern

Each of the three components mentioned above was used in the SVM classifier as a phase in the classification process. The Epileptiform pattern can be interpreted as a series of isolated events operating in each channel of data. The pattern can be measured with each channel in isolation as there must be some forms of visual identifiers that an EEG expert uses for labelling purposes. In this model segments of data for every channel are analysed and a determination is made on each. One of the problems with this model is that Artifacts and other brain activity could disturb or obscure Epileptiform events with similar characteristics. A component of this model would, therefore, be to remove the presence of Artifact Activity in the source data.

3.3 Common Processes

Any approach requires a certain amount of common processing as it would utilise the same data source and would be input into the same classification algorithm (SVM). These common processing steps are required by all approaches and involve the selection of data and its conversion into a usable format. This can be broken down into: Selection, Pre-processing, segmentation and reduction.

There are approximately 20 channels of data in the EEG that measure signals from different areas of the scalp. Each of these channels can be seen as either an independent representation of the electrical activity in that position of the brain or, alternatively, as a combination of activity from various parts of the brain that have propagated from their origins. There are originally 30 EEG channels but this is reduced to 20 due to the subtraction of common reference channels to create the bipolar channels. Each record of data consists of 30 seconds of recorded EEG data. It is not, however, possible to use all of this as the first 10 seconds of data is not labelled and so cannot be used to train the classification system. The first 10 seconds was, therefore, discarded.

The classification algorithm only recognizes two states, 1 and -1, which correspond to the different states of classification. In this example a label of 1 would signify Epileptiform Activity and -1 non-epileptic or normal activity. These two states are calculated by the classifier via the use of a signum function around the classifying hyperplane where cases above the plane are equal to 1 and those below are equal to -1. The EEG experts labelled the data into two binary states: 1 and 0 which are not compatible with the classifier. Additionally, the labelling process identifies Epileptiform Activity

in the last 10 seconds of each segment with higher resolution and specifies the centre of this activity. The centre of this activity is given a different label (127) which is reduced back to the normal epileptic label for the normal classification process. The binary states and labels need to be converted to ones recognizable by the classification algorithm.

In order to isolate each Epileptiform event from standard activity and to calculate localised characteristics it is important to implement a segmentation strategy. This segmentation process windows the data channels into more manageable and focussed sets of data.

It is impractical to use all of the data to train the system both because the classification algorithm cannot handle large amounts of data and some of the data needs to be left for validation. It is therefore necessary to select only a sample of the data that best represents the pattern. In accordance with this a record of data was taken from each of the 20 total patients and used in the training process.

3.4 Classification Method

The classification method utilised to implement the approach mentioned above was to segment each channel of data into windows from which features were then extracted.

This method uses the same structure as shown in the **Figure 7** below and the process is as follows:

1. The data is first loaded from binary files and then processing is done on it to convert it to a useable format.
2. The inputs for classification are created.
3. The classifier is trained then tested and the performance results are recorded.

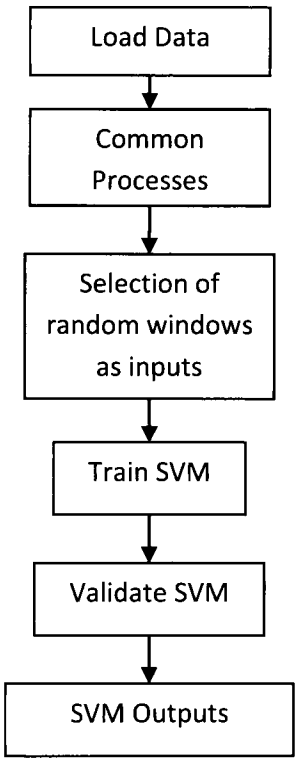

Figure 7: Overview of classification system

The validation process is identical to that of the training process until the data is separated prior to the creation of the SVM classifier. The training and validation sets are both tested against the same classifier in order to test the classification performance on the training data (i.e. how well the system has learnt the current pattern) and how well the system deals with unseen data.

The next sections detail the chosen method and then analyses its validity and shortfalls in terms of detecting Epileptiform Activity.

3.4.1 Windowed Feature Sets

This method can be divided into two parts which are nearly independent of each other. The first part is the creation of the input vectors and the second is the classification process which makes use of the input vectors. The sequence of events followed when creating the input vector is shown in the figure below. This method looks at windows of data in each segment randomly.

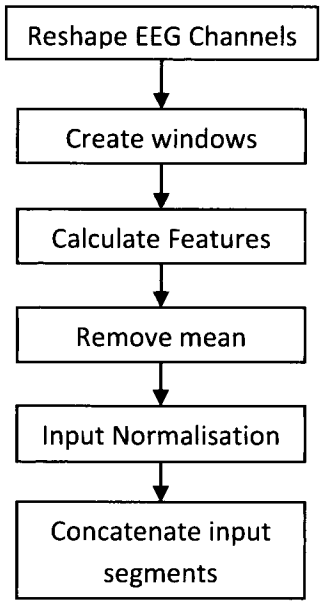

Figure 8: Windowed feature sets process

The windows of data are created using different techniques dependent on the specific feature extraction method used. These techniques determine the shape of the window. These windows are then stored in separate rows and can be seen as the initial input vectors prior to any calculations or data transformations.

The features of each window are calculated and stored as the input vectors. Data processing is then preformed on these vectors so that they can be input into the SVM classifier and so that they represent the Epilpetiform Activity more clearly. The mean value is first removed from the data and then the inputs are normalized between 1 and -1 which are the limits of the classification tool. It was seen that performing this normalization across a whole record instead of normalizing each window individually yielded better results. The last step is to concatenate all of these different records into one data set.

The SVM classification process is fairly simple in that a subset of the inputs created in the above process is selected, balancing the number of epileptic and non-epileptic cases. This is done as Epileptiform Activity occurs far less often than normal activity and so the classification could be skewed towards this. Validation is then performed to determine the optimal parameters for classification.

3.5 Components of Methods

The model used the three standard components (data processing, feature extraction and SVM classification) as fundamental building blocks. These components can be seen as black boxes by the overall classification system. The components don't need to know what happens outside of them, data is input to them which is processed and transformed into outputs.

3.5.1 Data Processing

This component includes all of the techniques used to manipulate and select the data so that it is in the correct format and so that the effect of the Epileptiform Activity is more pronounced. It also includes the Artifact extraction processes used to decrease the effect of Artifacts on the classification algorithm. This component is used in all parts of each system and so is not shown in **Figure 7** and **Figure 8** depicting the processes.

3.5.2 Feature Extraction

This component is made up of the extraction of the feature sets that summarise the information needed by the classifier in order to determine Epileptiform Activity. The dimension reduction of these feature vectors is also discussed. This component is used just before the classification.

3.5.3 Support Vector Machine Classification

This component consists of the choice of parameters and architectures used in the creation of the classifier. This component is where the actual classification occurs; only results and output processing are done after this. Classification can include more than one stage dependent on the method used.

Chapter 4 Data Processing

This chapter describes all of the processing steps used to clean and transform the data into a usable format so that the Epileptiform Pattern is more distinguishable. This includes: data selection, data pre-processing, data segmentation, data reduction and Artifact reduction.

The data used in this research was acquired from the Pattern Recognition Group of the University of Pretoria. The data consisted of multiple 30-second records of EEG recordings taken from 11 patients. These recordings were not taken while the patients were experiencing any seizure activity and hence the classification system will be trained to detect Inter-ictal Epileptiform Activity. There are 2 to 6 data records from each patient containing labeled Artifact, Epileptiform and Normal Activity. The first 10 seconds of each segment was not labeled and was therefore discarded from the analysis. The labels and the diversity of the data provide sufficient data in order to train a generalized detection system.

4.1 Data Selection

The data used in the classification algorithm was selected using records from all of the patients. Data records were randomly selected from the total pool of 11 patients to form input and target data sets.

4.2 Data Pre-processing

The pre-processing methods utilized were: data normalization, input vector whitening, unit standard deviation and data concatenation. The data was normalized between -1 and 1 which are the acceptable values for the SVM classifier signum function. The data was whitened by removing any non-zero correlations. A random vector is defined to be "white" if its elements are uncorrelated (identity covariance matrix) and have unit variance. This property corresponds to a flat power spectrum and is useful for data compression purposes. This whitening step is usually done as a preprocessing step to ICA. The other pre-processing methods were used to ensure that the data was standardized across channels and patients.

4.3 Data Segmentation

The EEG is non-stationary by nature which is a problem as most signal processing methods require data stationarity (e.g. FFT). A signal is said to be stationary if statistical properties such as mean and variance are time-invariant. A solution to the problem of non-stationarity is to take smaller portions of the signal through segmentation methods thereby ensuring stationarity on a micro rather than macro level. One segmentation method used frequently in signal processing is windowing functions.

In signal processing, a windowing function is one that is zero-valued or rapidly decaying outside of a given interval. The simplest form of the window function is rectangular in shape and is defined as constant-valued over a given interval and zero valued outside of this interval. When another function or a signal is multiplied by this window, the product is also zero-valued outside the interval and unchanged otherwise. This effectively reduces the signal to the dimensions of the window [29].

The problem with using rectangular windows, or any window for that matter, is that they affect the frequency spectrum of the signal they are applied to. Multiplication of a signal by the window function causes its Fourier transform to have non-zero values (commonly called leakage) at frequencies other than the given frequency. This effect tends to be more pronounced at frequencies

around the one under analysis and decreases with spectral distance. If the signal contains two sinusoids, with different frequencies, leakage can make it difficult to distinguish them spectrally. If the signal amplitudes are also dissimilar, then the leakage from the larger sinusoid can mask the spectral contribution of the smaller one. Leakage also causes a problem in signals with similar frequencies as they become indistinguishable. In order to deal with this functions that dampen the frequencies at the edges of the windows through rapidly decreasing sidelobes and emphasize those in the centre are used [31].

Although the rectangular window has excellent resolution characteristics for signals of similar strength it has weak resolution for signals with differing amplitudes and is subject to high leakage. This characteristic of is known as low dynamic range. The rectangular window also experiences leakage due to its high sidelobes. This work therefore focuses on using both the rectangular and Gaussian window functions whose characteristics are shown in the figures below. The frequency domain feature extraction methods (STFT and DWT) make use of the Gaussian window function to decrease the effect of leakage whereas the temporal methods do not utilize the frequency domain and hence can use a simple rectangular window. The label data was also windowed with any presence of Epileptiform Activity marking the window as epileptic. The minimum size of the window used was 200ms (40 samples) as this was the maximum resolution of the labeling process. The size of the window was selected based on analysis of the average size of epileptic activity which was approximately 50 samples [31].

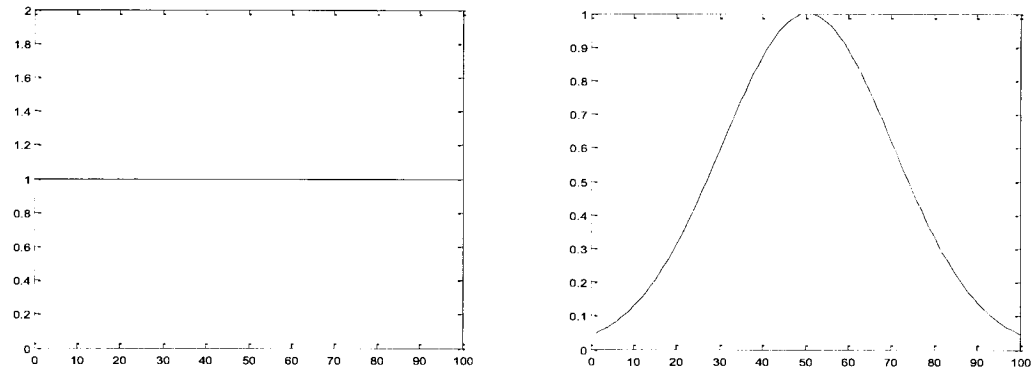

Figure 9: Rectangular and gaussian windows

Windowing was performed on the data using a sliding scale. This means that the starting point of the next window was not the end point of the last but rather a shifted amount from the start of the previous one. This compensated for the fact that the edges of a Gaussian window are tapered and so only the centre of the window is properly measured.

4.4 Data Reduction

The quantity of available data that could be fed into the classification algorithm would be too computationally expensive to process. In order to reduce the computational time of the classification process a reduction technique was implemented. The method adopted in this study was to calculate the average level of activity across all the data channels and identify which channels

consistently contained minimal Epileptiform Events. These channels were then removed from analysis.

Another reduction technique was to identify all the windows of data that contained a mixture of Epileptiform and normal activity. These windows were removed from analysis as they might confuse the classifier on the types of patterns to be identified.

4.5 Artifact Extraction

Artifacts cause many problems with analysis of EEG data as they can have large amplitudes which obscure the presence of Epileptiform Activity. The most common (and easiest) practice in dealing with Artifacts is rejection or removal of any contaminated data. This approach is not feasible as a lot of potentially useful data is discarded and the total signal characteristics may not be represented in the reduced data set. Another problem with this approach is that some of the Artifacts overlap with epileptic activity and so their presence is not easily detectable. This section outlines an automated method to remove the effects of the Artifacts with minimal loss to the other signals in the data.

4.5.1 Blind Source Separation

Blind source separation (BSS) is a statistical tool used to recover individual signals from a linear mixture without knowledge of the characteristics of these source signals (components) or the mixture. There has been extensive research into this field and there are many different methods of solving the BSS problem. This work will only focus on Independent Component Analysis (ICA) as it has been used extensively in the biomedical field, particularly with Artifact Reduction in EEG data. Two other methods are mentioned and briefly discussed to provide some examples of other solutions to this problem [32].

4.5.1.1 Independent Component Analysis

Independent component analysis (ICA) is a statistical modeling technique used to reveal hidden underlying factors from sets of random variables, measurements, or signals. The ICA model can be defined in terms of a linear mixture of signals with the following matrix equation:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{42}$$

Where **x** is the vector of measured or observable signals and **s** is the vector of the original signals or independent components (ICs) that have been mixed by a matrix **A**. The unknown variables of the above model are the original component signals and the mixing matrix. The individual components are assumed to be both non-Gaussian and mutually independent in order for this modeling technique to work. This mutual independence can be formalized to statistical independence. Statistical Independence between two signals is achieved if having information on or knowing the probability of one signal provides no information on the other [32].

The distributions of the ICs are unknown other than the fact that they cannot be normally distributed as ICA would then be impossible to perform. This is due to the fact that the joint density function of two normally distributed signals is symmetrical and hence provides no information on the directions of the columns of the mixing matrix [32].

There are two major limitations with the use of the ICA algorithm: The IC variances are indeterminate, the order of ICs is indeterminate. The variances of the ICs are indeterminate as both

A and **s** are unknown and changing the variance of **s** can be compensated for by changing the corresponding column of **A**. The order of the ICs is indeterminate due to the fact that the model cannot calculate the relative importance of the ICs within the observed signal [32].

The other limitations of the ICA model are: permutation and scaling; stationary mixing; assumption of independence; under- or over-determined representations and relevancy determination; ICs have low signal to noise ratios [32].

As independence is a necessary assumption for the ICA algorithm to function it is necessary to find out how to measure or determine this. The most common approach is to assume that the distributions of the ICs are as far from Gaussian as possible. An IC can therefore be estimated as a linear combination **y** of [32]:

$$f(t)=\sum_{j,k} c_{j,k}\psi_{j,k}(t) \tag{43}$$

in the data vector:

$$\begin{aligned} c_0 &= \frac{1}{4\sqrt{2}}\left(1+\sqrt{3}\right), \\ c_1 &= \frac{1}{4\sqrt{2}}\left(3+\sqrt{3}\right), \\ c_2 &= \frac{1}{4\sqrt{2}}\left(3-\sqrt{3}\right), \\ c_3 &= \frac{1}{4\sqrt{2}}\left(1-\sqrt{3}\right) \end{aligned} \tag{44}$$

as shown:

$$y=\omega^T x=\sum_i \omega_i x_i \tag{45}$$

where ω is the vector to be determined. This estimation can be seen as equivalent to one of the ICs with ω as one of the rows of the inverse of the mixing matrix.

The value of ω can be manipulated in order to ensure that the distribution of the estimated IC is as far from Gaussian as possible. The requirement that the ICs be uncorrelated makes it possible to narrow the search to only look at those areas in the optimization space whose estimates are uncorrelated with each other. In order to select the correct value for ω there needs to be a method of measuring how Gaussian a variable is. The table below shows some of the different measures used in the literature to measure non-gaussianity [32].

Measure	Method
Kurtosis	This is the fourth order cumulant and is only zero for Gaussian variables
Negentropy	This is a measure of signal structure in relation to Gaussian and is zero valued for Gaussian signals
Mutual Information minimization	This is a measure of dependence between variables, similar to negentropy
Maximum likelihood estimation	This is equivalent to the minimization of mutual information

Table 2: Measures of non-gaussianity [32]

After estimating the matrix **A**, its inverse (**W**) can then be computed from which the ICs can be obtained using the following relation [32]:

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x} \tag{46}$$

4.5.1.2 FastICA Algorithm

The FastICA Algorithm is an efficient and statistically sound method of implementing the ICA model. It attempts to separate the mixed signals by maximizing the mutual information or entropy between them. This algorithm can be explained by looking at the single "unit" or component example and then showing how this can be extended to more complex problems. FastICA attempts to find a weight vector ω such that the projection $\omega^T x$ maximizes non-gaussianity, which is the same as the estimation explained previously. The measure of non-gaussianity used is negentropy which is approximated using the following equation:

$$J(\omega^T x) \propto \left[E\{G(\omega^T x)\} - E\{G(v)\}\right]^2 \tag{47}$$

where E is the expected value of a function and *v* is a fixed Gaussian variable. Choosing a correct value for G provides an accurate approximation. The following two functions have been seen to be a good starting point in the selection of this parameter [33].

$$G_1(u) = \frac{1}{a_1}\log(\cosh(a_1 u))$$
$$G_2(u) = -\exp(-u^2/2) \tag{48}$$

where a_1 is commonly set to 1. The maximization of the negentropy is achieved by using an approximate Newton Iteration which uses derivatives of G (denoted as g) as follows:

1. Randomly select an initial weight vector ω
2. $\omega^+ = E\{xg(\omega^T x)\} - E\{g'(\omega^T v old x)\}\omega$
3. The new weight vector now becomes $\omega^+ / \|\omega^+\|$
4. Repeat the previous two steps until convergence

Convergence occurs when the previous and current values of ω have a dot product approximately equal to unity indicating that the vectors point in the same direction.

This estimation is just finds one of the ICs and needs to be extended to more components in order to be of any use. This extension is done by using several units with weight vectors b . A further step is

to decorrelate the calculated projections after every iteration so that the vectors do not converge to the same maxima [33].

4.5.1.3 Other Methods

There are different methods of performing BSS other than ICA. The following two methods are alternatives to the FastICA algorithm but were not used as they were seen to be more computationally expensive. Principal Component Analysis (PCA) reduces the number of dimensions in the data, removes eigenvalues and eigenvectors of the covariance matrix that are below a certain threshold. The highest eigenvectors are the signals that contribute the most to the overall power of the initial signal [34]. Temporal Decorrelation Separation (TDSEP) is based on several time delayed second order correlation matrices and consists of steps: whitening and approximate diagonalisation of matrices [35].

4.5.2 Application to the Electroencephalogram

The following assumptions about the ICs need to be met in order for optimal use of the ICA algorithm:

- The sources or hidden components are independent,
- The propagation delays of the brain are negligible,
- The number of Independent signal sources is no greater than the number of sensors or electrodes

The first assumption can be satisfied if we assume that EEG data can be partially modeled as a mixture of statistically independent brain signals. The second assumption is satisfied as signal conduction in brain tissue is almost instantaneous. The third assumption is difficult to satisfy as the number of signals contributing to the EEG readout is unknown. The main problem with the use of ICA in EEG then becomes that of determining how many input channels are required and also the mapping of discovered ICs to specific brain processes [36].

The ICA model can be applied to the EEG readout by following the following steps:

- Making the EEG signals measured at each electrode (channel) equal the input matrix x
- Modeling the ICs as the underlying brain processes
- Modeling the columns of the inverse matrix W^{-1} or mixing matrix as the propagated strengths of the ICs onto the scalp as measured by the electrodes

After the data has been broken down into ICs, the next step is to search for unwanted components and remove these from the data. Removal of these unwanted signals is achieved by setting the values of the components that match the search criteria to zero and then recalculating the input matrix without their effect. It is not possible to remove all of the interfering data, the purpose of this step is to reduce their effect so that the epileptic Epileptiform Activity is more noticeable.

This work utilises ICA decomposition for the following two uses:

1. Artifact Identification and Removal
2. Possible Epilpetiform Activity Identification

4.5.2.1 Artifact Extraction

As can be seen in the diagram below a large portion of the EEG data is contaminated by artifact activity (highlighted in blue) and it is therefore necessary to separate its effect without impacting the underlying signal.

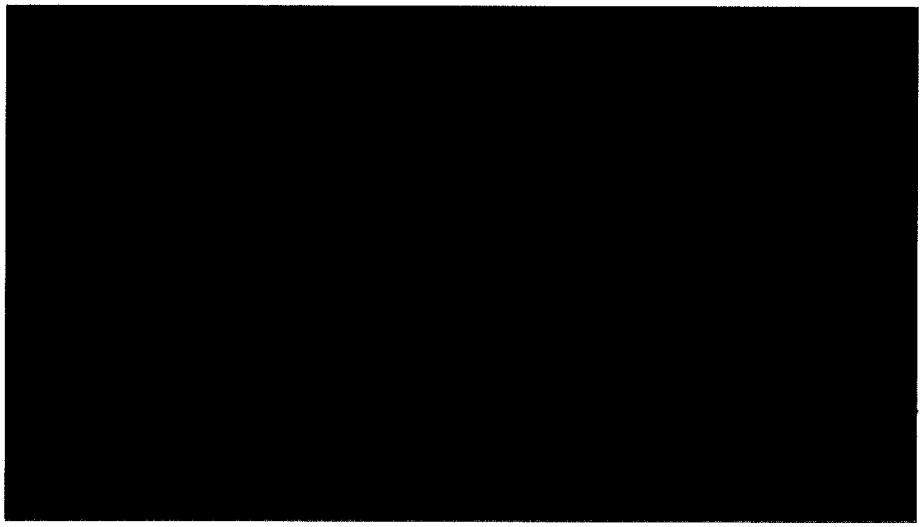

Figure 10: Example of Artifact activity in EEG data

The Artifacts selected for removal were Ocular (EOG) as these are the most common as well as most noticeable type. Muscular (EMG) Artifacts are also common but identification usually requires visual analysis for removal. This visual analysis is through the use of topographic scalp maps to identify their possible biological origin. EMG Artifacts are associated with activity originating from the frontalis and temporalis lobes (concentrated near the ears). This visual analysis is not feasible in an automated system.

The automatic identification of the ocular Artifacts was achieved by comparing some calculated parameters of the components which were generally indicative of EOG activity to certain thresholds. These parameters were:

The Ratio between peak amplitude and signal variance:

$$ratio = \frac{IC_{peak}}{\sigma^2} \tag{49}$$

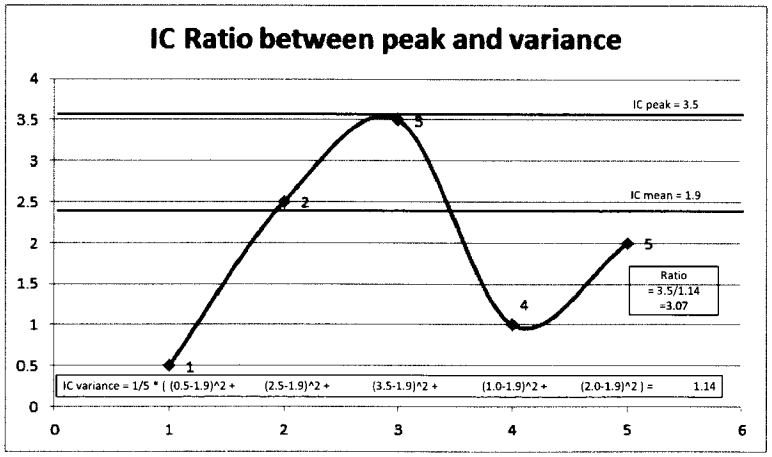

Figure 11: Calculation of IC ratio between signal peak and variance

Statistical Skewness of the IC (S)

$$skew = \frac{IC_{third\ moment}}{(IC_{variance})^{3/2}} = \frac{\mu_3}{(\sigma^2)^{3/2}} \tag{50}$$

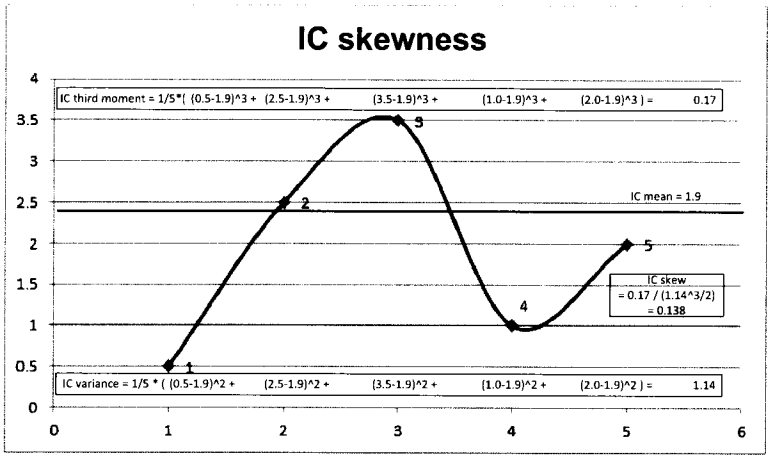

Figure 12: Calculation of signal skewness

Kurtosis of the IC (K)

$$kurtosis = \frac{IC_{fourth\ moment}}{(IC_{variance})^2} - 3 = \frac{\mu_4}{(\sigma^2)^2} - 3 \tag{51}$$

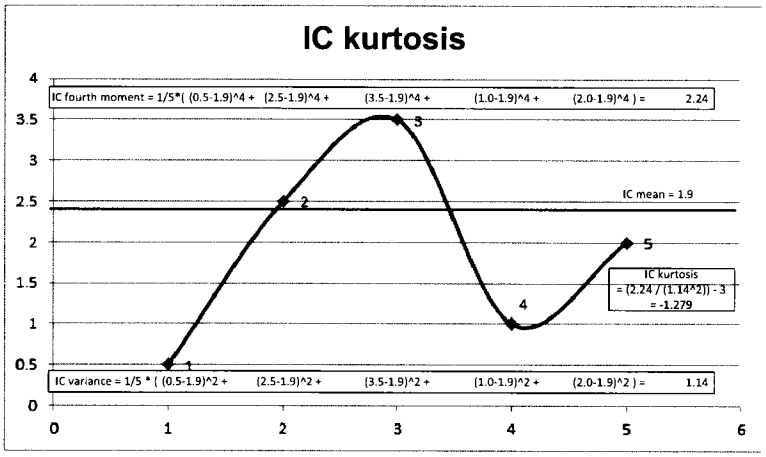

Figure 13: Calculation of signal kurtosis

The values for these thresholds were selected through experimentation. The components that were suspected of being related to Artifact Activity were discarded and the mixed signal recalculated.

4.5.2.2 Epilepsy Extraction

This method attempts to identify all potential Epileptiform Activity and remove the effect of any components that are not Epileptic in nature from the mixed signal. All of the ICs were compared to some parameters that broadly describe Epileptiform Activity. Any component that was below a certain threshold (through experimentation) was discarded. The thresholds were set quite conservatively as these parameters vary between patients and also between recordings.

In order to identify ICs that are indicative of Inter-ictal Activity a "spikyness index" I is introduced which is applied to each of the components as follows:

$$I_i = \frac{max|s_i(t)|}{1/T\int s_i(t)dt} \tag{52}$$

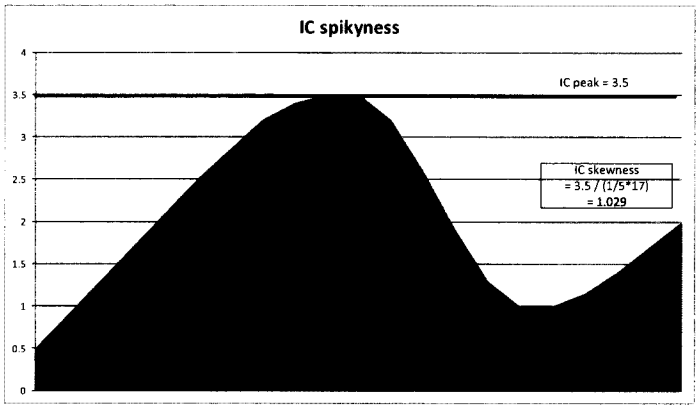

Figure 14: Calculation of signal spikyness

where the maximum is calculated over the entire observation interval T. This index does not just serve as an indicator for spike activity as any component with a strong maximum-to-absolute average ratio will produce large values.

The same measures as used for the Artifact extraction process are also looked at but with lower thresholds and look at all possible Artifact Activity as also possibly Epileptiform.

Chapter 5 Feature Extraction

Feature extraction is an important component of any classification system as it extracts information relevant to the Epileptiform Pattern. Although the raw data contains more information than that of the extracted feature sets, it also contains a lot of redundancy and so the relevant information can be obscured. The extraction of relevant features allows the classifier to more easily identify the desired pattern. This chapter describes the different feature extraction methods used for the detection of Epileptiform Activity as well as the process used to determine the optimal feature set.

5.1 Methods

The creation of a classifying hyperplane consists of learning the mapping between an input vector and a determination of whether or not it is indicative of Epileptiform Activity. Large dimensionality in the input space makes it difficult for the classification algorithm to find the optimal separating hyperplane as the data inter-relationships become more complex. The presence of high dimensionality can be compensated for by increasing the quantity of data used in the classifier training process or by reducing the dimensions of the input vectors. Increasing the number of training cases creates computational problems in terms of processing required to identify the separating hyperplane and becomes unmanageable with high volumes. Reduction of the dimensionality is a more useful approach as it reduces the amount of data to be processed as well as highlights the Epileptiform Activity. This reduction in size was achieved by calculating a feature set that effectively summarized the input vector into a lower dimensional one. This work looks at both the temporal, spectral and phase space aspects of the inputs as they provide useful information on the Epileptiform Pattern.

5.1.1 Electroencephalogram Features in the Time Domain

The time domain of the EEG is the observable signal and is used by EEG experts for analysis. This indicates that the signal contains sufficient information for the classification of Epileptiform Activity. Spatio-temporal attributes were calculated to effectively summarise the Epileptiform Activity.

5.1.1.1 Spatio - Temporal Parameters

Spatio-Temporal parameters were taken from sliding windows of data 300ms in duration. Parameters were selected that were indicative of Epileptiform Activity and were concatenated to form input vectors for the classification algorithm. Each channel of EEG data was windowed sequentially and the calculated input vectors were appended to the end of the previous channel, creating a continuous stream of vectors.

The parameters calculated were:

- Mean value of window (and relative to channel)
- Average sharpness of window (and relative to channel)
- Standard Deviation of window (and relative to channel)

The mean or average of the window provides information on what the average amplitude of the window is which can be useful as epileptic spikes can occur for a long period (up to 200ms) and so would cover most of the window. The sharpness of the window tells us whether the data in the window is changing rapidly. Standard Deviation measures how far the signal fluctuates from the mean with respect to power. This is useful as it provides an indication regarding the activity.

5.1.2 Electroencephalogram Features in the Frequency Domain - Fourier Transform

The spectral domain is a useful space for analysis as brain signals are characterised by their frequencies. The Fourier Transform (FT) is a tool that provides the means of transforming a signal defined in the time domain into its spectral or frequency domain equivalent. The FT is used to transform a continuous time signal into its equivalent frequency domain representation which describes the continuous spectrum of a non-periodic time signal. The FT X*(f)* of a continuous function *x(t)* in the time domain can be expressed as [37]:

$$X(f)=\int_{-\infty}^{\infty} x(t)e^{-i2\pi ft}dt \tag{53}$$

This form of the transform needs to be adjusted as the EEG data has been converted into a discretised format when input into the digital measurement system. The adjusted form is the Digital Fourier Transform (DFT) which calculates the FT of discrete signals as follows [37]:

$$X(mF)=\sum_{n} x(nT)e^{-inm2\pi nm} \tag{54}$$

5.1.2.1 Fast Fourier Transform

The DFT formula can be implemented using the Fast Fourier Transform (FFT) class of algorithms in a computationally efficient manner. FFT algorithms use the same transformation equation as the DFT, the difference being in the reduction of the required computational steps. The FFT reduces the steps required for N points from $2N^2$ to $2N\ln(N)$, whose difference can be seen in the figure below. Aliasing (leakage) problems due to sampling can be reduced by apodization using a tapering function but this comes at the expense of broadening the spectral response. This method was therefore not considered and the effects of aliasing were instead minimized by filtering out high frequency components prior to signal transformation.

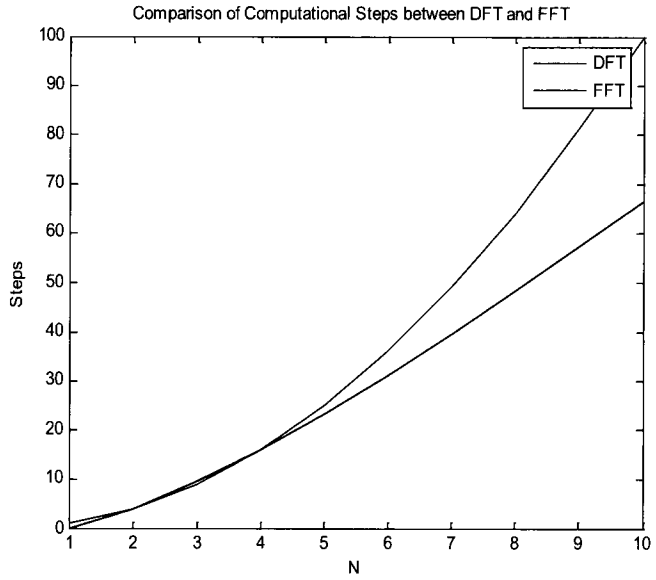

Figure 15: Comparison of computational steps between DFT and FFT

The reduction in computational steps in the FFT algorithms is achieved using the following two approaches: decimation in time and decimation in frequency.

There are two main approaches that FFT algorithms take: decimation in time, and decimation in frequency. The one used in this work is the Cooley-Tukey FFT algorithm which uses the decimation in time approach by first rearranging the input elements in bit-reversed order, then building the output transform. This is done by breaking up a transform of length N into two transforms of length N/2 using the identity (Danielson-Lanczos lemma):

$$\sum_{n=0}^{N-1} a_n e^{-2\pi i n k / N} = \sum_{n=0}^{N/2-1} a_{2n} e^{-2\pi i (2n) k / N} + \sum_{n=0}^{N/2-1} a_{2n+1} e^{-2\pi i (2n+1) k / N}$$

$$= \sum_{n=0}^{N/2-1} a_n^{even} e^{-2\pi i n k / (N/2)} + e^{-2\pi i k / N} \sum_{n=0}^{N/2-1} a_n^{odd} e^{-2\pi i n k / (N/2)} \quad (55)$$

The problem with frequency analysis is that it is primarily focused on activity that is present throughout the time period such as background activity and treats transient events as if they occurred over the whole period. This means that the Epileptiform Activity is scattered or lost in the frequency spectrum and so there is no way of knowing when a specific event occurred or how many events there are. A further problem is the non-stationarity of EEG signals which is a prerequisite for the transform. In order to get around these problems the signal is segmented into windows consisting of shorter time periods using a method called Short-Time Fourier Transform (STFT). This segmentation ensures that events can be analysed independently of each other and the temporal location of the events can be identified to a specific window [37].

The FFT produces a signal that contains data with elements in both the real and imaginary spaces. The absolute value of the FFT coefficients was calculated as the classification algorithm does not accept imaginary inputs. The average and first harmonic were taken as input features. Polet and Gunes [40] use Welch's method of spectrum estimation based on FFT estimation of windowed data in order to extract EEG features.

5.1.3 Electroencephalogram Features in the Frequency Domain - Wavelet Transform

The Wavelet Transform (WT) is a different method of analysing the spectral representation of a signal and is useful in that it allows for multi-resolution signal analysis. The advantage of WT over the STFT method described previously is that the window size is not fixed, allowing for better time and frequency resolution. There are two important properties of wavelets: translation and dilation. Translation is the process of shifting the wavelet along the time axis while dilation is the stretching or shrinking of the wavelet [38].

One of the problems associated with using wavelets is that they do not have well defined averages in the frequency domain; this means that they perform poorly in the detection of stationary or time frequency elements. Stretching a wavelet through dilation can also be problematic as this makes it less localized in time but more localized in frequency [38].

The standard form of the WT is the Continuous Wavelet Transform (CWT) and is used on continuous signals. EEG data is in a discretised format and so the Discrete Wavelet Transform (DWT) was used. The CWT also has problems in that it requires a lot of computation and it is inefficient in that the information at similar scales is highly correlated [38].

5.1.3.1 Discrete Wavelet Transform

The DWT is well suited to the analysis of non-stationary signals such as the EEG as this is not a prerequisite as it is for other such transforms. One of the shortcomings of the STFT is that the use of a fixed-length temporal window imposes a restriction on the time and frequency resolution of the attained transform. The DWT, in contrast, analyses the signal at multiple resolutions and high-frequency components have a sharper time resolution than low-frequency components. This is a useful property for EEG analysis as it contains signal components that differ significantly in duration and frequency content [38].

5.1.3.2 Theory and Computation

Only the discrete form of the DWT with orthogonal wavelet functions is described as orthogonal functions and produces a compact transform without redundancy at large scales. The term wavelet is used to describe a function $\psi \in L^2(\Re)$ such that appropriate translations and dilations of ψ form an orthonormal basis for the Hilbert space $L^2(\Re)$. The basic form of a wavelet is defined as [39]:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \quad a,b \in \Re, a \neq 0 \tag{56}$$

where a is the scale parameter responsible for dilation of the wavelet and b controls the position of the wavelet in time. In order to form an orthonormal basis in $L^2(\Re)$ suitable values for these parameters have to be chosen. A common choice is to select discrete values for *a* and *b* such that

$$\begin{aligned} a_j &= 2^j \\ b_{j,k} &= 2^j k \end{aligned} \tag{57}$$

where $j,k \in Z$ and j is referred to as the resolution level. Substitution of these values into the Wavelet Function yields the following form:

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right) \quad j,k \in Z \tag{58}$$

A linear combination of this function can be used to represent any finite-energy function *f*(t) as:

$$f(t) = \sum_{j,k} c_{j,k}\psi_{j,k}(t) \tag{59}$$

The wavelet coefficients $c_{j,k}$ provide a time-scale representation of the signal *f*(t). Scales are related to frequency and so this representation can be seen as similar to a time-frequency representation of the signal except with a variable time window. The multi-resolution behaviour of wavelets is [38]:

- Low scale (high frequency activity) is analysed with a fine time resolution and
- High scale (low frequency) activity is analysed with a coarse time resolution.

This difference in resolution for varying frequencies is due to the fact that high frequency activity usually occurs over shorter periods of time while low frequency activity is more spread out [38].

The wavelet function $\psi_{j,k}(t)$ that is used has to comply with certain mathematical requirements such as admissibility and regularity. Admissibility requires that the wavelet be a zero-mean, square integrable function. Regularity relates to the concept of vanishing moments which give the wavelet an oscillatory shape with a fast decay to zero [38].

There are many different types or families of wavelets which comply with these mathematical requirements. The family used in this study was the one proposed by Daubechies as they are a close approximation of the shape of epileptic spikes making them a good choice for feature extraction. In addition, the wavelet has good localization properties in both the time and frequency domains and has been used for EEG feature extraction in previous solutions. The fourth order Daubechies wavelet was found to be the most useful as the higher order wavelets were too sharp while the lower orders were too coarse for the analysis of Epileptiform Activity.

Daubechies wavelets are an extension of the general wavelet function in that they define the scaling parameter in terms of a function, ϕ, which is used to derive ψ as follows [38]:

$$\varphi(t) = \sqrt{2}\sum_{k=0}^{L-1} c_k \varphi(2t - k)$$
$$\psi(t) = \sqrt{2}\sum_{k=0}^{L-1} (-1)^k c_{1-k} \varphi(2t - k) \tag{60}$$

where L is related to the number of vanishing moments of the wavelet. The fourth order Daubechie (Daub-4) has a value of L = 4. The coefficients c_k for this wavelet are:

$$c_0 = \frac{1}{4\sqrt{2}}\left(1+\sqrt{3}\right),$$
$$c_1 = \frac{1}{4\sqrt{2}}\left(3+\sqrt{3}\right),$$
$$c_2 = \frac{1}{4\sqrt{2}}\left(3-\sqrt{3}\right),$$
$$c_3 = \frac{1}{4\sqrt{2}}\left(1-\sqrt{3}\right) \tag{61}$$

The figure below shows the scaling and wavelet function of a 4th order Daubechies wavelet.

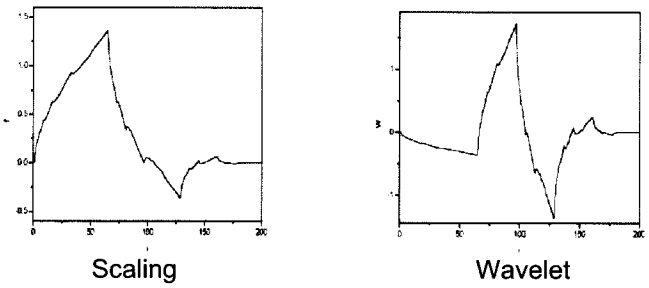

Figure 16: Scaling and Wavelet Functions for 4th Order Daubechies Wavelet [36]

The DWT of a discrete signal can be efficiently computed using a filter bank of quadrature mirror filters in what is known as a pyramidal algorithm as shown in the figure below. This process consists of a high pass and low pass filter with impulse response g[n] and h[n] respectively. Both impulse responses are derived from the coefficients of the wavelet family being used. As the filters are approximately halfband, the filtered signals are downsampled by 2 [38].

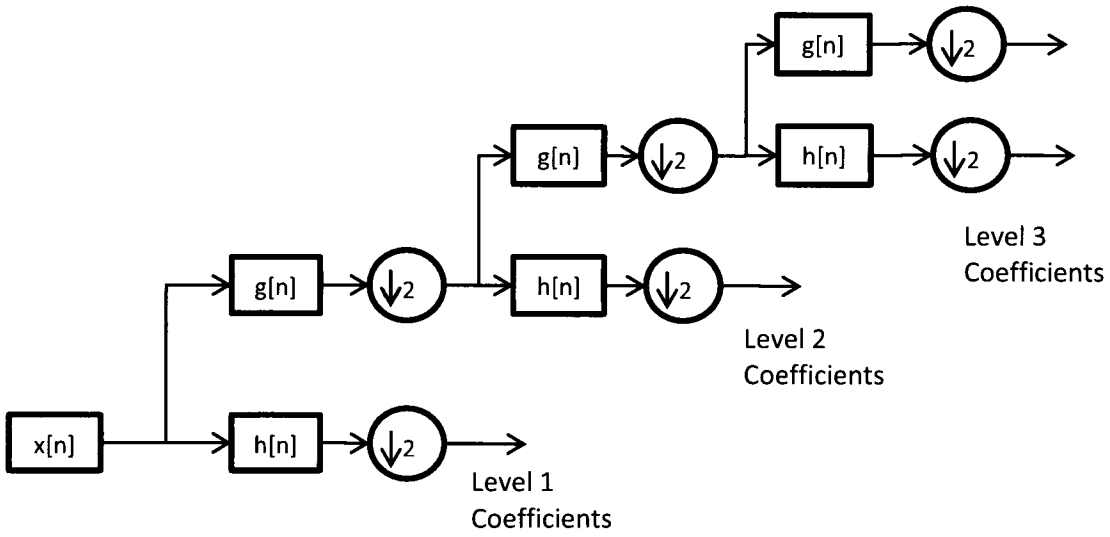

Figure 17: Filter bank used to calculate DWT [38]

The application of the pyramidal algorithm to a discrete signal of N samples is shown in the figure above. The signal is first decomposed into two components: detail and approximation wavelet coefficients of length N/2. These two components can be used to exactly recreate the original signal. The detail coefficients are the wavelet coefficients at resolution level 1 and represent a frequency band of $\pi/2 \rightarrow \pi$ rad/s. The approximation signal is decomposed at the next resolution level into another set of approximation and detail coefficients with frequency bands of $0 \rightarrow \pi/4$ and $\pi/4 \rightarrow \pi/2$ rad/s respectively and length N/4. The process is then repeated for as many resolution levels as are required. The implementation of the Daub-4 wavelets would therefore have to have four levels of filters [38].

5.1.3.3 Feature Extraction

Inter-ictal epileptic spikes are similar in shape and frequency spectrum to impulse functions in the time domain. This means that their effect spreads into a large range of frequencies which need to be analysed. This frequency range is known from previous solutions to match most closely to a resolution level 4 (3.6 – 8.24Hz) with possible components at level 3 (8.24 – 18.96 Hz). These frequency ranges for the resolution levels were calculated using the fact that the data was sampled at 200Hz. The maximum and average values of the level 4 coefficients were therefore used as input vectors to the classification algorithm. Mohamed et al [4] use the Discrete wavelet transform to extract epileptic features prior to classification using an artificial neural network.

5.1.4 Electroencephalogram Features in the Non-Linear Domain - Lyapunov Exponent

Intermittency of seizures are difficult to explain using linear systems as this will only occur in response to an input. In most types of epilepsy external triggers have not been identified. Non-linear systems allow for this type of behaviour [41].

Non-linear systems can be modelled using difference equations where changes over time are functions of one or more variables taken to different powers. One of the properties of these equations is that for some parameters they can behave chaotically. Characteristics of chaotic systems [41]:

- Strong dependence on initial conditions
- Ability to show self-organization. A change from a chaotic signal to an ordered one or vice versa is characterised by abrupt changes in phase

EEG shares many characteristics with that of a non-linear system [41]:

- Limit cycles
- Bursting behaviour
- Jump phenomena
- Amplitude dependent frequencies
- Frequency harmonics

This leads to the prediction that non-linear modelling could provide valuable information on the EEG.

A well-established method of analysing the dynamical behaviour of multi-dimensional systems is through the use of phase space portraits. Each time dependent variable is treated as a component of a vector in the phase space. Each vector represents an instantaneous state of the system. These vectors can be plotted sequentially representing the evolution of the state of the system over time. This can create an object confined over time to a sub-region of the phase space. These sub-regions are called "attractors" and their geometrical properties provide information about the global state of the system [41].

The complexity of an attractor can be determined from its dimension, the larger the more complicated. Lyapunov exponents can be used to measure the average rate of expansion and folding that occurs along different local directions within an attractor in the phase space. They can also be used to measure the chaoticity of the signal [41].

The Lyapunov exponents are calculated from the Jacobians of a dynamic's system differential equation. This can be shown using the following dynamic system [42]:

$$y_{t+1} = f(Y_t) + u_t \tag{62}$$

Where u_t is a random variable, $Y_t = (y_t, y_{t-1}, \dots, y_{t-d+1})$ is the data vector and d is an integer ≥ 1, f is a function of d variables and $t = 1,2, \dots$.

The Jacobian of this system is defined as:

$$J_t = Df(Y_t) = \begin{pmatrix} f_1' & f_2' & \cdots & f_{d-1}' & f_d' \\ 1 & 0 & & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 1 & 0 \end{pmatrix} \tag{63}$$

where f_d represents the partial derivative of the d^{th} variable.

If T_m is defined as:

$$T_m = J_m J_{m-1}, \dots, J_1 = Df^m(Y_1), \qquad m = 1,2,3 \dots, \tag{64}$$

The limit:

$$\lambda_i(Y) = \lim_{m\to\infty} \frac{1}{m} \ln|a_i(m, Y)|, \quad i = 1,2, \dots, d \tag{65}$$

defines the Lyapunov spectrum of the system where $a_i(m, Y)$ is the i^{th} largest eigenvalue of $Df^m(Y)$ [41].

The maximum (or dominant) Lyapunov exponent is the limit:

$$\lambda_1 = \lim_{m\to\infty} \frac{1}{m} \ln|T_m| \tag{66}$$

The estimation of the Jacobian therefore provides sufficient information to calculate the Lyapunov exponents of any time signal. This can be done by using the Taylor expansion of the system as an approximation to the characteristic function. The series can then be used to calculate the Jacobian matrix [42].

Epileptic activity has been seen to cause measurable changes in the chaoticity of a signal before, during and after. The L-max approximation per window was therefore used as an input to the system. Swiderski et al [43] use Lyapunov exponents to detect epileptic waveforms in EEG data.

5.2 Creation of Optimal Feature Set

A lot of features were used that could possibly separate Epileptiform Activity from Normal Activity. A potential problem with this is that some features might actually obscure the differences and therefore decrease the classification accuracy. It is therefore important to identify the set of features that most differentiate the two types of activity.

This work uses Probability Based Incremental Learning (PBIL) to determine which combination features produces the highest degree of differentiation. PBIL is a form of genetic algorithm that attempts to find the optimal solution through an iterative approach. PBIL was preferred over other genetic algorithm techniques due to its increased accuracy and faster processing rates [44].

Genetic algorithms are based on the evolutionary biology principles of natural selection and biology. All living organisms are made up of chromosomes which are made up of genes. These genes are a combination of traits from both parents whereby the weakest ones are eliminated. A genetic algorithm randomly selects a set of chromosomes and calculates the fitness of each based on a predetermined function. These chromosomes are then randomly bred to generate a new pool. Breeding consists of 2 parts [44]:

1. Crossover
2. Mutation

Crossover involves the exchange of genes between two parents whereas mutation is due to randomly changing binary values that server as a representation of the genes. These characteristics add diversity to the algorithm allowing it to settle into an optimal state. The progression of genes from one generation (iteration) to the next is determined by how well they perform against a fitness function [44].

PBIL is an adaptation of this whereby the population is replaced by a probability vector. In the feature selection case the elements of the probability vector define the probability of selection of each feature. The fitness of each trial is calculated using the Separability Index and the trial with the highest index is said to be the best of the generation. The probability vector is then adjusted to reflect the best trial. This means that the best features would increase in probability while the worst would decrease. Mutation is then applied to decrease the chance that a local rather than global optimum is achieved [44].

The fitness function used to determine the best feature set in each generation was Thornton's Separability Index. This calculates the percentage of data points whose classification is the same as its nearest neighbours. A high index indicates that the classes looked at are highly separable and a classification algorithm using those features will therefore perform well. Thornton's Separability Index is calculated using [45]:

$$s = \frac{\sum_{i=1}^{n}(f(x_i)+f(x_i')+1)\mathrm{mod}2}{n} \tag{67}$$

Where x' is the nearest neighbour to x and n is the total number of points.

In this case the components of the vector are made up of the probability that a feature is selected and the separability index is calculated on the entire training set.

5.3 Summary

The feature vector created from all the extraction methods described in this chapter can be seen in **Table 3** below.

Feature
Average sharpness of window
Average sharpness of window relative to channel
Maximum Laypunov exponent of window
Average Laypunov exponent of window
Standard deviation of window
Standard deviation of window relative to channel
Maximum coefficient from DWT analysis
Average of coefficients from DWT analysis
Maximum harmonic of window
Average harmonic of window
Mean value of window
Mean value of window relative to channel

Table 3: List of features extracted from data

Chapter 6 Support Vector Machine Classification

This chapter discusses the steps taken in order to identify the optimal separating hyperplane from the input vectors calculated in the feature extraction stage. The choices with regard to parameters are also discussed and justified.

6.1 Classification Process

The classification process consisted of using the input vector datasets to identify the optimal separating hyperplane for Epileptiform Activity. A SVM toolbox was used as the implementation of the algorithm within a MATLAB environment. The data was split into two sets: training and validation. These sets were made up of data taken from all the patients that was pre-processed and then split into two parts. This split methodology was used in order to test the generalization properties of the separating hyperplane.

The problem with detecting the Epileptiform Activity is that there are far fewer cases of this present in the data in comparison with normal or Artifact Activity. This imbalance in the data has a negative impact on classification as the algorithm will be skewed towards non-Epileptiform Activity. In order for the classifier to better differentiate between the two and to improve its generalization performance it is necessary to balance the number of cases that are input into the classification system. This was done by building the sets through random selection from each Activity Type.

As discussed in **Chapter 2** the SVM classifying hyperplane is created by calculating the Lagrange multipliers derived from the training inputs. These Lagrange multipliers are then used to calculate the classification prediction via multiplication by a Hessian matrix constructed from both the training data and the data being tested [13].

6.2 Input and Target Vectors

The training set used by the classification algorithm consisted of two components: inputs and targets. The inputs to the classification were selected from the outputs of the feature extraction methods discussed in **Chapter 4**. These were:

1. Spatio-Temporal parameters describing the morphology and behaviour of the windowed data.
2. The highest and average harmonic of the windowed frequency response
3. The highest and average coefficients of the 4th order Daubechies 4 DWT wavelet
4. The maximum Lyapunov exponent of the window

The target vectors are the labels that describe the classification of the windows according to EEG expert diagnosis. This is used for performance measurement and to train the classification algorithm to learn which inputs contain Epileptiform Activity. If the window contained Epileptiform Activity then it was given a label of **1** and a label of **-1** otherwise. This labelling system was used in order to align with the outputs from the sigmoid function used by the SVM classification algorithm to calculate the predicted state of the vector. The sigmoid function produces an output of **1** for anything above its positive threshold and **-1** for anything below its negative threshold. Anything in between these thresholds produces an output of **0**.

6.3 Unbalanced Data sets

The imbalance in the data can be compensated for by evening-out the number of Epileptiform and non-Epileptiform cases being input into the classification algorithm and system. This is achieved through under-sampling of non-Epileptiform cases and oversampling of Epileptiform ones. The sampling process was implemented using random selection.

6.5 Hyperparameter Selection

As shown in **Chapter 2** the hyperparameter C is an important choice as it governs the bounds of the Lagrangian multipliers and hence the margin as shown in the equation below [13].

$$\gamma = \left(\sum_{i \in SV} \alpha_i^* - \frac{1}{C} \langle \alpha^* \cdot \alpha^* \rangle \right)^{-1/2} \tag{68}$$

where the vector α contains the Lagrange multipliers. The value of this parameter was selected through experimentation and performance analysis. It was observed that higher values of C produced better results in terms of generalization. A value of 5 was chosen for this parameter.

6.6 Kernel Selection

The kernel function is used on the input vectors in order to map them to a higher dimensional plane in which the classes are more linearly separable. It was necessary to select the most suited from a number of different kernel functions and also to select the dimension of this new space.

The kernel mappings selected were Gaussian RBF and linear whose transform equations are [13]:

$$K(x,x') = \exp\left(\frac{-\|x - x'\|^2}{\sigma^2} \right),$$
$$K(x,x') = x.x' \tag{69}$$

The reasons for this were that the RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle cases where the relation between class labels and attributes is nonlinear. This kernel also has similar performance or is similar to other kernels such as the linear and sigmoid versions. The second reason for using the RBF kernel is the difference in the number of hyperparameters required by each kernel, where more parameters are directly related to greater model complexity. Kernel functions like the polynomial kernel have more hyperparameters than the RBF kernel. The last reason for choosing the RBF kernel was that it has less numerical difficulties than other kernels. The linear kernel was selected due to its low dimensionality as well as high historical performance [46].

In the case of the RBF kernel function the dimension is varied by changing the width or sigma parameter. Selection of a large value sigma produced high specificity but low sensitivity whereas a low value produced high training results but low generalization performance. The value chosen was $\sigma = 1$ which is between these two extremes.

6.7 Unclassifiable Data

The use of soft margin classification is necessary as it is highly probable that Epileptiform and non-Epileptiform Activity are not completely separable, partly due to inaccuracies in the labelling process. The use of hard margin classification in this case would give rise to a large number of misclassified cases and also a great degree of uncertainty on the accuracy of the classification. In using soft margin classification it is possible to identify these unreliable or uncertain data points. They can then be ignored in the output as unclassifiable data. This approach is valid if the quantity of unclassifiable data is small. The classification algorithm would also be better aligned to the more definite cases.

Soft margin classification was implemented by selecting a threshold distance away from the classifying hyperplane inside which the class of the input vector was indeterminate or unclassifiable.

6.8 Validation

Extensive validation was then performed to determine the optimum parameters and kernels for classification of Epileptiform Activity. Various parameters and options both in the classification and data processing stages were tested and their performance measured.

Chapter 7 Classification Results and Comparison

This chapter looks at the performance results of all the models and methods described in the previous chapters. The best options are then identified and the obtained results are compared to previous or existing solutions. The last section looks at some of the areas where further work can still be done to improve or extend the solution.

7.1 Selected Systems

The performance classification method described in **Chapter 3** will be examined in this chapter. The extracted features, as described previously in **Chapter 5,** are:

Feature
Average sharpness of window
Average sharpness of window relative to channel
Maximum Laypunov exponent of window
Average Laypunov exponent of window
Standard deviation of window
Standard deviation of window relative to channel
Maximum coefficient from DWT analysis
Average of coefficients from DWT analysis
Maximum harmonic of window
Average harmonic of window
Mean value of window
Mean value of window relative to channel

Table 4: List of input features to SVM

These features were input into the PBIL process to determine which are best in separating the Epilpetiform pattern from normal activity.

The effect of the Artifact extraction / reduction techniques introduced in **Chapter** 4 as a pre-processing step is also analysed.

7.2 Selected Data

Data was taken from EEG channel readings as shown in the figure below from multiple patients. The figure depicts epileptiform activity in red with the yellow lines denoting the centre of the event. As can be seen it is not possible to visually differentiate epileptiform activity from other forms and so features must be extracted.

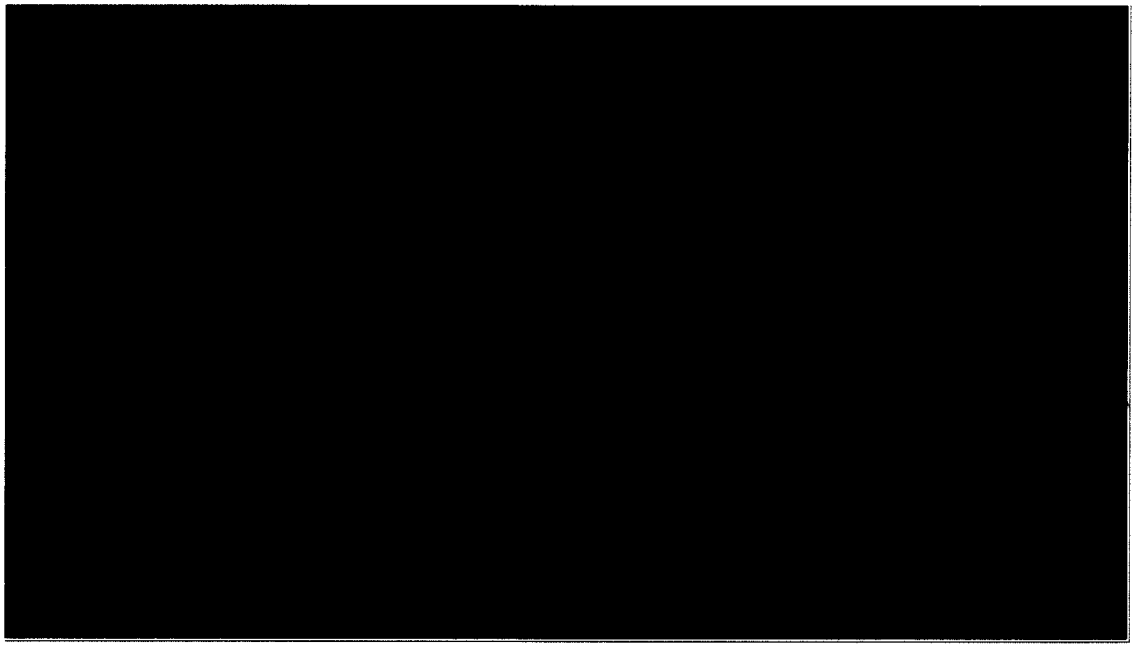

Figure 18: Average epileptiform activity per EEG channel

The quantity of data under analysis was reduced by identifying which channels consistently contained few Epileptiform labels. This was done by analysing a large quantity of data (40 data segments). The results are shown in the figure below. As can be seen from the figure below; there is almost no activity in channels 5, 19 and 20. These channels are therefore removed from analysis.

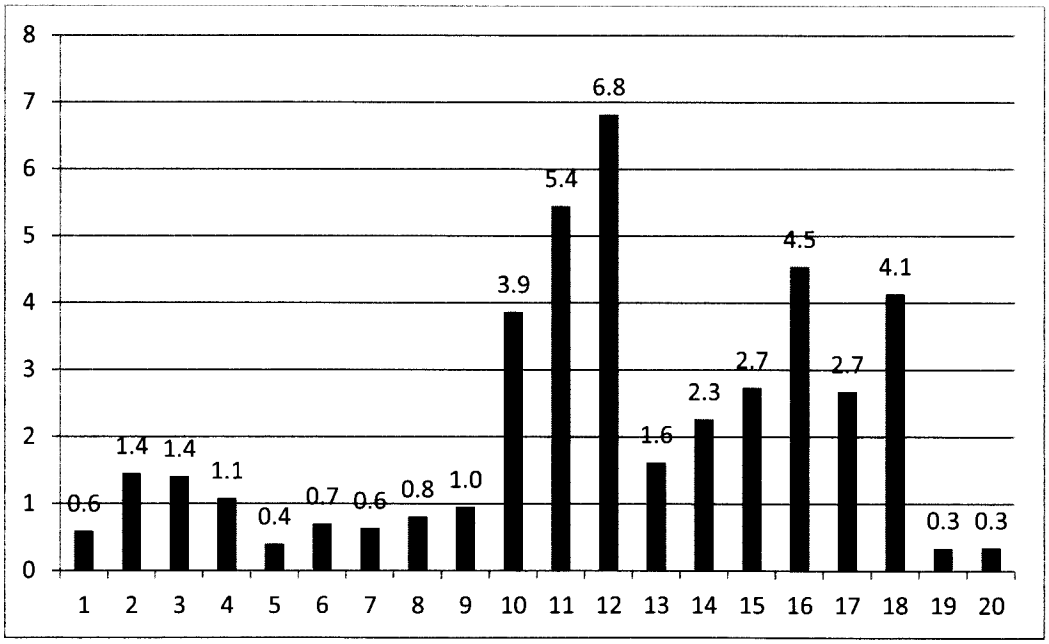

Figure 19: Average epileptiform activity per EEG channel

7.3 Performance Evaluation Methods

In order to measure the performance of a classification system and compare it to existing systems it is necessary to utilize a standardized analysis method. This allows for a quantitative analysis of the system performance that can be easily compared to existing solutions.

7.3.1 Measures

Given a classifier and a set of instances (the validation and test sets), a two-by-two confusion matrix (also known as a contingency table) can be constructed representing the dispositions of the set of instances. This matrix forms the basis of analysis for many common metrics. The confusion matrix is shown in the figure below.

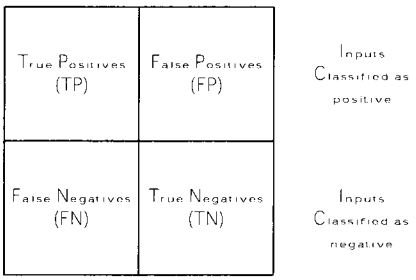

Figure 20: Layout of Confusion Matrix

The confusion matrix itself is not used directly in this work. The approach taken instead is to take the values (such as True Positives) to calculate performance measures for the system. These measures are shown in the table below [47].

Measure	Value
False Positive rate	FP/N
Sensitivity	TP/P
Specificity	TN / N
Precision	TP / (TP + FP)
Accuracy	(TP + TN) / (P + N)
F – measure	2/(1/precision + 1/sensitivity)

Table 5: Definition of Performance Measures

Where N = Total Negatives (TP + FN) and P = Total Positives (FP + TN)

The False Positive Rate is a ratio of the percentage of negatives that are misclassified as positive. The Sensitivity measures the ratio of how many positives are correctly classified to the total positive classifications. A high sensitivity means that the system has a high degree of certainty that the values classified as Epileptiform Activity are correctly classified.

The Precision and F-measure values are not used as they do not show any valid information due to the large imbalance between epileptic and non-epileptic events. The performance of the classifier is then evaluated and visually represented using ROC analysis.

7.3.2 ROC Analysis

Receiver Operating Characteristic (ROC) graphs are two-dimensional graphs in which the TP rate is plotted on the Y axis and FP rate is plotted on the X axis. This graph is a visual representation of the

trade-off or cost ratio between these two properties. The area enclosed below the curve represents the performance of the classifier averaged over all possible cost ratios. This research uses this analysis method to measure the accuracy of the different methods and compare them [48].

7.4 Results

This section lists all the results of the different systems and feature extraction methods and analyzes these data. The best features obtained by the PBIL algorithm are also presented together with their separability index. The effect of using a soft margin threshold whereby any output that is too close to the classifying hyperplane is labeled as unclassifiable is also discussed. The use of this threshold is due to the fact that vectors close to the hyperplane could be incorrect.

7.4.1 Optimal Feature Set

As described in **Chapter 5** the PBIL algorithm was run on the feature set in order to determine which features would be of most benefit to the classifier. This feature set was the one that best separated the data.

Feature	Used
Average sharpness of window	No
Average sharpness of window relative to channel	No
Maximum Laypunov exponent of window	Yes
Average Laypunov exponent of window	No
Standard deviation of window	Yes
Standard deviation of window relative to channel	Yes
Maximum coefficient from DWT analysis	Yes
Average of coefficients from DWT analysis	No
Maximum harmonic of window	Yes
Average harmonic of window	No
Mean value of window	No
Mean value of window relative to channel	No

Table 6: Best features for random sample of patients

The Separability Index achieved with these features was 77.6%. This high degree of separability suggests that the training process should produce an accurate classifier. It was noticed that running the PBIL algorithm multiple times produced varying results containing different sets of features. This was due to the random data sampling and was countered by looking at a wider range of data. This result was then re-used with smaller samples.

7.4.2 Results with best features

A classifier was created from the feature set obtained using the PBIL feature selection process.

7.4.2.1 Training Set

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	84.3	89.2	91.4	91.7
Specificity (%)	87.3	90.7	93.9	94.0
Accuracy (%)	85.8	90.0	92.7	92.8
Unclassifiable (%)	4.1	18.9	25.1	28.4

Table 7: Classification results for training set from random sample of patients

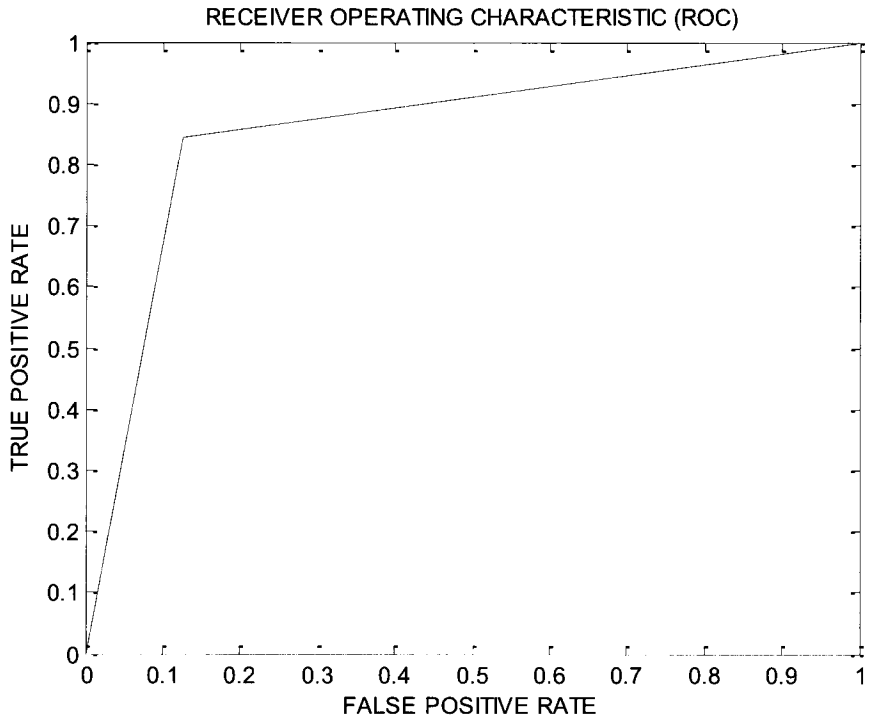

Figure 21: ROC Curve for training set from random sample of patients

7.4.2.2 Validation

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	76.8	82.3	84.3	86.2
Specificity (%)	85.0	88.6	90.6	91.2
Accuracy (%)	80.9	85.6	87.6	88.7
Unclassifiable (%)	3.9	18.6	27.6	31.1

Table 8: Classification results for validation set from random sample of patients

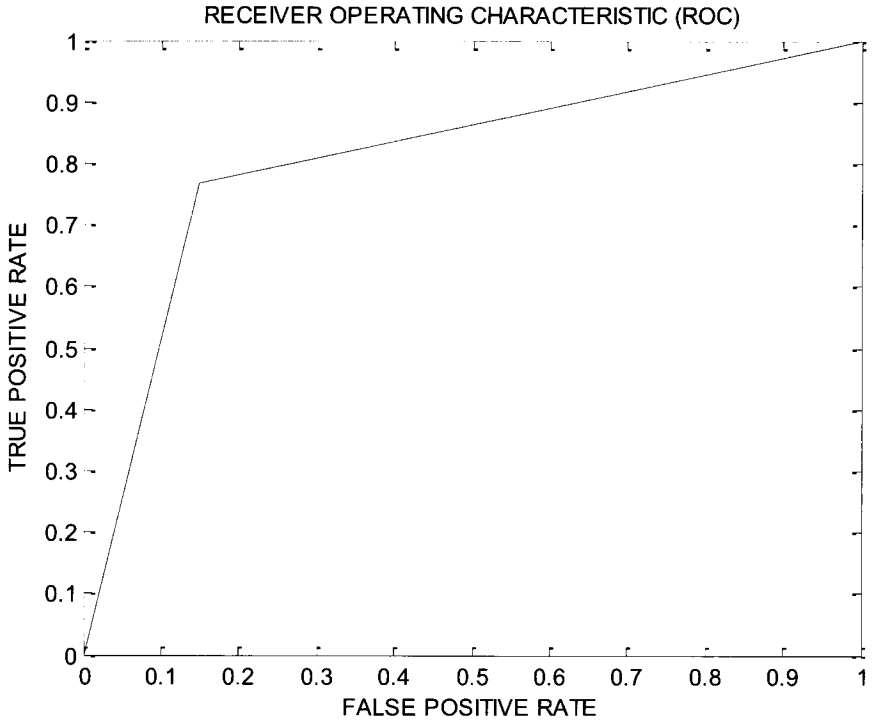

Figure 22: ROC curve for validation set from random sample of patients

The classifier produced a high degree of accuracy which was in line with the separabilty score obtained with the PBIL feature selection process. The relatively small difference between the training and classification results shows that the classifier has good generalization capabilities. The accuracy increased with a larger threshold on the soft margin with minimal improvement after 0.8.

7.4.3 Results with Artifact extraction

The artifact extraction techniques were run on the random subset of data taken from a sample of 20 patients. A new optimal feature set was then identified for this data as the extraction process changed the underlying patterns.

7.4.3.1 Best Feature Set

Feature	Used
Average sharpness of window	No
Average sharpness of window relative to channel	No
Maximum Laypunov exponent of window	Yes
Average Laypunov exponent of window	No
Standard deviation of window	No
Standard deviation of window relative to channel	Yes
Maximum coefficient from DWT analysis	Yes
Average of coefficients from DWT analysis	No
Maximum harmonic of window	No
Average harmonic of window	Yes
Mean value of window	No
Mean value of window relative to channel	No

Table 9: Best features for random sample of patients after artifact extraction

The Separability Index achieved with these features was 59.7%. This low degree of separability suggests that the training process would produce an inaccurate classifier.

7.4.3.2 Training Set

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	71.1	78.4	80.9	80.4
Specificity (%)	68.4	74.6	77.1	77.5
Accuracy (%)	69.7	76.5	79.0	79.0
Unclassifiable (%)	6.6	35.1	48.3	53.6

Table 10: Classification results for training set from random sample of patients after artifact extraction

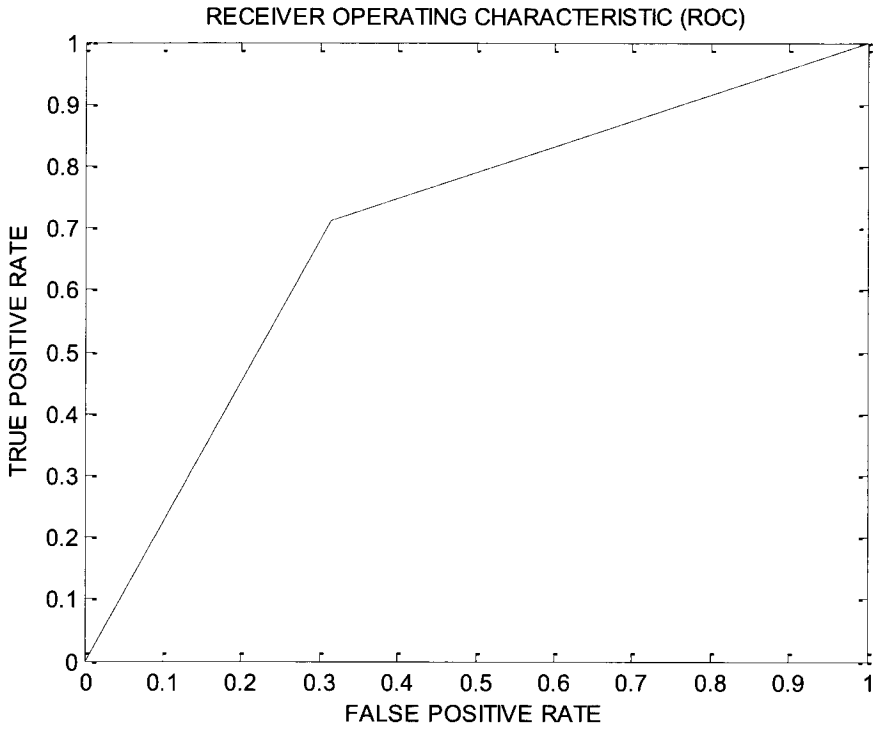

Figure 23: ROC curve for training set from random sample of patients after artifact extraction

7.4.3.3 Validation

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	34.9	30.3	25.8	22.3
Specificity (%)	89.8	97.0	98.9	98.8
Accuracy (%)	62.7	67.8	68.5	68.9
Unclassifiable (%)	7.9	33.3	44.1	48.5

Table 11: Classification results for validation set from random sample of patients after artifact extraction

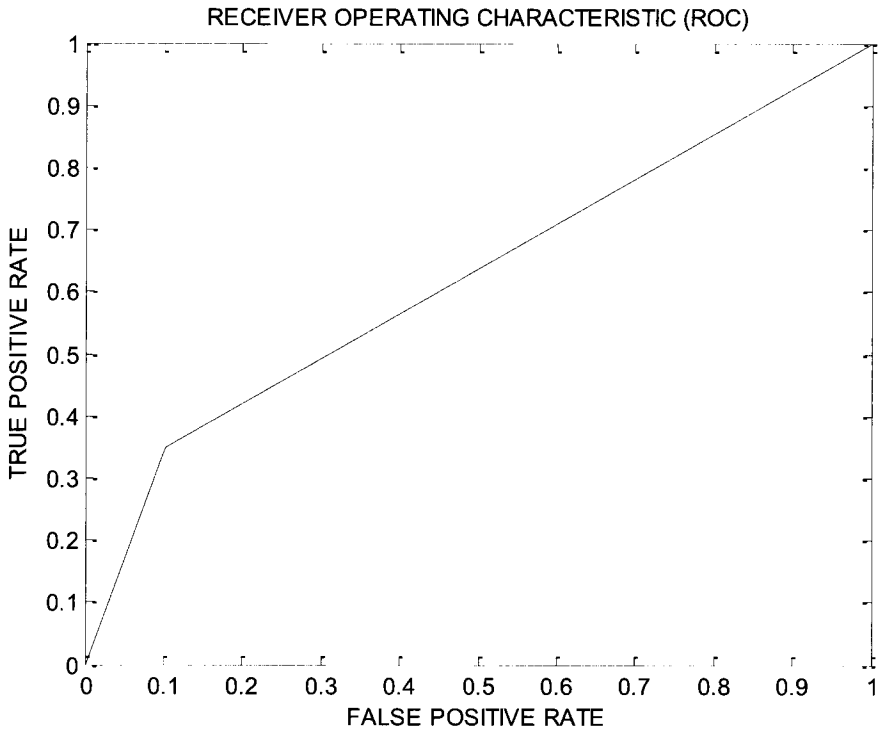

Figure 24: ROC curve for validation set from random sample of patients after artifact extraction

The classifier produced a low degree of accuracy which was in line with the separabilty score obtained with the PBIL feature selection process. The large difference between the training and classification results shows that the classifier has bad generalization capabilities and has overfit the training data. This indicates that the Artifact extraction process increased the already present variability across patients. The accuracy increased with a larger threshold on the soft margin with minimal improvement after 0.8. The sensitivity of the validation set was below 30% and so this classifier was too inaccurate to be of use.

7.4.4 Results with all features

7.4.4.1 Training Set

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	80.2	84.6	88.1	89.3
Specificity (%)	89.5	93.3	94.5	95.5
Accuracy (%)	85.0	89.1	91.3	92.4
Unclassifiable (%)	3.6	16.5	23.9	27.8

Table 12: Classification results for training set from random sample of patients without feature selection

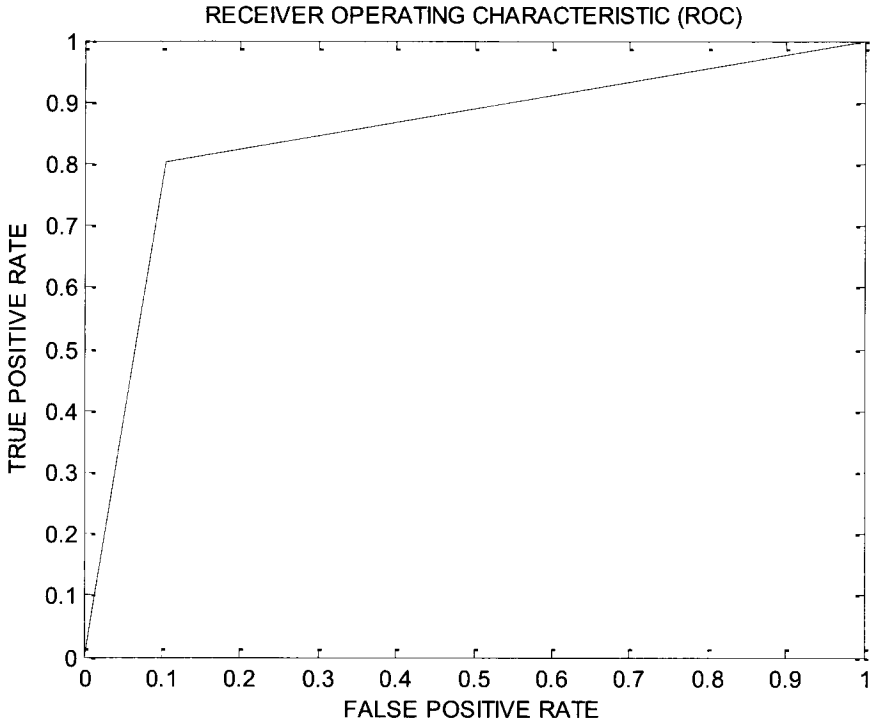

Figure 25: ROC curve for training set from random sample of patients without feature selection

7.4.4.2 Validation

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	66.0	67.6	70.2	71.0
Specificity (%)	88.3	93.2	94.9	95.0
Accuracy (%)	77.3	81.0	83.4	84.0
Unclassifiable (%)	3.3	19.0	26.1	30.0

Table 13: Classification results for validation set from random sample of patients without feature selection

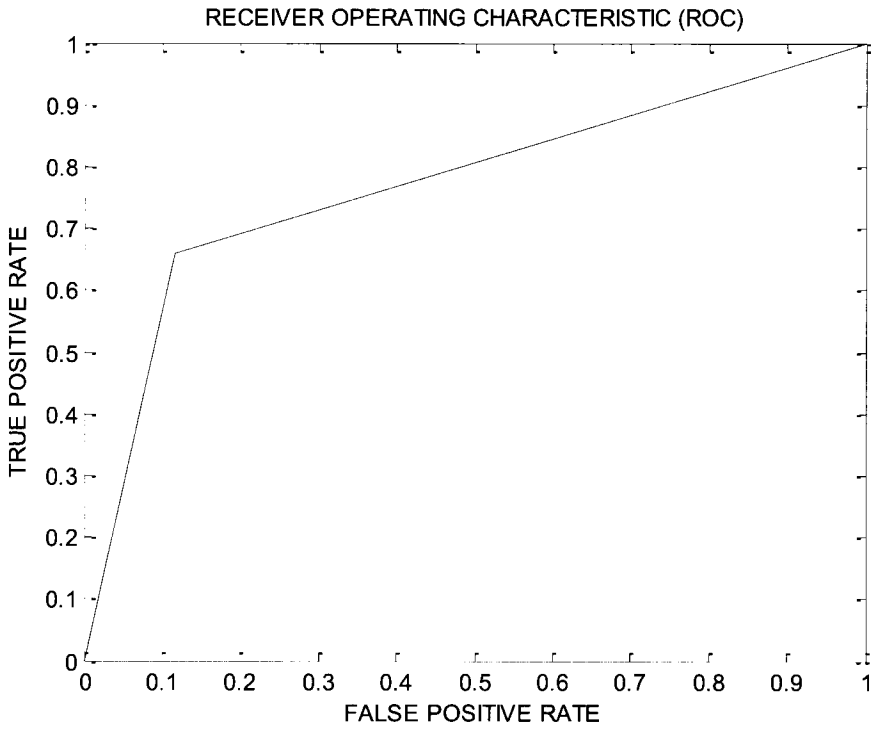

Figure 26: ROC curve for validation set from random sample of patients without feature selection

The classifier produced a variable degree of accuracy which was due to the extra features disguising the underlying pattern. The accuracy increased with a larger threshold on the soft margin with minimal improvement after 0.8.

7.4.5 Results with a Single Patient

7.4.5.1 Training Set

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	90.4	93.3	95.0	94.8
Specificity (%)	82.3	88.0	92.0	93.7
Accuracy (%)	86.4	90.7	93.5	94.3
Unclassifiable (%)	2.2	16.7	22.8	27.2

Table 14: Classification results for training set from one patient

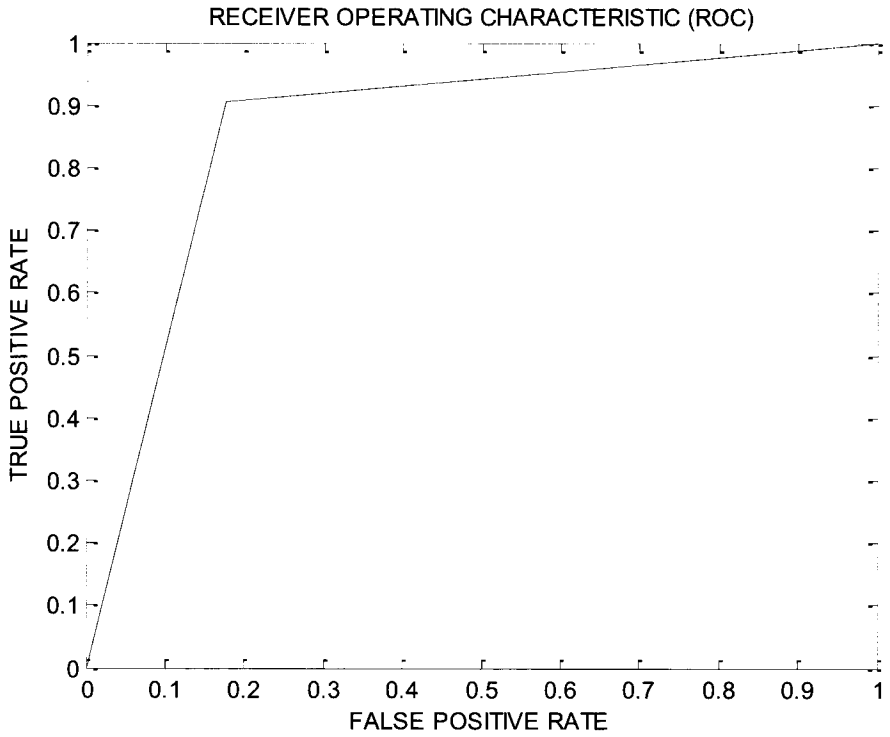

Figure 27: ROC curve for training set from one patient

7.4.5.2 Validation

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	68.8	73.0	75.2	74.7
Specificity (%)	84.4	88.0	91.2	91.0
Accuracy (%)	76.7	81.0	84.0	83.7
Unclassifiable (%)	4.7	25.6	35.8	38.6

Table 15: Classification results for validation set from one patient

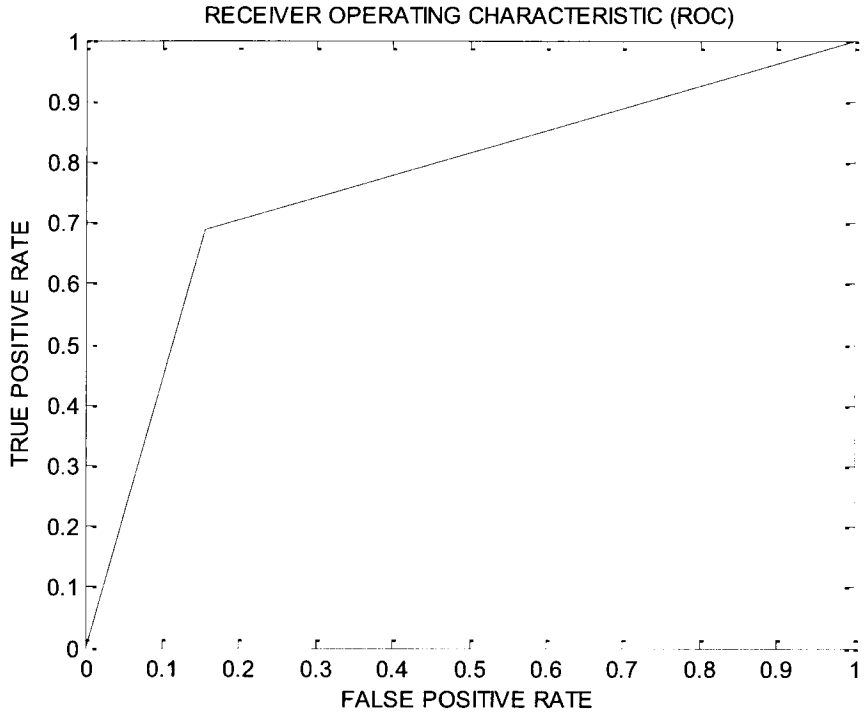

Figure 28: ROC curve for validation set from one patient

The classifier produced a high degree of accuracy which was in line with the separabilty score obtained with the PBIL feature selection process. The difference between the training and classification results shows that the classifier has average generalization capabilities. This decrease in performance was due to the smaller number of cases used in the training. The accuracy increased with a larger threshold on the soft margin with minimal improvement after 0.8.

7.4.6 Kernel Selection

All of the previous tests used the Gaussian Kernel function. This section now details the performance of the Linear Kernel.

7.4.6.1 Training Set

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	79.4	83.2	85.4	86.2
Specificity (%)	85.8	90.1	90.7	92.2
Accuracy (%)	82.6	86.7	88.0	89.2
Unclassifiable (%)	5.1	21.9	28.9	32.8

Table 16: Classification results for training set from random sample of patients with a linear kernel

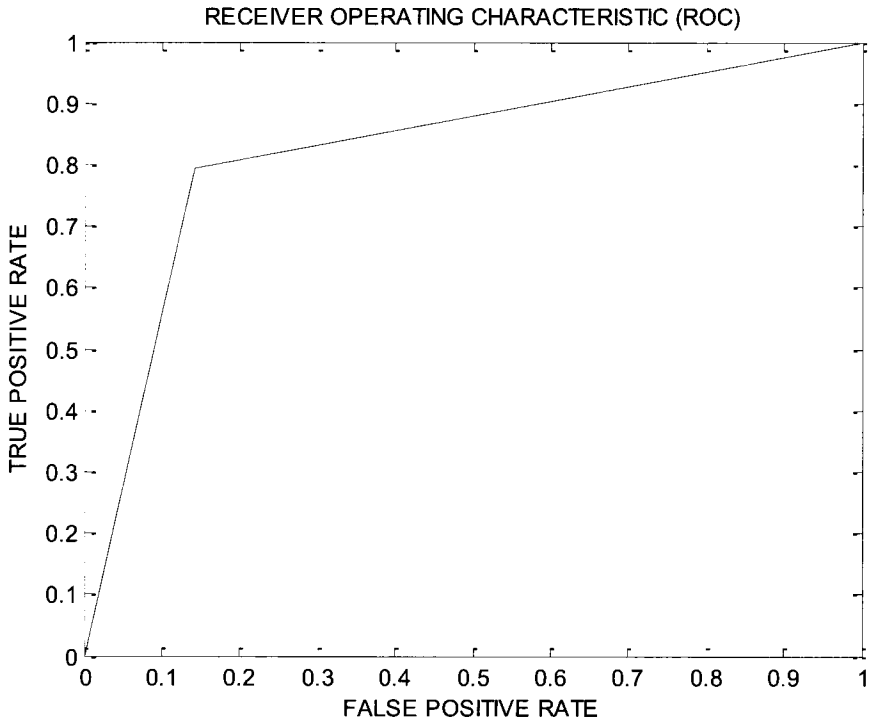

Figure 29: ROC curve for training set from random sample of patients with a linear kernel

7.4.6.2 Validation

Measure	Soft Margin			
	0.1	**0.5**	**0.7**	**0.8**
Sensitivity (%)	74.1	80.1	82.1	82.7
Specificity (%)	85.3	89.8	91.8	92.7
Accuracy (%)	79.6	85.2	87.2	87.9
Unclassifiable (%)	3.1	20.0	27.6	30.8

Table 17: Classification results for validation set from random sample of patients with a linear kernel

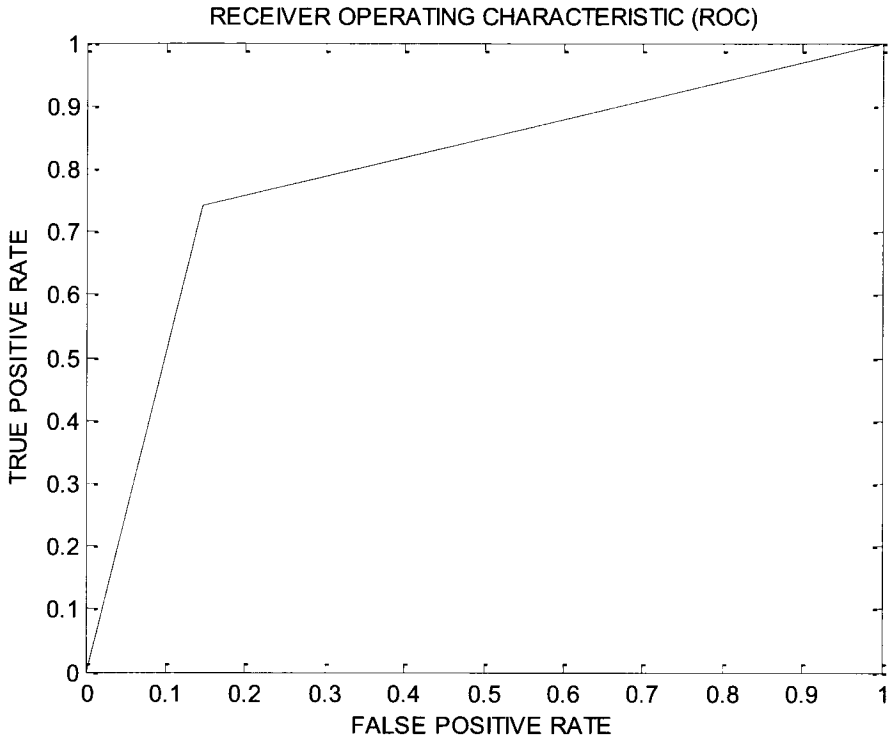

Figure 30: ROC curve for validation set from random sample of patients with a linear kernel

The classifier produced a high degree of accuracy which was in line with the separabilty score obtained with the PBIL feature selection process. The relatively small difference between the training and classification results shows that the classifier has good generalization capabilities. The accuracy increased with a larger threshold on the soft margin with minimal improvement after 0.8.

7.4.7 Comparison of Results

7.4.7.1 Kernel and Feature Selection

The Gaussian Kernel performed slightly better than the Linear Kernel, suggesting that it maps to a better separating feature space. The best features identified were:

- Maximum Laypunov exponent of window
- Standard deviation of window
- Standard deviation of window relative to channel
- Maximum coefficient from DWT analysis
- Maximum harmonic of window

This shows that the chaoticity, frequency and variability of the signal provide reproducibly accurate indicators of Epileptiform Activity.

7.4.7.2 Artifact Extraction

Artifact extraction significantly decreased the classification performance to an unusable level. This could be because the process is very difficult to automate and the Artifact Activity too similar in characteristics to Epileptiform Activity. It would require a more detailed pattern recognition approach than the thresholding technique used. Another problem with these methods is that they are computationally intensive as ICA needs to be performed on all patients or data sets prior to any other analysis.

7.4.7.3 Single Patient

Analysing data from a single patient reduced the accuracy of the classifier. This was due to the fact that the training process had less data to work with. Obtaining larger samples of data per patient would increase the performance of this method. This would potentially produce better results than looking at multiple patients due to the decrease variability in the data. A possible way to obtain more data would be to group patients based on their activity type.

7.4.7.4 Soft Margins

Thresholding values above a certain point (0.8) produced marginal improvements in performance at the cost of an increase in unclassifiable cases. The value of 0.8 was therefore chosen as optimal. The data lost due to being unclassifiable was seen to be an acceptable tradeoff for improved accuracy.

7.5 Comparison with Existing Solutions

The performance of this solution is comparable to other existing solutions. The 88.7% accuracy of the Gaussian kernel with optimised features was chosen as the upper comparison benchmark. This is in line with the performance of the ANN based classifiers discussed in **Chapter 2** which have over 90% accuracy. This shows that SVMs perform comparably to other classification methods with regard to EEG data and are therefore a viable tool.

This system is more flexible than some of the other existing solutions as it is not rule based and also does not rely on specific Epileptiform waveform shapes. These characteristics allow for a more robust system as there is great variation in waveform shape and quantity in the same patient as well as across different patients.

The high rate of unclassifiable data makes this system inappropriate for definitively establishing the presence of Epileptiform Activity in a given patient. The classification accuracy over the rest of the data is however sufficiently high to compensate for this. This method is also not useful to label specific localised elements as epileptic or not, it is instead useful for detecting the presence of Epileptiform Activity in a patient's EEG scan.

7.6 Possible Further Work and Improvements

There are various areas in the research that could be expanded upon or improved. This section discusses some of the more important areas identified.

A spike detection method could be investigated based on the fact that Epileptiform activity can resemble spike-like waveforms in the EEG signal. A problem with this method is that it could identify epileptic spikes in the EEG but would be less able to detect the wave complexes described in **Chapter 2**.

Epileptiform activity is not isolated to a specific point and propagates across multiple channels simultaneously from the point of origin. A method could be investigated that modelled the activity across multiple channels as components of a single signal.

The Artifact and Epilepsy Extraction Methods did not provide significant improvements to the performance of the system. This was due to the rule based approach taken in identifying Epilepsy or Artifact components produced by ICA decomposition. In order to improve the performance of these systems there are two possible methods that could be investigated:

1. Identify features and thresholds that better describe the components being identified (Artifact and Epilepsy)

2. Create a new automated classifier trained to detect the different component types. In order for this to work it is necessary to create or use a training set / library where the relevant components have been labelled.

Chapter 8 Conclusions

8.1 General Performance

The proposed system consists of three main components:

1. Artifact extraction
2. Feature Extraction and selection
3. SVM training

The Artifact extraction component did not function as required. The reason for this was that the data whitening step required by ICA removed some of the information contained in the original data set and so the selected feature set was not sufficiently descriptive. The other reason for the poor performance was that the rules used for identification were insufficient for the differentiation of Artifact Activity from Epileptiform data. This in turn leads to less information available for the feature extraction methods.

The feature extraction methods measuring the chaoticity, frequency and variability of the signal effectively extracted the features relevant to detecting Epileptiform Activity.

The SVM classification system was successfully trained and tested and displayed good generalization performance.

The benefits of this system are:

1. The system is computationally efficient
2. The system is accurate

The limitations of this system are:

1. The system cannot be used on patients that demonstrate excessive activity in the EEG as this would mask potential Epileptiform Activity
2. Large amounts of Artifacts could decrease the accuracy of the system

In view of the good performance of the solution this system could be used as a screening tool for detecting patients that possibly have epilepsy. The system cannot be used to rule out epilepsy in a patient due to the high rate of unclassifiable data.

8.2 Performance in Relation To Objectives

The system that was developed achieved the objectives for the research as defined in the first chapter. Each objective is discussed and analyzed.

1. The objective of creating a system to automatically detect Inter-ictal Epileptiform Activity in EEG data using Support Vector Machines was achieved.

2. The system that was created was computationally efficient. The total process (training + validation + testing) took an average of 20 minutes to run. This computational time increased exponentially with more data used.

3. The best system identified in this research produced an 88.7% accuracy which is sufficient in order to screen patients for Epileptiform Activity.

4. An accurate feature set was found providing a high degree of separability. These features measured the chaoticity, frequency and variability of the signal.

5. The performance of this solution was reviewed in relation to other existing solutions and was found to be comparable. The accuracy obtained is in line with previous solution using the same data set but with neural networks as well as other solutions in the literature. In order to achieve this accuracy 31.1% of the input data was unclassifiable and so discarded. Decreasing the Accuracy also decreased the amount of unclassifiable data. This could therefore be adjusted depending on the purpose of the application. The high percentage of data that was discarded limits the uses the usability of the system.

References

[1] *Epilepsy: aetiogy, epidemiology and prognosis*, http://www.who.int/publications, World Health Organisation, Fact sheet N°165, Revised February 2001

[2] Epilepsy Facts, http://www.emedicinehealth.com/epilepsy/article_em.htm, emedicinehealth, Last Accessed 30 May 2010

[3] Binnie C.D., Hermann S., *Modern electroencephalography: Its role in epilepsy management*, Institute of Epileptology, King's College London, London, UK, 1999

[4] Mohamed N., Rubin D. M., Marwala T., *Detection of Epileptiform Activity in Human EEG Signals Using Bayesian Neural Networks*, School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2006

[5] About Epilepsy, http://ucepilepsycenter.com/about-epilepsy, Epilepsy Center, Last Accessed 30 May 2010

[6] Epilepsy Information & Support, http://www.epilepsy.org.za/home/, Epilepsy South Africa, Last Accessed 30 May 2010

[7] Teplan M., *Fundamentals of EEG Measurement*, Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, Bratislava, Slovakia, 2002

[8] Honavar V., *Artrticial Intelligence: An Overview,* Artificial Intelligence Research Laboratory, Center for Computational Intelligence, Learning, and Discovery, Department of Computer Science, Iowa State University, 2006

[9] Luger G., Stubblefield W., *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (5th ed.),* The Benjamin/Cummings Publishing Company, Inc., 2004

[10] Basheera I.A., Hajmeerb M., *Artificial neural networks: fundamentals, computing, design, and application*, Engineering Service Center, The Headquarters Transportation Laboratory, CalTrans, Sacramento, USA

[11] Ethem A., *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*, MIT Press, 2004

[12] Kecman V., *Support Vector Machines Basics*, School of Engineering, The University of Auckland, 2004

[13] Gunn S. R., *Support Vector Machines for Classification and Regression*, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton, 1998

[14] Lin W., *A Case Study on Support Vector Machines versus Artificial Neural Networks*, University of Pittsburgh, 2004

[15] Panu E., *Support Vector Machines - Backgrounds and Practice,* Rolf Nevanlinna Institute, Helsinki, 2001

[16] Scholkopf B., Smola A J., *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond,* The MIT Press, Cambridge, Massachusetts, 2002

[17] Abe S., *Support Vector Machines for Pattern Classification,* Springer, London, 2005

[18] Ille N., Berg P., Scherg M., *Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies.*,J Clin, Neurophysiology vol. 19, pp. 113–24, 2000

[19] Jung T.P., Makeig S., Humphries C. Lee T.W., *Removing electroencephalographic artifacts by blind source separation*, Psychophysiology vol. 37, pp. 163–178, Cambridge University Press, 2000

[20] Joyce C.A., Gorodnitsky I.F., Kutas M., *Automatic removal of eye movement and blink artifacts from EEG data using blind component separation,* Psychophysiology vol. 41, pp. 313–25, 2004

[21] Zhou W., Gotman J., *Automatic removal of eye movement artifacts from the EEG using ICA and the dipole model*, Progress in Natural Science vol. 19, 2009

[22] Schlo¨ gl A., Keinrath C., Zimmermann D., Scherer R., Leeb R., Pfurtscheller G., *A fully automated correction method of EOG artifacts in EEG recordings*, Clinical Neurophysiology vol. 118, pp. 98–104, 2007

[23] Wallstroma G.L., Kassb R.E., Millerc A., Cohnd J.F., Foxe N.A., *Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods*, International Journal of Psychophysiology vol. 53, pp. 105–119, 2004

[24] Webber W.R., Litt B., Wilson K., Lesser R.P., *Practical detection of Epileptiform discharges (EDs) in the EEG using an artificial neural network: a comparison of raw and parameterized EEG data,* Electroencephalography and clinical Neurophysiology vol. 91, pp. 194–204, 1994

[25] Ozdamar O., Kalayci T., *Detection of spikes with artificial neural networks using raw EEG,* Comput. Biomed Res. Vol. 31, pp. 122–142, 1998

[26] Ko C.W., Chung H.W., *Automatic spike detection via an artificial neural network using raw EEG data: effects of data preparation and implications in the limitations of online recognition,* Clinical Neurophysiology vol. 111, pp. 477–481, 2000

[27] Gotman J, Gloor P., *Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG.,* Electroencephalography and clinical Neurophysiology vol. 41, pp. 513–529, 1976

[28] Polat K., Gunes S., *Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform,* Applied Mathematics and Computation vol. 187, pp. 1017–1026, 2007

[29] Jando G., Siegel R.M., Horvath Z., Buzsaki G., *Pattern recognition of the electroencephalogram by artificial neural networks, Elecroenceph,* Clinical Neurophysiol. vol. 86, pp. 100-109, 1993.

[30] Kalayci T., Ozdamar O., *Wavelet preprocessing for automated neural network detection of EEG spikes,* IEEE Eng. Med. Biol. Mag., pp. 160-166, 1995

[31] Ubeyli E.D, *Combined neural network model employing wavelet coefficients for EEG signals classification,* Digital Signal Processing vol. 19, pp. 297–308, 2009

[32] Patnaika L.M., Manyamb O.K., *Epileptic EEG detection using neural networks and post-classification,* Computer methods and programs in biomedicine vol. 91, pp. 100–109, 2008

[33] Acir N., Guzelis C., *Automatic spike detection in EEG by a two-stage procedure based on support vector machines,* Computers in Biology and Medicine vol. 34, pp. 561–575, 2004

[34] Harris FJ., *On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform,* Proceedings of the IEEE vol. 66, pp. 51–83, 1978

[35] Hyvärinen A., Oja E., *Independent Component Analysis: Algorithms and Applications*, Neural Networks Research Centre, Helsinki University of Technology, Neural Networks vol. 13, pp. 411-430, 2000

[36] Hyvärinen A., Oja E., *A Fast Fixed-Point Algorithm for Independent Component Analysis*, Laboratory of Computer and Information Science, Helsinki University of Technology, 1997

[37] Duhamel P., Vetterli M., *Fast fourier transforms: A tutorial review and a state of the art*, Columbia University, New York, 1989

[38] Daubechies I., *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[39] Adeli H., Zhou Z., Dadmehr N., *Analysis of EEG records in an epileptic patient using wavelet transform*, The Ohio State University, Columbus, 2002

[40] Polat K., Gunes S., *Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fat Fourier transform*, Department of Electrical and Electronics Engineering, Selcuk University, Turkey, 2007

[41] Iasemidis L.D., Sackellares J.C., *Chaos Theory and Epilepsy*, Department of Neurology, Neuroscience, and Electrical Engineering, University of Florida, 1996

[42] Lai D., Chen G., *Statistical Analysis of Lyapunov Exponents from Time Series: A Jacobian Approach*, University of Texas Houston, 1998

[43] Swiderski B., Osowski S., Rysz A., *Lyapunov Exponent of EEG Signal for Epileptic Seizure Characterization*, Warsaw Technol. Univ., Poland, School Of Computer Science, Carnegie Mellon University, Pennsylvania, 1994

[44] Baluja S., *Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning*, School Of Computer Science, Carnegie Mellon University, Pennsylvania, 1994

[45] Greene J., *Feature subset selection using Thornton's separability index and its applicability to a number of sparse proximity-based classifiers*, Department of Electrical Engineering, University Of Cape Town, South Africa

[46] Burges C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition*, Bell Laboratories, Lucent Technologies, Data Mining and Knowledge Discovery vol. 2, pp. 121–167, 1998

[47] Sokolova M., Japkowicz N., Szpakowicz S., *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*, Lecture Notes in Computer Science vol. 4304, pp. 1015-1021, 2006

[48] Fawcett T., *An introduction to ROC analysis*, Institute for the Study of Learning and Expertise, Palo Alto, USA, 2005

Abstract

This dissertation evaluates the effectiveness of using Support Vector Machines (SVM) to identify Inter-ictal epileptic activity in Electroencephalogram (EEG) data. There are existing systems that already do this but identifying the best solution requires comparative studies. A sample of data was randomly selected from 20 patients. A number of features were extracted and a PBIL algorithm was then used to identify the set of features that provided the best separability within the dataset. These were found to be related to the chaoticity, frequency and variability of the signal. The features were then used to train an SVM model. This method resulted in 88.7% accuracy but left 31.1% of inputs unclassifiable. This performance is comparable with existing solutions.

The strengths of the designed system were computational efficiency and system accuracy. The limitations were that the high degree of Artifacts masked indicative patterns and therefore decreased the classification accuracy in some cases. This system could be used as a screening tool for detecting patients that possibly have epilepsy. The system cannot be used to rule out epilepsy in a patient due to the high rate of unclassifiable data.

Some areas were identified where future work could be done, namely a spike detection method, multi-channel analysis, and improved Artifact extraction classification. These areas could improve the general performance of the classification system in terms of more clearly separating epileptic from non-epileptic activity.

Abstract

This dissertation evaluates the effectiveness of using Support Vector Machines (SVM) to identify Inter-ictal epileptic activity in Electroencephalogram (EEG) data. There are existing systems that already do this but identifying the best solution requires comparative studies. A sample of data was randomly selected from 20 patients. A number of features were extracted and a PBIL algorithm was then used to identify the set of features that provided the best separability within the dataset. These were found to be related to the chaoticity, frequency and variability of the signal. The features were then used to train an SVM model. This method resulted in 88.7% accuracy but left 31.1% of inputs unclassifiable. This performance is comparable with existing solutions.

The strengths of the designed system were computational efficiency and system accuracy. The limitations were that the high degree of Artifacts masked indicative patterns and therefore decreased the classification accuracy in some cases. This system could be used as a screening tool for detecting patients that possibly have epilepsy. The system cannot be used to rule out epilepsy in a patient due to the high rate of unclassifiable data.

Some areas were identified where future work could be done, namely a spike detection method, multi-channel analysis, and improved Artifact extraction classification. These areas could improve the general performance of the classification system in terms of more clearly separating epileptic from non-epileptic activity.