TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MORTEZA ZABIHI

PATIENT-SPECIFIC EPILEPTIC SEIZURE DETECTION IN LONG-TERM EEG RECORDING IN PAEDIATRIC PATIENTS WITH INTRACTABLE SEIZURES

Master's thesis

Examiners: Professor Serkan Kiranyaz, Professor Turker Ince, and Professor Moncef Gabbouj
Examiner and topic approved in the Natural Sciences Faculty Council meeting on 8 May 2013.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master Degree Programme in Biomedical Engineering
Zabihi, Morteza: Patient-specific epileptic seizure detection in long-term EEG recording in paediatric patients with intractable seizures
Master of Science Thesis, 61 pages
March 2013
Major: Medical Informatics
Examiner: Professor Serkan Kiranyaz, Professor Turker Ince, Professor Moncef Gabbouj
Keywords: Electroencephalograph, Epilepsy, Intractable Seizure, Conditional Mutual Information Maximization, Support Vector Machine, Sensitivity, Specificity

Over recent years, due to the increase in the epileptic patient population, issues of diagnosing and treatment of epilepsy have become more and more prominent and much research has been done in this field in consequence. However, there are still many gaps and lack of knowledge in interpreting Electroencephalograph (EEG) signals in order to solve the problem.

Particular problems in this area include difficulties in detecting the seizure events (due to the different seizure types and their variability from patient to patient or even in an individual over time), and dealing with long-term EEG recordings, which is an onerous and time consuming task for electroencephalographers.

The thesis discusses the two problem areas using EEG data from four subjects with overall 21 hours of recording from the CHB-MIT scalp benchmark EEG dataset. We propose a patient specific seizure detection technique, which selects the optimal feature subsets, and train a dedicated classifier for each patient in order to maximize the classification performance. To exploit the characteristics of a patient's EEG pattern as much as possible, we used a large set of features in the proposed framework, namely time domain, frequency domain, time-frequency domain and nonlinear features, and selected the most crucial features among them by using Conditional Mutual Information Maximization (CMIM) technique. We further performed extensive comparative evaluations against 6 other feature selection methods to demonstrate the superiority of the CMIM.

Support Vector Machine (SVM) with the linear kernel is used as the classifier. The experimental results show a delicate classification performance over the test set, i.e. an average of 90.62% sensitivity and 99.32% specificity are acquired when all channels and recordings are used to form a composite feature vector. In addition, an average sensitivity and specificity rates of 93.78% and 99.05% are obtained using CMIM, respectively.

PREFACE

This master thesis was carried out in the MUVIS group at the signal processing department of Tampere University of Technology. It was funded by the Academy of Finland.

I wish to express my sincere gratitude to Professor Serkan Kiranyaz, whose inspiring guidance and inexhaustible patience have helped me cultivate a systematic approach towards the research. I am indebted to him for the freedom he gave me in implementing my ideas. I have also acquired valuable insights through his instruction in academic approach.

I am very much grateful to Professor Turker Ince for his warm support and his valuable suggestions. Besides, I would also like to thank Professor Moncef Gabbouj for providing me with the opportunity for the thesis work. He gave me a lot of trust and flexibility in the project.

I have been extremely fortunate in having valuable friends who have made a homely atmosphere for me during my stay in Tampere. I express my special gratitude to my friend Ali Bahrami Rad for his many useful practical tips and help in various stages. I am very grateful to my close friends Pooya Saketi and Milad Mostofizadeh who have been sources of constant encouragement and friends in the true sense of the word. In addition, I am lucky to have friends like Mozafar Iqbal and Reza Mohseni who are not here in Finland, but long distance has never managed to diminish our friendship.

The pleasant and supportive atmosphere throughout my stay in the Signal Processing Department is appreciated. I am grateful to my colleagues, especially Stefan Uhlmann and Guanqun Cao for their kind help and comments during these months.

I am immensely indebted to my parents, brothers and sisters for their unflinching belief in my endeavours.

Tampere, 1st of May, 2013
Morteza Zabihi

CONTENTS

1. INTRODUCTION

1.1. Epilepsy

1.1.1. Definition and History

The definition and the history of epilepsy are intertwined sothat the definition of epilepsy has formed over time. Thus, in this section both the definition and history of epilepsy are brought together.

The origin of "Epilepsy" is from the Greek word *"epilēpsia*", which means to seize upon or to take hold of. In other words, epilepsy was believed to be an illness in which the patient was seized upon, presumably by a supernatural force. In 400 B.C., Hippocarates wrote a book about epilepsy. He was the first person who refused religion as a justification for the source of diseases [1]. Superstitious interpretations about epilepsy had existed until 1859, when three English neurologists – J. H. Jakson, R. Reynolds, and Sir W. R. Gowers developed a scientific approach to epilepsy and seizure occurrence [2].

From the earliest attempts to define epilepsy, it was found out that defining it is challenging. In 1877, H. Nothangel pointed to this problem and claimed that it is not possible to develop an adequate definition, at least in a brief manner. However, it is necessary to have a definition for epilepsy even if it is limited. One solution is to define epilepsy according to its clinical symptoms. In 1875, epilepsy was defined by J. S. Jewell as a sudden attack of loss of consciousness or at least impairment of consciousness. At about the same time, it was claimed by W. A. Hammond that in addition to unconsciousness, spasm is another symptom of epilepsy.

Another way of defining epilepsy is based on pathophysiology. For instance, the definition by J. H. Jackson, which defines epilepsy as "*a sudden, excessive, and rapid discharge of the grey matter of some part of brain*", can be put in the same category. Three years later in 1876 he completed his previous definition, which is also used nowadays: "*Epilepsy is a chronic disorder in which there are recurring, sudden, excessive, and rapid discharges of the grey matter of some parts of the brain, the clinical manifestations of which are determined by the anatomical site in the brain*" [3].

It must be noted that seizure is not the name for any diseases. In essence, epileptic seizures are an abnormal brain activity that is caused by many diseases. Epilepsy in medical applications defined as a disorder in neurological condition in which electrical malfunction of neurons in the brain is counted as the main reason. In essence, epilepsy is an abnormality in the normal electrical activity pattern of nerve cells. The most apparent symptom of epilepsy is seizure occurrence that is defined as "*a transient occur-*

rence of signs of synchronous neuronal activity in the brain" in [4]. In addition, it should be noted that seizure does not necessarily mean epilepsy. Febrile seizures, non-epileptic events, and eclampsia are some examples to show the difference between seizure and epilepsy.

1.1.2. Types of Epilepsy

Epileptic seizures have different types and can be classified in various ways. In the most common way, seizures are classified into two classes according to the location of occurrence in the brain: partial (or focal) and generalized. However, The International League Against Epilepsy (ILAE) has recommended abandoning terms of generalized and partial since there are many cases that cannot be fitted into these categories [5]. Instead, new terms such as genetic, structural-metabolic and unknown are defined to fit more seizures. A brief explanation about the former and latter classification is given in this section.

Partial or focal seizures occur in one hemisphere of the brain. In this category, seizures are named according to the area of the brain that generates the seizure. For instance, focal frontal lobe seizures are categorized under partial seizures. Partial seizures are divided into two sub-sections called simple and complex focal seizures. In simple focal seizures, the patient does not lose his or her consciousness. The symptoms could be unusual feelings or sensations (e.g., sudden and unexplained feeling of joy, anger, sadness, or nausea, or sense or feeling of things that do not actually exist) [6]. It should be noted that in the latter classification of seizures, simple and complex do not fit in the category of partial seizures [5].

In generalized seizures, the entire left and right lobes of brain show abnormal activity of neurons and the patient may lose his or her consciousness in this sub-section. Absence seizure, tonic seizure, clonic seizure, myoclonic seizures, atonic seizures, tonic-clonic seizures are classified under the generalized seizures [7]. Some properties of generalized seizures are listed in Table 1.1. In Table 1.2, new terms such as genetic, structural-metabolic and unknown, which are defined by ILAE, are briefly described.

Table 1.1. *Name and symptoms of generalized seizures according to ILAE classification*

Name of seizure	**Properties**
Absence seizure (petit mal seizure)	Happen few times or more than 100 times during a day, staring, rapid blinking, repetitive eye and extremity movement, no memory of what happened, mostly in children 4-12 years old, brief loss of consciousness
Tonic seizure	Stiffness of muscles, rigidity
Clonic seizure	Muscles repetitively jerk and relax
Tonic clonic seizure (grand mal seizure or convulsive seizure)	First stiffness of arms or legs (tonic stage) and then limbs and head begin jerking (clonic stage)
Myoclonic seizure	Parts of a person's body jerk e.g., arms or legs twitch
Atonic seizure (drop attack, astatic or akinetic seizures)	Limpness of whole or part of body suddenly

Table 1.2. *Genetic, structural-metabolic and unknown seizures and their description according to ILAE classification* **[5]**.

Name of seizure	Description
Genetic Seizure	The epilepsy is a direct result of a genetic cause, a gene and the mechanisms should be identified; e.g., channelopathies
Structural-metabolic Seizure	The epilepsy is the secondary result of a separate structural or metabolic condition;
Unknown Seizure	Indicates that further investigation is needed to identify the cause of the epilepsy.

In Table 1.3, the last classification proposed by the Commission on Classification and Terminology can be seen. Although this classification is officially recommended, it has not found a general acceptance yet. This might be due to use of new terms and the elimination of some familiar expressions such as "*grand mal*" and "*petit mal*" while many neurologists still use these terms even if they are not correct. The more the separation of seizures is studied, the more nebulous the classification task becomes, and there are many limitations in every type of classification [2].

In addition to the mentioned types of epileptic seizures, there is another seizure type named intractable seizure. The National Association of Epilepsy Centers (NAEC) considers intractable epilepsy as an epilepsy in which a person's seizures do not come under control after nine months of treatment under the care of a neurologist. In essence, intractable epilepsy can be defined as the severity of the seizure; for instance, in a case in which there are two individuals with the same type of seizure, one with severely incapacitating seizure and the other person with controllable seizure, the former type of seizure is named intractable, while the latter one is not recognized as an intractable seizure [8]. Thus, a patient with intractable seizure may suffer from different types of seizures.

Table 1.3. *Common classification of epileptic seizures* **[2]**

I.	Partial (focal, local) seziures	A. Simple partial seizures (consciousness not impaired)	1. With motor signs
			2. With somatosensory or special sensory symptoms
			3. With autonomic symptoms or signs
			4. With psychic symptoms
		B. Complex parietal seizures	1. Simple partial onset, followed by impairment consciousness
			2. With impairment of consciousness at onset
		C. Partial seizures evolving to secondarily generalized seizures	1. Simple partial seizures evolving to generalized seizures
			2. Complex partial seizures evolving to generalized seizures
			3. Simple partial seizures evolving to complex partial seizures evolving to generalized seizures
II.	Generalized (conclusive or nonconclusive) seziures	A. Absence seizures	1. Typical absences, alone or in combination
			2. Atypical absence
		B. Myoclonic seizures	
		C. Clonic seizures	
		D. Tonic seizures	
		E. Tonic-clonic seizures	
		F. Atonic seizures (astatic)	
III.	Unclassified epileptic seziures		
IV.	Addendum, with respect to occurance of seizures (cyclic, fortuitous) or perception by triggering eventts		

1.1.3. Importance and Challenges of Epilepsy Detection

Studies and research on the detection and treatment of epilepsy have tremendously increased in recent years. The considerable number of epileptic patients and numerous side effects of seizure attacks mean that more attention is given to diagnosis and curing the disease.

According to a World Health Organization (WHO) report, more than 50 million people have epilepsy, of whom 85% are from developing countries, where 60% to 90% receive no treatment due to inadequacies in health resources. Around 2.4 million new cases add to epileptic patients each year globally; and it begins in childhood or adolescence in at least 50% of the cases. Fortunately, 70% to 80% of epileptic patients can have normal lives if properly treated, which shows the importance of epilepsy diagnosis [9].

Different seizure types (Section 1.1.2) affect patients differently. For instance, memory may be affected by a generalized tonic-clonic or a complex partial seizure or in patients with atonic seizure limpness of the body during the seizure event is regular. Medication also has side effects which may include inattention or restlessness. Epilepesy could be the symptom of many other diseases. Epileptic patients may suffer from other disorders such as hyperactivity disorder (ADHD), autism, or mental retardation, which cause, for instance, decrease in the performance of learning in children.

In addition to medication, a surgical approach is used for epilepsy treatment. There are five main types of surgical treatment [10]:

1. Focal resection for hippocampal sclerosis and other lesion on the mesial temporal lobe
2. Focal resections for other overt lesions (lesionectomies) in temporal neocortex or other cortical areas
3. Non-lesional focal resections (where there is no lesion on imaging, but epileptic tissue is localized by functional methods and/or on clinical grounds)
4. Hemispherectomy, hemispherotomy and other large multilobar resections
5. Functional procedures-multiple subpial transection, corpus callosectomy, focal ablation, focal stimulation, vagal nerve stimulation.

In both surgical and medication treatment, the detection of seizure type and its localization in the brain play a crucial role.

Different clinical methods are used to detect seizures such as Magnetic Resonance Imaging [11], Positron Emission Tomography [12], three-dimensional accelerometer [13], and Electroencephalography (EEG). The EEG is the most common method for seizure detection, especially for long-term monitoring since it is cost-effective, fast, and does not have any side effects.

There are various signal processing algorithms for the diagnosis of epileptic seizures using EEG signals. These algorithms may be real time or non-real time. In addition, the detection may only detect the onset of the seizures or the whole seizure section may be detected. Some algorithms are limited to one or two types of special seizures or are lim-

ited to short-term EEG signals. According to different needs, different methods and algorithms are demanded.

1.2. Electroencephalography

This section is organized as follows. In section 1.2.1 a brief explanation about the anatomical and physiological of nerve cells is given and Electroencephalography (EEG) is defined. In section 1.2.2 a summarized history of EEG and its development is given. The most common standard electrode positioning is introduced in section 1.2.3, and finally in section 1.2.4 the application of EEG in analysing epilepsy especially the detection of seizure in recent research is discussed.

1.2.1. Definition

The brain is made up of many cells including nerve cells. The number of nerve cells in the brain is about 10^{11}. These cells are responsible for transmitting nerve signals to and from the brain. The nerve cell consists of three parts based on their structure and function:

1) The cell body (soma)
2) Axon
3) Dendrites

The cell body includes the nucleus, mitochondria, endoplasmic reticulum, ribosomes, and some other organelles. These organelles are similar to other cells. The cell body is surrounded by a membrane, which plays an important role in generating electrical signal. The membrane is composed of two layers (inner and outer layers) of phospholipid molecules, which are protein (about 60% of the membrane), and lipid or fat (about 40% of the membrane). It is constructed in such a way that it allows the passage of certain charged components (ions such as Na^+, K^+, Cl^-, Ca^{++}) of the solutions. The membrane, however, does not allow the passage of all the ions present in the solutions and is thus a selectively permeable membrane. The impermeability of the membrane is typically related to the size of a particular ion. Also, the total number of charged molecules on either side of the membrane is equal. A consequence of the selective permeability of the membrane barrier is the development of an electrical potential between the two sides of the membrane. These electrical potentials are also known as "*action potentials*", "*nerve impulses*" or "*spikes*". In essence, action potential depends on the voltage-gated sodium and potassium channels.

The axon is a long nerve fiber that transfers the electrical signal from the cell body to another nerve or to a muscle cell. Axon in mammalian usually is $1 - 20\ \mu m$ in diameter. Some kinds of axon are covered by an insulating layer called the "*myelin sheath*", which is made of "*Schwann cells*". The myelin sheath is divided into sections along the axon, where each of these intervals is called "*node of Ranvier*". Dendrites are the exten-

sion of the cell body, which receive electrical impulses from other cells and transfer them to the cell body. In Figure 1.1, different parts of the cell are shown [14].

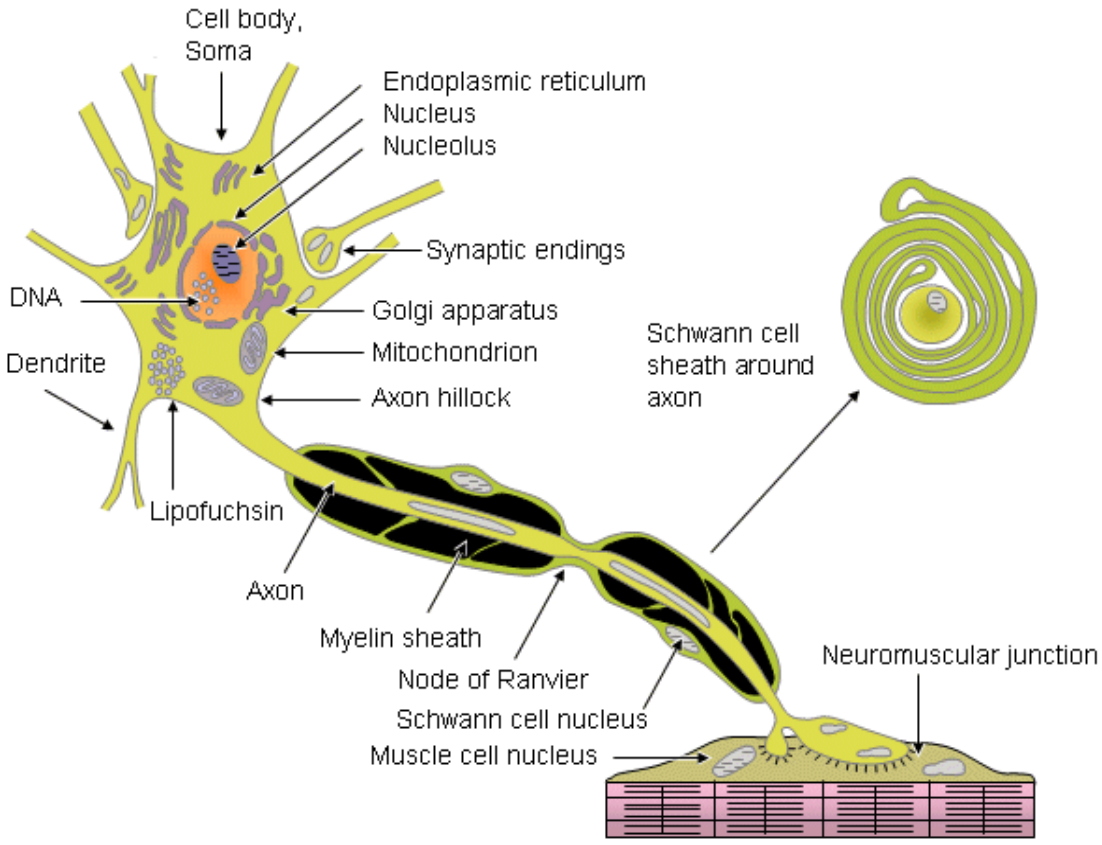

Figure 1.1 *The different parts of a neuron* **[15]**

Several terms must be defined first, in order to describe the changes in the action potential:

1) Polarization: the stage when the membrane maintains the difference in electrical charge between outside and inside, is called polarization. With respect to the outside of the membrane, the inner side has slightly negative electrical potentials. This difference potential is named the resting potential, which is mainly the result of negatively charged proteins inside the cell. In a typical neuron, resting potential is around $-70\ mV$.
2) Depolarization: the stage when the membrane becomes less negative that at resting potential (e.g., a change from -70 to $-60\ mV$) is called depolarization. This term also refers to the state when the inside potential with respect to the outside of the membrane becomes $+30\ mV$.
3) Repolarization: this happens when the membrane returns to resting potential after the depolarization stage.
4) Hyperpolarization: the stage when the inside of the membrane becomes more negative with respect to the outside (Figure 1.2).

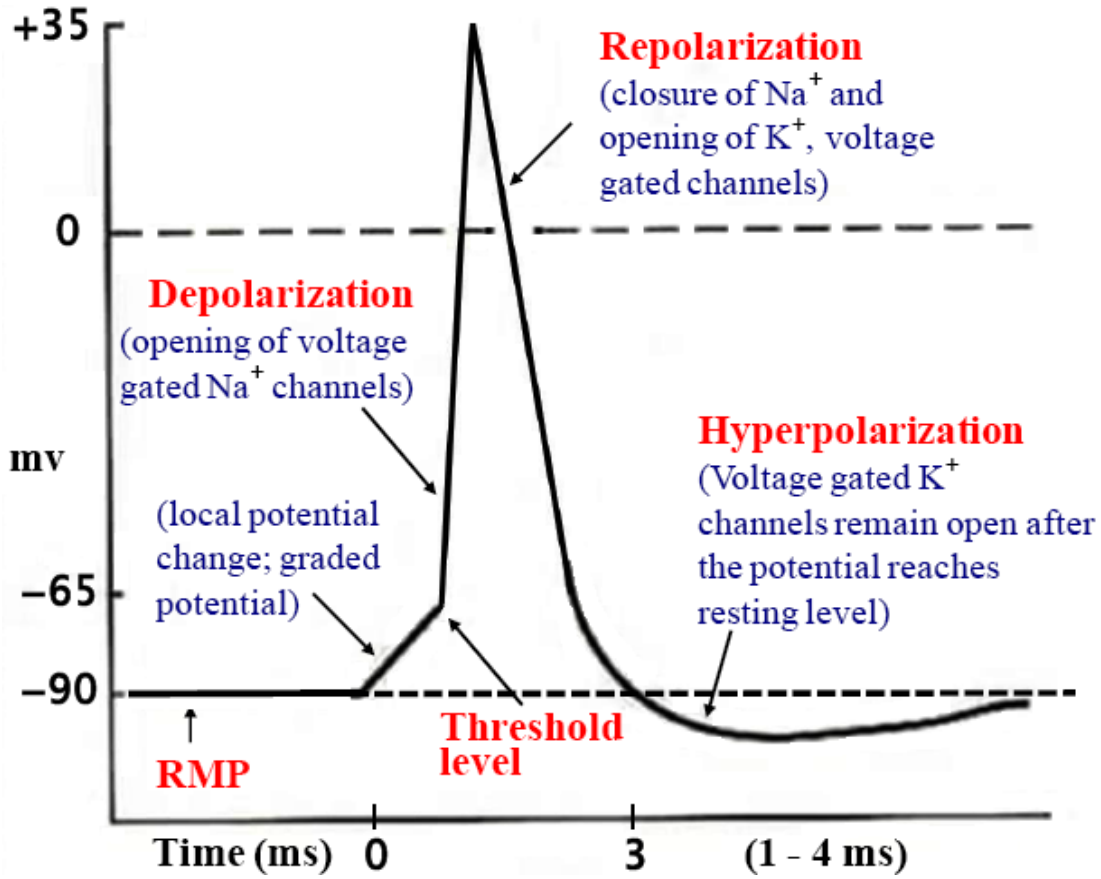

***Figure 1.2**. The action potential of a nerve cell with its different stages.*

A generation of bioelectric signal in nerve cells can be summarized as follows: in the resting state there are more K^+ ions than outside of the cell and in contrast, there are more Na^+ ions existing outside the cell membrane. In this state (i.e., resting state) the inside of the cell has about -70 to $-90\ mV$ electric potential (shown by RMP in Figure 1.2). The generation of bioelectric signal starts with the depolarization process, when Na^+ ions go from outside to inside from higher to lower concentration and make the inside of the cell positive. In repolarization the opposite process happens; K^+ channels are open during repolarization and let the K^+ ions go outside the cell, which restore the ionic balance (sodium-potassium ion pump). This kind of pump pumps the K^+ and Na^+ ions inside and outside the cell, respectively. In the hyperpolarization process, K^+ channels remain open, which makes the inside potential a little more negative in contrast to the resting state. This process is contiguous during the production of action potential.

The amplitude of the nerve impulse (action potential) is about 100mV and it lasts about 1ms [15]. Recording of the collective electrical activity of these nerve cells in the cerebral cortex is named Electroencephalography (EEG). It should be noted that the activity of single neuron cannot be measured on the scalp due to thick layers of tissue (e.g., fluids, bones, and skin) [16]. EEG is recorded by placing several electrodes on the scalp. The applications of EEG can be categorized into 3 main parts:

1) Measuring "*spontaneous activity*", which measures the EEG signals on the scalp. This is the most common application of EEG. The term spontaneous implies the continual activity of the nerve cells in the brain.
2) Measuring "*evoked potentials*", which measures the components of the EEG that arise in response to a stimulus (this stimulus could be electric, visual, auditory, etc.).
3) Measuring the "*Single neuron electrical behavior*", which measures the electrical activity of a single neuron using microelectrodes in order to build models of cell networks. This kind of measuring is usually not used in the clinical environments.

Generally, EEG signals are unpredictable in terms of amplitude, duration, or morphology. Thus, it is said that EEG signal is a stochastic process. It should be stressed that the EEG process is not necessarily a random process, but it may have such a high degree of complexity that only this description in statistical terms is definable. In addition, EEG signal can be assumed as a deterministic signal but made of many components that make the signal too complex [17]. In Figure 1.3 sample EEG signals from a healthy subject are shown.

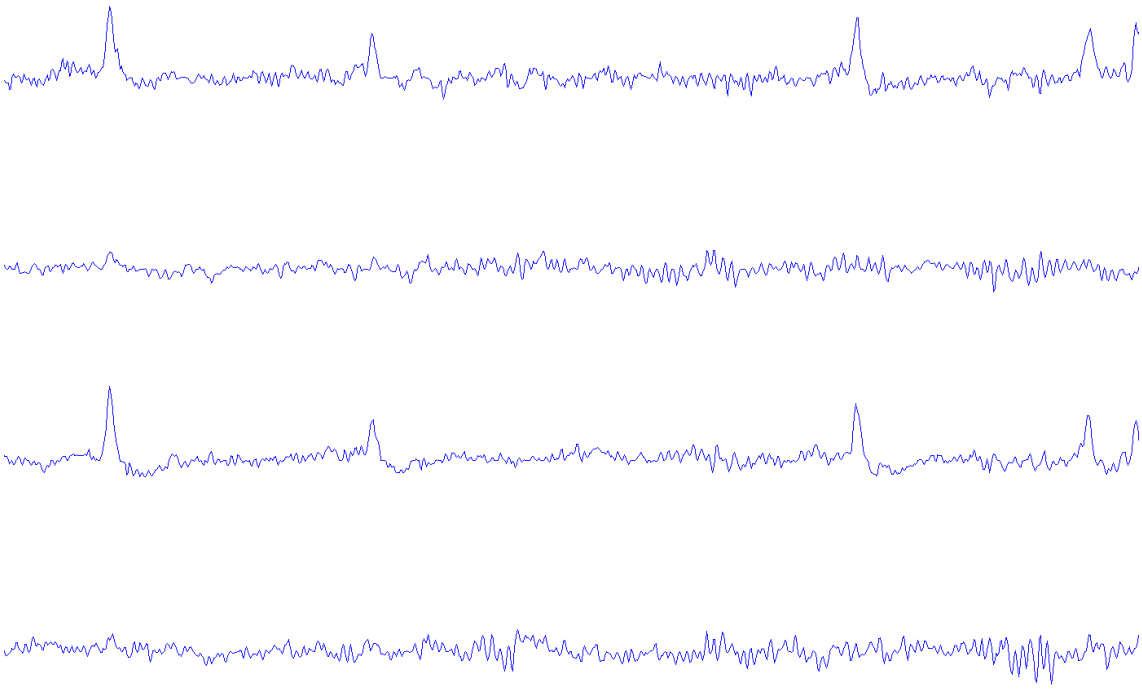

Figure 1.3. *Healthy subject EEG signals from 4 frontal and occipital channels*

1.2.2. History

The history of EEG has been a continuous process and vast developments in interpretation of EEG signals have been done. However in this section a brief history of EEG is given, which is summarised from [2], [18], and [19].

The pioneers who utilized the electrical signals of muscle nerves using a galvanometer and established the concept of neurophysiology are C. Matteucci and D. B. Rey-

mond. However, the concept of action current was introduced by H. V. Helmholz. R. Caton (1842-1926), an English physician, used a galvanometer and placed 2 electrodes on the scalp of a human subject and recorded the brain activity of the brain the first time in 1875, and the term electroencephalography has been used since then. He also described the negative variations of electric current in the grey matter while it is in a state of functional activity.

G. Fritsch (1838-1927) and J. E. Hitzig (1838-1907) discovered the concept of evoked potential. In 1877, V. Y. Danilevsky (1852-1939) studied both evoked potential and spontaneous electrical activity of the brain in animals. N. Cybulski (1854-1919) investigated the evidence of an epileptic seizure in a dog caused by electrical stimulation. The work of Fritsch and Hitzig was continued by D. Ferrier and G. F. Yeo in 1880, who performed electrical stimulation of the cerebrum in apes.

H. Berger (1873-1941) is well known by almost all electroencephalographers. His report in 1929 introduced the alpha rhythm. He also investigated the effect of hypoxia on the human brain and the sleep spindles. He was also interested in localizing brain tumours using EEG and finding a correlation between mental activities and changes in EEG signals. In 1938 he reported the advantages and disadvantages of the two types of electrodes he used for EEG recording: chloride silver needles and silver foil sheets. Berger's contribution to experimental and clinical EEG can be summarized as follows: 1) studies of the effects of the skull on EEG voltages; 2) development of needle cup, and plate electrodes; 3) development of electrode-restraining devices; 4) normative values for EEG background rhythms; 5) use of simultaneous EEG with EKG and blood pressure; 6) simultaneous surface and invasive recording; 7) investigations of EEG topography; 8) simultaneous EEG and movement recording in focal motor epilepsy; 9) use of signal processing to extract EEG parameters; 10) estimation of the Fourier transform of the EEG.

A. E. Kornmuller recognized the importance of using multi-channels for EEG recording to cover the wider brain area. J. F. Toennies (1902-1970) built the first ink-writing oscillograph, which was called a "*neurograph*". W. G. Walter discovered delta waves in EEG signals. In 1934, H. Davis observed the alpha rhythm. Hallowel and P. Davis were the earliest investigators of human sleep using EEG. A. L. Loomis, E. N. Harvey and G. A. Hobart studied human sleep EEG patterns mathematically for the first time.

E. D. Adrian (1889-1977) developed the method of recording single neuron action potentials. He used a capillary electrometer in conjunction with a vacuum tube amplifier. He collaborated with electrical engineer, B. Matthews, who introduced the use of differential input amplifiers to electrophysiology.

Around 1935 in North America, research on EEG began and rose to international fame with the works of H. Davis, F. A. Gibbs, and E. Gibbs at Harvard University. The most development in EEG in the 1930's was in instrumentation improvements. Nowadays, EEG signals are recorded with instruments which are equipped with many signal processing tools and also big memory space for long-term recordings. Also, EEG is

used with Magnetic Resonance Imaging (MIR) systems, simultaneously to obtain more information.

1.2.3. 10-20 Standard Electrode Positioning System

EEG signals are recorded using electrodes attached to the scalp. The electrodes most commonly are made of platinum, gold or silver-silver chloride. The choice of electrode depends on the cost and quality. The most commonly used electrode is silver-silver chloride. These electrodes have acceptable qualities with minimal drift of electrical potentials and a long Time Constant (TC). In order to decrease the impedance between the scalp and electrodes an electrolyte (conductive jelly or electrolyte gels) is used between the scalp and electrodes. These electrolyte gels generally contain Cl as the principal anion for better conductivity. High impedance of electrodes tends to produce artefacts with slight movement of the body or even electrode wires. Impedance of less than $5\ k\Omega$ is the best. However, impedance of less than $8\ k\Omega$ is tolerable. Modern instruments are equipped with methods of checking impedance.

A common artefact, which is named "*electrode pop*", can be caused by not enough volume of electrolyte or by unstable electrode surface contact with the skin. It should be noted that this artefact is not related to the impedance of electrodes, therefore it still may occur even when the impedances are low.

There are many special kinds of electrodes for different purposes such as subdermal needle electrodes and nasophqryngeal electrodes. Also, for long-term EEG monitoring, especially for detection of a seizure in patients with intractable epilepsy who qualify for surgical treatment, sphenoidal, foramen ovale, tympanic, ethmoidal, depth, and subdural electrodes are used. These different kinds of electrodes make the interpretation of EEG more accurate [20].

For clinical purposes, in order to ensure standardized reproducibility (i.e., the subject can be compared over time and also compared with other subjects), positions of the electrodes on the scalp follow international standard placement, which is called the 10-20 standard system for electrode placement. In Figure 1.4 the locations of these electrodes are shown.

As can be seen in Figure 1.4, each electrode indicates the general area (F- frontal, C- central, P- parietal, T- temporal, O- occipital, A- earlobes) and the electrodes in the left hemisphere have odd numbers, and the electrodes in the right hemisphere have even numbers. 10-20 stands for the percentage of the distance between neighboring electrodes relative to the distance between the beginning and end of a row [21].

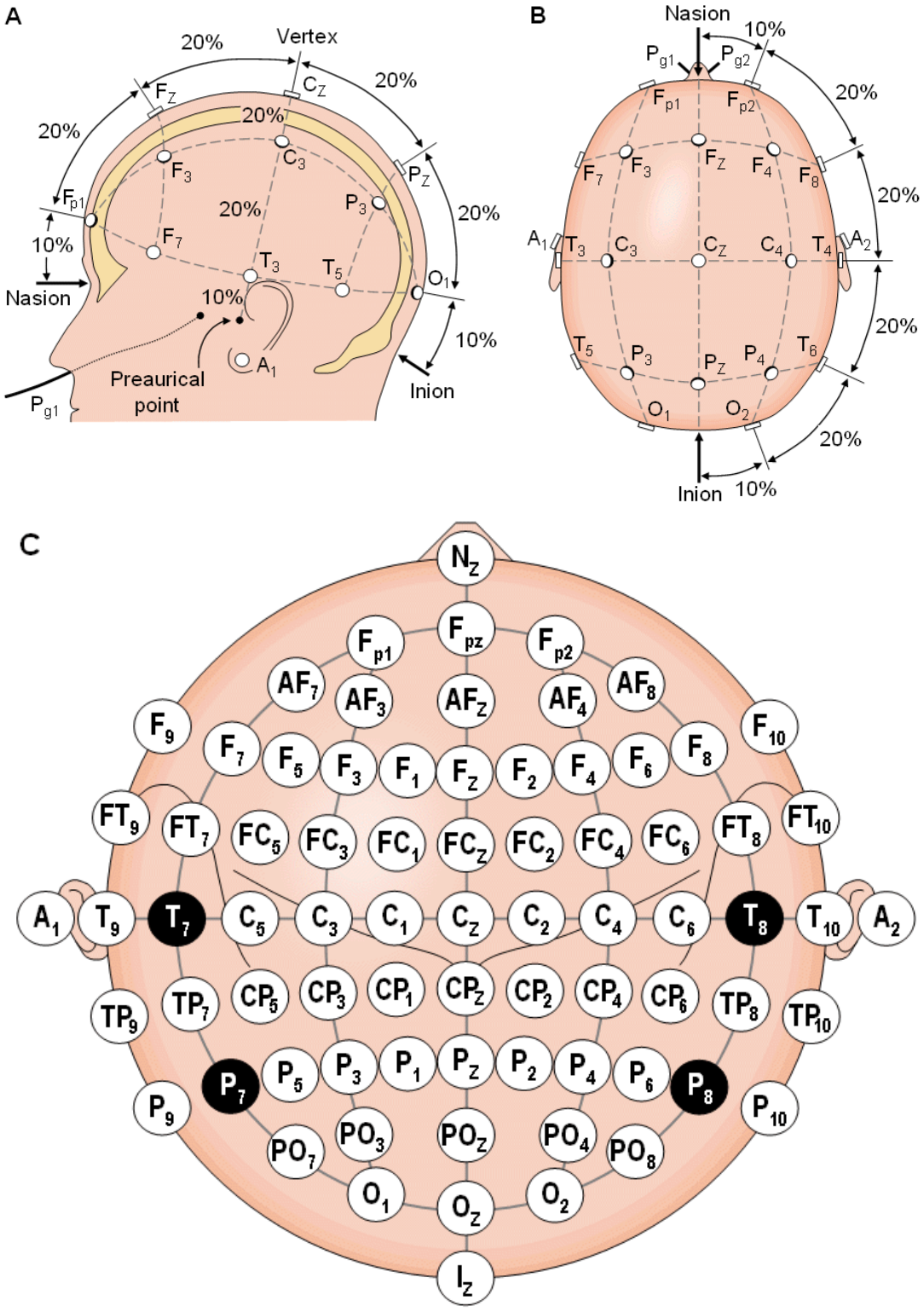

Figure 1.4. *10-20 standard system seen from (A) left and (B) above the head and (C) location and nomenclature of the intermediate 10% electrodes* [15].

1.2.4. EEG Application to Epilepsy

In recent years numerous studies of epilepsy have been done using EEG signals with different approaches. In this sub-section some recent studies are reviewed and compared to our study.

In 2005, Kannathal et al. [22] used entropy estimators (spectral entropies, Renyi's entropy, state space reconstruction, Kolmogorov-Sinai entropy, and approximate entropy) in order to detect epilepsy in EEG. The EEG data used for the study was obtained from Bonn University. The dataset contains 2 sets of five normal subjects and five epileptic patients. From each set, about 30 single channel EEG segments, where the duration of each segment was 23.6 seconds, were selected manually in order to finally have noise-free segments. The entropy measures are extracted from each segment and in order to distinguish their significance t-test was used. In addition, an adaptive neuro-fuzzy classifier was used in order to classify the normal and epileptic segments. The correct classification percentage of the normal class was 93.02% with 60 segments for training and 43 segments for testing, and the correct classification percentage of the epileptic class was 91.49% with 60 segments for training and 47 segments for testing. Since the noise free segments were visually selected, this method is not feasible for automatic seizure detection.

Mohseni et al. [23] compared different feature-based seizure detection methods. The dataset contains 3 sets of data (A, D, E). Each set (A and B) contains the EEG signals of five healthy and five epileptic patients, respectively. These 2 sets (i.e., A and B) contain no seizure occurrences while set E contains seizure attacks. In total, the used dataset has 100 segments of EEG signal from both healthy and epileptic subjects. Each segment is formed by 256 discrete data. Specificity and sensitivity of 97.38% and 96.13% were obtained using a recurrent neural network. Lyapunov exponent-based features were fed into the classifier as inputs [24]. In [25], using wavelet coefficients (D3, D5 and A5) of EEG signals as an input of multilayer perceptron neural network, with the Levenberg-Marquardt algorithm the sensitivity and specificity of 92.8% and 92.3% were obtained, respectively (60% of data for training and the rest 40% for testing). In this study, four channels of F7-C3, F8-C4, T5-O1, and T6-O2 were used. In addition, in [26] Mohseni correctly obtained 98.25% classification (with 50% training and 50% testing) using features based on pseudo Winger-Ville distribution and a feed-forward back propagation neural network as classifier.

Greene et al. [27] evaluated 21 features and their different combinations, which were extracted from 2 seconds artefact free segments. The classification using linear discriminant classifier was performed on 17 neonates with 9 channels, and 81.08% sensitivity and 82.23% specificity were achieved for the combined features. Kuhlmann et al. [28] used a combination of relative average amplitude, relative power, and coefficients of variation of amplitude as better-performing features for seizure detection purpose. They obtained 81% average sensitivity for the combination. The analyses were performed on 21 patients with 525 hours EEG recording and 88 clinical seizure occur-

rences in total. The classifier-based framework developed by Saab and Gotman [29] was employed and 81% sensitivity obtained (train set contains 367 hours of 14 subjects with 58 seizures, and test set contains 158 hours of 7 subjects with 30 seizures).

In 2012, Pan et al [30] used a feature selection method based on mutual information, selecting the relevant features among different features (median absolute deviation, Wavelet coefficients, frequency regularity, spatiotemporal correlation, power, entropy, and hybrid). Support vector machine was used as a classifier in this study and average sensitivity of 88.99% and average specificity of 93.82% were acquired. The dataset contains 7 subjects with 43 seizure occurrences. Features are extracted from 2 second time windows with 1 second overlap. For each patient, the entire sessions are divided into sub-sessions based on the number of seizure events. Each sub-session starts from the starting point of seizure and ends at the starting point of the next seizure occurrence. It should be noted that in this study the performance of classification was evaluated based on leaving one session out as cross validation, which means each time one sub-session was kept as a test dataset and the rest of the sub-sessions were used as a training dataset. In essence, this study only detects the onset of seizure events since each sub-session contains one seizure and the dataset was classified based on sub-sessions.

Few studies have been analysed on CHB-MIT Scalp benchmark EEG dataset [31]. Shoeb and Guttag [32] used 916 hours of EEG signals from 23 patients. They measured the energy of each 2 seconds epoch, which was passed through a filterbank that spans the frequency range 0.5-25 Hz, as feature. They also used SVM with RBF kernel (kernel parameter γ=0.1 and error C=1) for classification purpose. First 20 seconds of seizure segments and all nonseizure segments were used for training. In order to evaluate the performance of the designed detector 2 different approaches were used. To compute the sensitivity they used N_{Ns} nonseizure records of the patient EEG where N_{Ns} is the number of 1-hour records without seizure events, and they used $N_s - 1$ seizure records where N_s is the number of 1-hour records with at least one seizure event. They repeated the training task N_s times so that each seizure record is tested. To estimate the specificity they used in the opposite way, i.e., they used N_s seizure and $N_{Ns} - 1$ nonseizure records for training, and only one nonseizure record for testing. This process was then repeated N_{Ns} times. An average of 96% sensitivity with the mean latency of 4.6 seconds was reported. In other words, they detect 96% of 173 seizure events (166 seizure events detected) and 7 seizure segments were entirely missed. The false detection rate varies dramatically from patient to patient, which its median value was 2 per 24 hours (i.e., 2 nonseizure segments among 24 hours were detected as seizure events).

In [33] Khan et al. proposed a framework to detect the seizure onsets. For this purpose, they extracted six features from each 1 second epoch, which were kurtosis and skewness of the raw EEG, relative energy and coefficient of variation of Daubechies 4 coefficients (A5 0-4 Hz, D5 4-8 Hz, and D3 16-32 Hz). In order to normalize the epoch features they took a 25 seconds window as a background window and considered 15 seconds between each epoch and the background window to prevent seizure onset into the background. They applied the algorithm only to the first 10 patients of CHB-MIT

Scalp benchmark EEG dataset. For classification purpose, they used a linear classifier. They classified the features for each patient separately and used 80% of seizures for training and the other 20% for testing, and repeated this process until every seizure got tested. Also in order to eliminate the artifact and noise, they declared the seizure in each epoch if it was present in at least 60% of the channels. They obtained mean latency of 3.2 and reported 100% sensitivity.

In all these perior works the majority of data (e.g., 80%) is used as training, which is not feaseable in medical environment. Also in those studies that detect the only the onset of seizure events the sensitivity rate indicates the rate of correctly detected onsets of the seizure segments without any indication of the seizure duration. In most cases, the appearance of EEG signals at the onset of a seizure occurrence changes significantly, which simplifies seizure onset detection. For instance, in Figure 1.5 the onset of seizure arrives with an increment of amplitude of signal and decrement in frequency in two channels of F7-T7 and T7-P7. As can be seen, This is the reason that makes onset seizure detection an easier task in contrast to seizure segment detection.

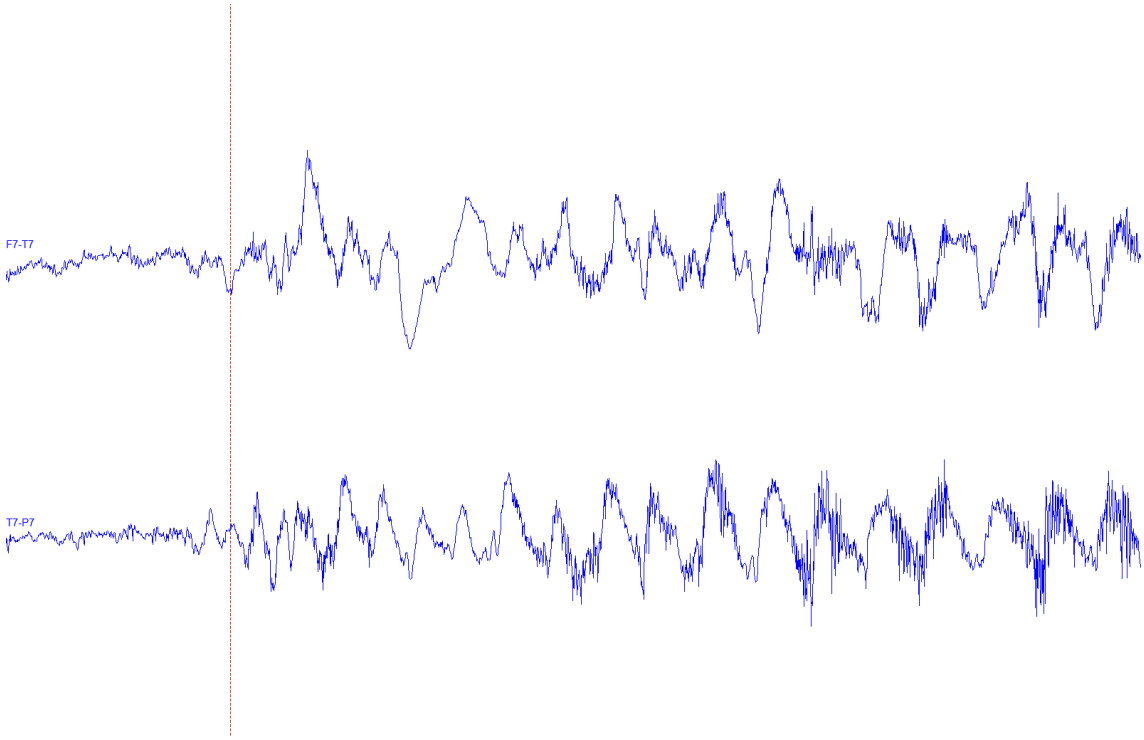

Figure 1.5. *The onset of seizure on two channels of F7-T7 and T7-P7is shown with red line (recording chb03-04)*

1.3. Thesis Outline

This thesis is organized as follows: In Section 2 the information of the benchmark EEG dataset is provided also different feature extraction techniques are described. In Section 3, several state-of-the-art feature selection methods are introduced. In Section 4, the proposed EEG classification method is presented. Finally, in Sections 5 and 6, the experimental results are presented and discussed with conclusive remarks.

2. EEG DATA PROCESSING

2.1. The Benchmark EEG Dataset

The benchmark EEG dataset recorded at Children's Hospital, Boston consists of EEG recordings of pediatric subjects with intractable seizures. The subjects were monitored for several days in order to characterize their seizures and assess their candidacy for surgical treatment [31].

The EEG dataset is collected from 22 subjects (5 males, ages 3-22; and 17 females, ages 1.5-19). One subject is recorded in two different time periods, thus there are overall 23 cases. The EEG recordings are in "*edf*" files. EDF format stands from European Data Format, which is a standard file format designed for the exchange and storage of medical time series.

The sampling frequency of signals is $256\,Hz$ with 16-bit resolution. Most of the cases have around 23 EEG recordings. The 10-20 international system and nomenclature was used for the recordings. In addition to EEG signals, a few records contain ECG signal and Vegan Nerve Stimulus (VNS) signal. In addition, all recordings have labels for each second to show whether it is a seizure or non-seizure segment.

In this study, four subjects with overall 21 hours of recordings (subject 1 and subject 3 with 6, subject 5 with 5, and subject 8 with 4 recording hours) are studied. The subjects are selected such that EEG recordings contain at least 1 seizure occurrence in each hour. The 18 processed channels are, FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ and CZ-PZ.

2.2. Feature Extraction Methods

In this section, time domain, frequency domain, time-frequency domain and nonlinear as four typical feature extraction methods are used in this study. In addition, Cepstral features are proposed as new features in EEG signal processing. The signal is windowed with non-overlapping rectangular windows of 1 second (256 samples of discrete data). In the time-frequency domain the Hamming window is used.

In the very first step, the EEG signal of each channel is band-passed filtered between 0.5 and 30 Hz using a linear phase FIR filter with the Parks-McClellan algorithm. This frequency band has been used in most seizure detection studies [34], [35]. The artefacts have not been removed in order to evaluate the performance of features and the proposed method. All the processing in this section is done by MATLAB version 7.13. All the extracted features according to their types are listed in Table 2.1.

Table 2.1. *List of the extracted features*

Morphological	LAT, LAR, AAMP, ALAR, PAR, NAR, TAR, ATAR, TAAR, AASS, PP, PPT, PPS, ZC, ZCD, SSA
Time	Skewness, Kurtosis, NO. maxima and minima, Mean, Variance, Standard deviation, Coefficient of variation, RMS, Shannon entropy, Approximate entropy, Energy; Standard variation, Variance and Energy of auto-covariance
Frequency	Maximum, Minimum, and Mean of Power Spectrum, Spectral entropy, Median frequency
Time-Frequency	Relative scale energy, Shannon entropy, Coefficient of variation of 5 Approximations and 5 Details coefficients with db1, db2, db3, and db4, Frequency regularity index, Maximum, Minimum, Variance, Mean, Standard deviation, No. of exterma and energy of 5 Approximations and 5 Details using db4, Energy of STFT in 4 frequency bands of Delta (1-4 Hz), Theta (4-7), Alpha (7-13 Hz), and Beta (13-30 Hz), Energy in Winger-Ville distribution in 3 frequency bands of Delta (1-4 Hz), Theta (4-7) , Alpha (7-13 Hz), and Beta (13-30 Hz)
Non-Linear	Average of Lyapunov exponents
Cepstral	MFCCs, first and second order derivative of MFCC coefficients in 5 frequency bands of 1-4 Hz, 4-8 Hz, 8-12 Hz, 12-20 Hz

2.2.1. Morphological Features

Sixteen morphological features are extracted from each 1 second non-overlapping window, which were used by Kalatzis et al. previously [36]. These features are expressed as follows:

Latency (LAT): the time where the maximum value of the signal s appears:

$$t_{s_{max}} = \{t | s(t) = s_{max}\}. \tag{2.1}$$

Latency/Amplitude Ratio (LAR): the ratio of the latency to the maximum signal value:

$$LAR = \frac{t_{s_{max}}}{s_{max}}, \tag{2.2}$$

where s_{max} is equal to $\max\{s(t)\}$.

Absolute Amplitude (AAMP): the absolute value of the signal and is equal to

$$AAMP = |s_{max}|. \tag{2.3}$$

Absolute Latency/Amplitude Ratio (ALAR):

$$ALAR = \left|\frac{t_{s_{max}}}{s_{max}}\right| = |LAR|. \tag{2.4}$$

Positive Area (PAR): the sum of positive signal values:

$$PAR = \sum_t 0.5(s(t) + |s(t)|). \quad (2.5)$$

Negative Area (NAR): the sum of negative signal values:

$$NAR = \sum_t 0.5(s(t) - |s(t)|). \quad (2.6)$$

Total Area (TAR): the sum of negative and positive values of signal:

$$TAR = PAR + NAR. \quad (2.7)$$

Absolute Total Area (ATAR): the absolute of the total area value:

$$ATAR = |TAR|. \quad (2.8)$$

Total Absolute Area (TAAR):

$$TAAR = PAR + |NAR|. \quad (2.9)$$

Average Absolute Signal Slope (AASS):

$$AASS = \frac{1}{n}\sum_t \frac{1}{\tau}|s(t+\tau) - s(t)|, \quad (2.10)$$

where n is the number of samples of the digital signal in each time frame (256 samples in 1 second), and τ is the sampling interval of the signal, which is equal to one in this study.

Peak-to-Peak (PP): the difference between the maximum and minimum signal values

$$PP = s_{max} - s_{min} \quad (2.11)$$

Peak-to-Peak Time window (PPT):

$$PPT = t_{s_{max}} - t_{s_{min}} \quad (2.12)$$

Peak-to-Peak Slope (PPS):

$$PPS = \frac{PP}{PPT} \quad (2.13)$$

Zero Crossings (ZC): the number of times that $s(t) = 0$, in the peak-to-peak time window

$$ZC = \sum_{t=t_{s_{min}}}^{t_{s_{max}}} \delta_s \tag{2.14}$$

Zero Crossing Density (ZCD): zero crossing per time unit, in the peak-to-peak time window

$$ZCD = \frac{ZC}{PPT} \tag{2.15}$$

Slope Sign Alterations (SSA): the number of slope sign alterations of two adjacent points of signal:

$$SSA = \sum_{t} 0.5|\frac{s(t-\tau)-s(t)}{|s(t-\tau)-s(t)|} + \frac{s(t+\tau)-s(t)}{|s(t+\tau)-s(t)|}| \tag{2.16}$$

where τ is the sampling interval of the signal.

2.2.2. Time domain Features

Statistical features are used as typical features since they are easy to implement, have a lower computational burden and show significant properties of EEG signals. In this practice, skewness [37], kurtosis [37] [38], number of maxima and minima [37] [39], mean [40], variance [40] [41], standard deviation [38] and coefficient of variation [40] of each non-overlapping moving window with a length of 1 second as seven statistical features are extracted. In addition, root mean square amplitude [37] the energy of the signal [38] [41], Shannon entropy [38] [39], and approximate entropy [39] are extracted from each time frame as features in the time domain.

Skewness: Symmetry is an important parameter of the distribution. According to the definition, the mean of a skewed variable is not located at the center of distribution. It should be noted that two signal segments could have the same mean and standard deviation but different values for skewness. Skewness is expressed as:

$$s = \frac{E(x-\mu)^3}{\sigma^3}, \tag{2.17}$$

where μ , x and σ are the mean, signal and the standard deviation of x, respectively. Also $E(f)$ is the expected value of f. This equation can be extended as follows:

$$s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2})^3}, \tag{2.18}$$

Negative and positive values of s indicate that the signal is skewed left or right, respectively.

Kurtosis: The degree of "*peakedness*" of a distribution is presented by Kurtosis and formulated as follows:

$$k = \frac{E(x-\mu)^4}{\sigma^4}, \tag{2.19}$$

where μ , x and σ are the mean, the signal and the standard deviation of x, respectively. Equation (2.19) can be extended as follows:

$$s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}. \tag{2.20}$$

Number of maxima and minima: This is the number of maxima and minima of each 1 second, non-overlapping time frame.

Mean:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.21}$$

where x is the signal and n shows the number of samples in each time window ($n = 256$).

Variance: Variance measures how far a sample in the signal is from the mean value, and defined as:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{2.22}$$

Standard deviation:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{2.23}$$

Coefficient of variation: Coefficient of variation is the ratio of standard deviation to the mean of the signal and defined as:

$$CV = (\frac{standard\ deviation}{mean})^2 = (\frac{\sigma}{\mu})^2. \tag{2.24}$$

Root mean square amplitude: The root mean square amplitude (RMS) of a signal is defined as the averaged amplitude of signal in a given interval and is expressed as:

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}, \tag{2.25}$$

where n is the number of sample in the determined period, and x_i is the i[th] sample of the signal.

Energy: The energy of the signal in the time domain from non-overlapped windows with length of 1 second is computed as a feature in this study. Since often seizure occurrence increases the energy of the signal during its onset, it is typical to use energy for seizure detection, and defined as follows:

$$E = \sum_{t} |s(n)|^2. \tag{2.26}$$

Shannon entropy: Shannon entropy determines the expected value of the included information in a message in communication field and was introduced by Shannon. In essence, entropy is a measure of uncertainty of a signal. This concept has been used in several studies in epilepsy detection related fields through EEG processing.

$$ShEn = \frac{H_{sh}}{\log k}, \tag{2.27}$$

where

$$H_{sh} = -\sum_{i} p_i \log p_i, \tag{2.28}$$

where p_i is the probability density function of the original signal. In Equation (2.28), for normalization, H_{sh} should be divided by $\log k$, where k is the number of bins that the original signal is divided into (see Section 3.1.1).

Approximate entropy: Approximate entropy was introduced by S. M. Pincus to remove the limitation of other entropy measures. Approximate entropy is a method to quantify the amount of regularity or randomness of a signal; for instance, for comparing two sets of data with equal component numbers and values, which one of them is regular (i.e., periodic) and the other irregular. In this case, statistics features such as mean and variance are not able to detect the difference between these two sets of data while approximate entropy can distinguish the difference. It is defined as:

$$ApEn\left(m, r_f, N\right) = \Phi^m\left(r_f\right) - \Phi^{m+1}\left(r_f\right), \tag{2.29}$$

Where $\Phi^m\left(r_f\right)$ is expressed as:

$$\Phi^m\left(r_f\right) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(r_f), \tag{2.30}$$

and $C_i^m(r_f)$ is equal to:

$$C_i^m(r_f) = \frac{number\ of\ such\ j\ that\ d\ [x_m(i), x_m(j)] \leq r_f}{N - m + 1}, \tag{2.31}$$

where m and r_f are positive integer and real number, which represent the length of compared data and filtering level, respectively. N is the number of samples involved. Parameter d is the distance between vectors $x_m(i)$ and $x_m(j)$ and determined as:

$$d\ [x_m(i), x_m(j)] = \max_{k=1,2,\dots,m}(|s(i+k-1) - s(j+k-1)|), \tag{2.32}$$

where $x_m(i)$ is equal to:

$$x_m(i) = \{s(i), s(i+1), \dots, s(i+m-1)\};\ \ 1 \leq i\ \leq N - m + 1. \tag{2.33}$$

Whenever the $ApEn$ is near to zero, and has a small value, it shows that it is regular and predictable. The signals with higher $ApEn$ are likely to have noise contamination. In this study, $ApEn$ of each moving non-overlapping window with a length of 1 second is computed as a feature. Also, according to reference [42] $m = 2$ and $r_f = 0.2\ SD$.

Standard variation, Variance and Energy of auto-covariance: standard variation, variance and energy of the auto-covariance of each segment. The cross-covariance of two matrices of x and y is defined as:

$$cov(x, y) = E[(X - \mu_X)(Y - \mu_Y)^*], \tag{2.34}$$

where μ_X and μ_Y are the expected values of x and y and $*$ denotes the complex conjugate. If x is equal to y, then it is named auto-covariance of signal x. In other words, auto-covariance is the covariance of signals x against to the time-shifted version of the signal. Auto-covariance is expressed as follows:

$$C_{XX}(t, s) = E[(X_t - \mu_t)(X_s - \mu_s)] = E[(X_t X_s)] - \mu_t \mu_s. \tag{2.35}$$

2.2.3. Frequency Domain Features

Sometimes the frequency domain is more understandable than the time-domain representation of EEG signals. However, the frequency domain for electroneurophysiologists is mostly defined as traditional EEG waves such as alpha, beta, theta and gamma. In this sub-section the maximum, minimum, and mean of power spectral density, spectral entropy [38] [39], and median frequency [38] as the most common features from frequency-domain are extracted.

Power Spectral Density: There are several approaches in order to obtain power spectral density, or simply power spectrum. However, the direct approach is the magnitude squared of the Fourier transform of the interested signal, and is equal to:

$$PS(f) = |X(f)|^2, \tag{2.36}$$

where $X(f)$ is the Fast Fourier Transform (FFT) of the original signal,

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{(-j2\pi fn/N)}, \tag{2.37}$$

where N is the total number of points and f shows the family member. The maximum, minimum, and mean of the power spectrum are calculated in this study as three features.
Spectral entropy: Spectral entropy is defined as the same as Shannon entropy with this difference that P_i is the power density of the power spectrum.

$$H_{Sp} = -\sum_{i=f_l}^{f_h} P_i \log P_i, \tag{2.38}$$

where f_l and f_h are the low and high frequency bands. For normalization purpose, the Spectral entropy is defined as:

$$SpEn = \frac{H_{sp}}{\log N_f}, \tag{2.39}$$

where N_f is the number of frequency components between f_l and f_h.
Median frequency: Median frequency is defined as the particular frequency, which divides the total area under $PS(f)$ into two parts of equal size,

$$\int_0^{\omega_{MDF}} PS(f)d\omega = \int_{\omega_{MDF}}^{\pi} PS(f)d\omega, \tag{2.40}$$

where ω_{MDF} is median frequency.

2.2.4. Time-frequency Features

In time-frequency analysis both the time and frequency domains are studied at the same time. Therefore, time-frequency space can bring more valuable presentation in contrast to traditional interpretations. In this subsection, different time-frequency distributions based features are extracted, which are listed in the following.
Short Time Fourier Transform Based Features: The idea of the Short Time Fourier Transform (STFT) originates from the Fourier transform. In STFT, the Fourier transform is restricted to a fixed time interval, which also tries to provide information about

the time domain at the same time and cover the Fourier transform defect. This is defined as:

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-j\omega n}, \tag{2.41}$$

where x is the digital signal, and ω is the basis (window) function. Since sharp cut off causes artificial discontinuities and creates undesired problems, Hamming, as a smooth cut off function (window), is chosen with a length of 1 second in this study. There are some limitations in STFT, which are caused by the fixed length of the window. The narrower the window in the time domain the more resolution is obtained, while in the frequency domain accuracy is missed. In contrast, the wider the window in the time domain the more resolution in the frequency domain is obtained, while less information will be provided from the time domain. The following feature is extracted using STFT.

a) ***Energy of STFT in Four Frequency Bands of Delta (1-4 Hz), Theta (4-7), Alpha (7-13 Hz), and Beta (13-30 Hz):*** The energy of four frequency bands of Delta, Theta, Alpha, and Beta are calculated in 1 second epoch length.

Wigner-Ville Distribution: Wigner-Ville is a non-stationary analysis tool, the same as STFT. The Winger-Ville distribution of a signal $z(t)$ is defined as:

$$W_z(t,\omega) = \int_{-\infty}^{\infty} z\left(t+\frac{\tau}{2}\right) z^*\left(t-\frac{\tau}{2}\right) e^{-i\omega\tau} d\tau, \tag{2.42}$$

Equation (2.37) shows that the expression of Wigner-Ville distribution resembles the autocorrelation, but is not exactly the same. This kind of autocorrelation is named *instantaneous autocorrelation*:

$$k_z(t,\tau) = z\left(t+\frac{\tau}{2}\right) z^*\left(t-\frac{\tau}{2}\right), \tag{2.43}$$

where τ is the time lag and $*$ represents the complex conjugate of the signal z [43]. In essence, the Winger-Ville distribution is the Fourier transform of the instantaneous autocorrelation. In other words, it is the power spectrum calculation using the mentioned autocorrelation function.

In contrast to STFT, the window in Winger-Ville distribution is usually a shifted version of the signal itself, which compares the information of the signal with its own information at other times and frequencies.

a) ***Energy in Wigner-Ville Distribution in Four Frequency Bands of Delta (1-4 Hz), Theta (4-7), Alpha (7-13 Hz), and Beta (13-30 Hz):*** The energy of the four frequency bands of Delta, Theta, Alpha, and Beta are extracted from the Wigner-Ville distribution.

Discrete Wavelet based features: Wavelets are a type of waveform of limited duration with zero mean and nonzero norm. These waves tend to be irregular and asymmetric. In

wavelet analysis, the signal is decomposed into shifted and scaled versions of the original wavelet. In essence, as in Fourier transform, wavelet transform summarizes the correlation between the signal and some basic functions with certain physical properties (e.g., frequency, scale or position). Discrete Wavelet Transform (DWT) is an orthogonal function and similar to Discrete Fourier Transform (DFT), whereas the basis functions in DWT are a set of functions which are defined by a recursive difference equation, while in DFT the basis function is a sinusoid.

In DWT digital filtering is used due to a time-scale representation of the digital signal. In other words, the signal is passed through filters with different cut off frequencies at different scales, where the resolution of the signal (i.e., amount of detail information in the signal) is determined by the filtering operations, and the scale is defined by down-sampling and up-sampling operations. Referring to the Mallat algorithm, the digital signal in the time domain is passed through low-pass and high-pass filters, successively as shown in Figure 2.1 As can be seen in Figure 2.1, the discrete signal $x[n]$ is passed through low-pass filter G_0, and high-pass filter H_0. At each decomposition level, the high-pass filter produces detailed information, $d[n]$, and approximations, $a[n]$, are produced by low-pass filter.

The maximum number of decomposition levels depends on the length of the signal. The DWT of original signal $x[n]$ is obtained by concatenating all the $a[n]$ and $d[n]$ coefficients from the last up to the first level of decomposition.

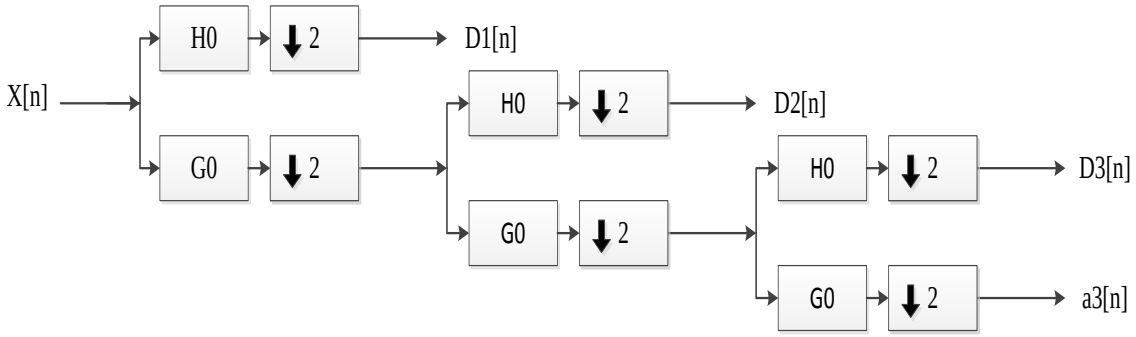

Figure 2.1. *Three-level wavelet decomposition tree*

There are various numbers of basis functions (mother wavelets) that are used for wavelet transformation. Haar, Daubechies, Coiflet, Symlet, Meyer, Morlet and Mexican Hat are some wavelet family instances of mother wavelets. Thus, the appropriate mother wavelet must be chosen according to the particular application [43].

In this study, several features are extracted from the coefficients of details and approximations that are produced by Daubiches 1 (db1), Daubiches 2 (db2), Daubiches 3 (db3), and Daubiches 4 (db4) as four wavelet transformations.

a) Shannon Entropy, Coefficient of Variation of 5 Approximations and Details Coefficients with db1, db2, db3, and db4: Two features of Shannon Entropy and Coefficient of Variation are extracted from both approximation and detail coefficients [23].

b) Maximum, Minimum, Variance, Mean, Standard Deviation, No. of Exterma and Energy of Five Approximations and Details Using db4: Maximum, minimum, var-

iance, mean, standard variation, number of exterma and energy as 7 statistical features are extracted from both detailed and approximation coefficients using the db4 mother wavelet [44].

c) ***Relative Scale Energy:*** Relative scale energy is the ratio of the energy of coefficients in the given scale to the energy of the wavelet coefficients in all scales. It is a measure of rhythmicity and defined as follows:

$$e_r(i) = \frac{e(i)}{\sum_{j=1}^{M} e(j)}, \tag{2.44}$$

where M is the number of wavelet bands, and $e(i)$ is the energy of ith band and equal to

$$e(i) = \sum_{k=1}^{N_i} D_{ik}^2 \frac{\Delta t}{N_i}, \tag{2.45}$$

where N_i is the number of wavelet coefficients in band i, and D_{ik} are the coefficient's values in band i. Δt is 1 second window length in this study [45].

d) ***Frequency Regularity Index:*** The frequency regularity index is expressed as follows:

$$R_i = \max_{\tau} \left| \frac{\int_{-\infty}^{\infty} s(t+\tau) y(t) dt}{\sqrt{\int_{-\infty}^{\infty} x^2(t) dt \times \int_{-\infty}^{\infty} y^2(t) dt}} \right|, \tag{2.46}$$

where $y(t)$ is 1 second EEG signal and $x(t)$ is expressed as follows:

$$x(t) = \sin(2\pi f_c t), \tag{2.47}$$

where f_c is the central frequency of the band. In other words, R_i is the maximum normalized cross-correlation between $y(t)$ and $x(t)$. The higher the value of R_i the more similarity between $y(t)$ and $x(t)$, which also can be interpreted as concentration of power of $y(t)$ in a narrow range near central frequency. Similar to relative scale energy, the frequency regularity index represents rhythmic synchronization [30].

The wavelet packet method is used in order to calculate the frequency regularity index. The wavelet packet method is a generalization of wavelet decomposition, which provides a greater range of possibilities for further signal analysis. In wavelet transform, the signal is split into an approximation and detail. In the next decomposition level, the approximation itself is split into approximation and detail and this process continues until the desired decomposition level is reached. Using the wavelet packet method, not only the approximation, but the detail coefficients can be split as well.

2.2.5. Nonlinear features

Lyapunov exponents measure the Sensitive Dependence on Initial Conditions (SDIC). SDIC means trajectories that start arbitrarily near to each other will separate exponentially fast [46]. The Lyapunov exponent is calculated as:

$$\lambda = \lim_{\substack{t\to\infty \\ |\Delta X_0|\to 0}} \frac{1}{t} ln \frac{|\Delta X(X_0, t)|}{|\Delta X_0|}, \tag{2.48}$$

where X_0 and $X_0 + \Delta X_0$ are two points in a space, and $\Delta X(X_0, t)$ is the function of separation of the two orbits that are generated by X_0 and $X_0 + \Delta X_0$. In this study the average of Lyapunov exponents is calculated for each time segment.

2.2.6. Cepstral Features

In addition to such traditional features we also extract Mel Frequency Cepstral Coefficient (MFCC). In [47], MFCC was used to detect robust emotion in EEG signal. MFCC is a representation of the power spectrum based on a linear cosine transform. In order to calculate the MFCCs, First the incoming frames are Hamming windowed in order to enhance the harmonic nature. In addition, Hamming window can reduce the effects of discontinuities and edges that are introduced during the framing process. Especially in logarithmic domain, the windowing effects can be encountered significantly. In order to perform filtering in the time domain, the frame is zero-padded to get the size as a power of 2 and then FFT is applied to get into the spectral domain for plain multiplication with the filterbank. The mel (melody) scaled filterbank is a series of filterbank, which has the central frequencies uniformly distributed in mel-frequency (mel(f)) domain where,

$$mel\,(f) = m_f = 1127 \log\left(1 + \frac{f}{700}\right) and\ \ f = 700\left(e^{\frac{m_f}{1127}} - 1\right). \tag{2.49}$$

Once the filtering is applied, the energy is calculated per band and Cepstral Transform is applied on the band energy values. Cepstral Transform is a discrete cosine transform of log filterbank amplitudes:

$$c_i = (\frac{2}{\rho})^{\frac{1}{2}} \sum_{j=1}^{\rho} \log m_j \cos\left(\frac{\pi . i}{N}(j - 0.5)\right), \tag{2.50}$$

where $0 < i \leq \rho$ and ρ is the number of filter banks. A subset of c_i is then used as the feature vector for this frame. In this study, in addition to the six MFCCs, the first and second order derivative of MFCCs in 5 frequency bands of 1-4 Hz, 4-8 Hz, 8-12 Hz, 12-20 Hz are calculated.

2.2.7. Normalization and Feature Smoothing

After extracting the features from each of the 18 channels, all feature vectors were normalized between -1 and 1. Moreover, in order to smoothen the features and enhance

the discrimination between the seizure and non-seizure segments we applied a moving average filter with a 20 seconds window and 1 second overlapping. In Figure 2.2 the energy of the Theta band (4-7 HZ) is shown before and after using the moving average filter. It is fairly obvious that after using the moving average filter the discrimination between the seizure and non-seizure segment is enhanced and the noise level in the non-seizure segment is significantly reduced. Note that in this example the seizure starts from the time frame 2996 and ends at 3036 seconds. Also in Figure 2.3 the enhancement of smoothing for variance feature can be seen obviously (in this example the seizure starts from the time frame 417 and ends at 532 seconds).

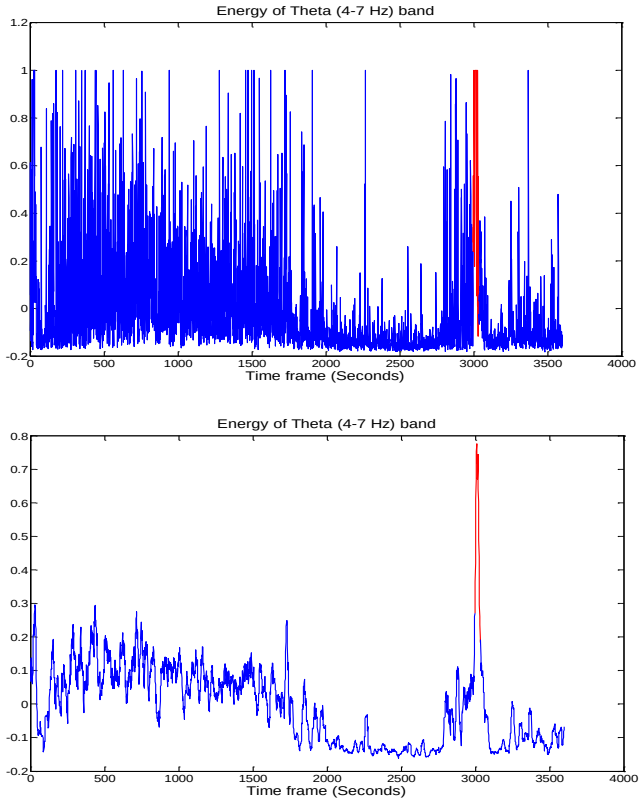

Figure 2.2 *Before (top) and after (bottom) using the moving average filter. The seizure segment is shown by red color.*

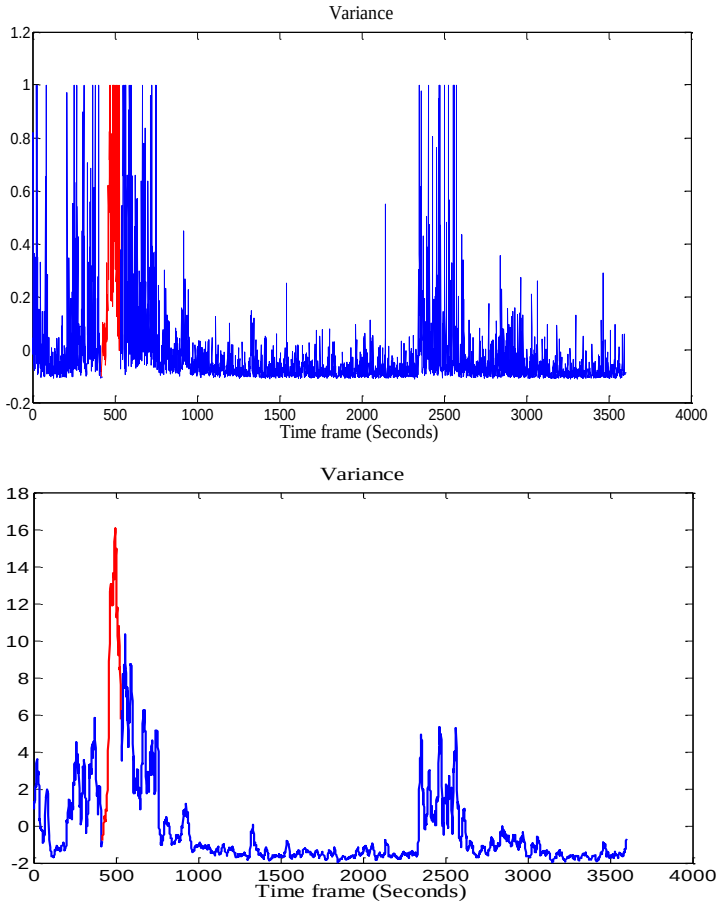

Figure 2.3. *Before (top) and after (bottom) using the moving average filter. The seizure segment is shown by red color.*

3. FEATURE SELECTION

3.1. Introduction

In many problems there is no possibility to access the predefined rules for an event, and therefore the task is to define the rules by relatively reasonable number of observations, which are named "*training*". In essence, using merely training data associated with labeled outcomes in order to find a relationship between the input and output data is the main goal of machine learning. According to section 2.2, however, adding those extracted features is reasonable at first glance, provoking the Curse of Dimensionality, and thereby the relevant features are obscured by irrelevant and (or) redundant features. The two following questions may then arise: "*what are irrelevant and redundant features?*" and "*how can relevant features be selected among other features?*". In this section a brief introduction of three main categories of feature selection methods (filters, wrappers, and embedded) is provided. In addition, the theoretical background and algorithms of the 7 feature selection methods are described.

Relevant features are those features which are informative and descriptive. Redundant features are defined as features that carry the same information as one or some other features. Both feature types can be determined by feature selection. Also, data reduction, feature set reduction, performance and speed improvements, and visualization can be counted as other motivations for using feature selection. In fact, feature selection means selecting the features that lead to "*largest possible generalization*" or "*equivalently to minimal risk*" [48]. There are different feature selection approaches that can be classified as follows:

3.1.1. Filter Methods

These methods rank all the feature set using methods, including correlation coefficients, which measure the dependence of individual features with the predefined label (target). These individual ranking (univariate) methods have limitations since a feature that is individually irrelevant may become relevant in the context of other features, and also some individually relevant features may not be useful because of possible redundancies. Therefore "*multivariate*" methods are proposed because they consider not only individual dependencies. The main problem is how to calculate the relevancy between features and targets. Different approaches are explained briefly as follows:

Relevance indices based on correlation:

Correlation based filters can be counted as the simplest approach among the others. The Pearson correlation coefficient [49] is common in calculating the dependency between two features and is defined as:

$$\rho(X,Y) = \frac{E(XY)-E(X)E(Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}} = \frac{\sum_i (x_i-\bar{x_i})(y_i-\bar{y_i})}{\sqrt{\sum_i (x_i-\bar{x_i})^2 \sum_j (y_j-\bar{y_j})^2}}, \tag{3.1}$$

where $E(XY)$ is the cross-correlation between X and Y. $\sigma^2(X)$ $(= E(X^2))$ and $\sigma^2(Y)$ $(= E(Y^2))$ are the variances of signals X and Y. X is the feature vector with values x, and Y is the class vector (target) with values y. If X and Y are linearly dependent then $\rho(X,Y)$ is ±1 and if they are completely uncorrelated then $\rho(X,Y)$ is equal to zero. This criterion works well until the relation between values of the feature and class vectors remain monotonic.

Another simple criterion based on the mean of class distributions and called signal-to-noise ratio is defined as:

$$\mu(X,Y) = \frac{\mu(y_+)-\mu(y_-)}{(\sigma(y_+)+\sigma(y_-))}, \tag{3.2}$$

where $\mu(y_+)$ is the mean value of class y_+ and $\sigma(y_+)$ is the variance of this class. Equation (3.2) is similar to the Fisher linear discriminant [50] criterion, which is defined as:

$$J(X,Y) = \frac{|\mu(y_+)-\mu(y_-)|^2}{(\sigma(y_+)^2+\sigma(y_-)^2)}, \tag{3.3}$$

where the nominator and denominator are named between-class and within-class variances, respectively. Therefore in Equation (3.3), the projection is desired where observations from the same class are close together and at the same time, the projected means of different classes are far as possible. Also, a two-sample T-test uses a slightly different equation and is defined as:

$$T(X,Y) = \frac{\mu(y_+)-\mu(y_-)}{\sqrt{\frac{\sigma(y_+)^2}{m_+}+\frac{\sigma(y_-)^2}{m_-}}}, \tag{3.4}$$

where $m_\pm$ is the number of samples in class $y_\pm$.

It is also necessary to show how significant differences are in $\rho(X,Y)$ and other index values. One simple test is defined as:

$$\mathrm{P}(X\sim Y) = \mathrm{erf}(|\rho(X,Y)|\sqrt{\frac{m}{2}}), \tag{3.5}$$

where erf is the error function. In other words, $\mathrm{P}(X\sim Y)$ estimates the probability that variables X and Y are correlated and can be used to rank the feature vectors in descending order. Usually a threshold is set for $\mathrm{P}(X\sim Y)$ since the probability of $\mathrm{P}(X\sim Y)$ is so close to 1. Also, other methods such as cross-validation or wrapper may be used for ranking.

As mentioned in the first part of this section, in addition to measuring the relevancy between feature vectors and classes, it is important to measure the redundancy between redundant (selected) feature vectors. For this purpose Equation (3.6) is proposed [51].

$$J(X_k, Y) = \frac{kr_{ky}}{\sqrt{K+(K-1)r_{kk}}}, \tag{3.6}$$

where $r_{ky} = \bar{\rho}(X_k, Y)$ is the average of correlation coefficients between feature vectors and classes, and $r_{kk} = \bar{\rho}(X_k, X_k)$ is the average correlation coefficients between two feature vectors.

Relevance indices based on distances:

In this part, three main methods are introduced, which measure the dependence among features and classes based on evaluating their probability distributions. Kolmogorov proposed a difference between the joint and the product distributions as follows:

$$D_K(Y, X) = \sum_i \sum_{j=1}^{K} |\mathcal{P}(y_j, x_i) - \mathcal{P}(x_i)\mathcal{P}(y_j)|, \tag{3.7}$$

where $\mathcal{P}$ shows the probability. If $D_K(Y, X)$ is equal to zero, then the feature vector is completely irrelevant. It should be noted that $D_K(Y, X)$ does not depend on the number of samples. Also it is bounded as shown in Equation (3.8).

$$0 \leq D_K(Y, X) \leq 1 - \sum_i \mathcal{P}(x_i)^2. \tag{3.8}$$

If two classes have the same priori probabilities, Equation (3.7) reduces to:

$$D_K(Y, X) = \frac{1}{2}\sum_i |\mathcal{P}(x_i|y = 0) - \mathcal{P}(x_i|y = 1)|. \tag{3.9}$$

The second criterion is the Bayesian measure or the average Euclidean norm of conditional distribution [52] and is defined as:

$$J_{BM}(Y, X) = \sum_i \mathcal{P}(x_i) \sum_{j=1}^{K} \mathcal{P}(y_i|x_i)^2. \tag{3.10}$$

In fact, $J_{BM}(Y, X)$ measures the concentration of the conditional probability distribution for different values of x_i.

The third criterion, called Jeffreys-Matusita distance (JM-distance) [48], is defined as:

$$D_{JM}(Y, X) = \sum_i \sum_{j=1}^{K} [\sqrt{\mathcal{P}(y_j, x_i)} - \sqrt{\mathcal{P}(x_i)\mathcal{P}(y_j)}]^2, \tag{3.11}$$

if $D_{JM}(Y, X)$ is equal to zero, then feature X is irrelevant and if it is growing to one, it is a highly relevant feature vector. D. R. Wilson et. al [53] designed the Value Difference

Metric (VDM) to evaluate the redundancy between features using conditional probabilities:

$$VDM(X, X'; Y)^2 = \sum_i \sum_{j=1}^{K} |\mathcal{P}(y_i, |x_i) - \mathcal{P}(y_j|x_i')|. \tag{3.12}$$

Relevance indices based on Information Theory:
The most common criteria are based on information theory. Information is equal to negative of entropy and is defined as:

$$H(X) = -\sum_{i=1}^{K} \mathcal{P}(x_i) \log_2 \mathcal{P}(x_i), \tag{3.13}$$

where $\mathcal{P}(x_i)$ is the prior probability for all X values and in the discrete contribution it is equal to the fraction of samples X in the class to all samples. Also information on the joint distribution of classes and features is defined as:

$$H(Y, X) = -\sum_i \sum_{j=1}^{K} \mathcal{P}(y_i, x_i) \log_2 \mathcal{P}(y_i, x_i), \tag{3.14}$$

where $\mathcal{P}(y_i, x_i)$ is the joint probabilities of feature vector X after the Y values (target) are given. The lower value of $H(Y, X)$, makes the feature vector X more valuable. Mutual Information is defined based on Equation (3.13) and (3.14):

$$MI(Y, X) = H(Y) + H(X) - H(Y, X) = -\sum_{i,j} \mathcal{P}(y_j, x_i) \log_2 \frac{\mathcal{P}(y_j, x_i)}{\mathcal{P}(y_j)\mathcal{P}(x_i)}. \tag{3.15}$$

According to the definition of Mutual Information, a feature vector is more significant if the $MI(Y, X)$ is larger. Kullback-Leibler divergence is similar to Mutual Information and its original form is:

$$D_{KL}((\mathcal{P}(X))||(\mathcal{P}(Y)) = \sum_i \mathcal{P}_Y(y_i) \log \frac{\mathcal{P}_Y(y_i)}{\mathcal{P}_X(x_i)} \geq 0. \tag{3.16}$$

Equation (3.16) can be extended into the following form, which, as mentioned, is similar to Mutual Information,

$$D_{KL}((\mathcal{P}(X, Y))||(\mathcal{P}(X)\mathcal{P}(Y)) = \sum_i \sum_{j=1}^{K} \mathcal{P}(y_j, x_i) \log \frac{\mathcal{P}(y_j, x_i)}{\mathcal{P}(y_j)\mathcal{P}(x_i)}. \tag{3.17}$$

P. Smyth et. al [54] proposed J-measure:

$$J_J(X) = \sum_i \mathcal{P}(x_i) \sum_j \mathcal{P}(y_j|x_i) \log \frac{\mathcal{P}(y_j|x_i)}{\mathcal{P}(y_j)}. \tag{3.18}$$

There are many other criteria that have also been proposed, such as average weight of evidence [55] and Minimum Description Length (MDL) [56] based on information theory.

Relevance indices based on Decision Trees:

Decision trees select the features in a top-down portioning procedure. There are various algorithms using the decision tree scheme for feature selection, but only a few of them are introduced in this section.

1R algorithm ranks feature vectors according to the error rate, i.e., when the predefined impurity values for a feature vector is achieved, then that feature vector is removed. 1R consists of single-level trees and feature vectors are analyzed to see which vectors are dominating in a single class. Different criteria can be used for its performance [57]. The C4.5 tree algorithm [58], opposite to 1R algorithms, ranks the most important features which are close to the root node.

Classification And Regression Trees (CART), as a non-parametric learning technique [59], uses Gini index to evaluate the impurity of feature vectors, which is defined as:

$$J_{Gini}(Y) = 1 - \sum_i \mathcal{P}(y_j)^2, \tag{3.19}$$

where $\mathcal{P}(y_j)$ is the class probability distribution for a node. Now if a feature vector is split into several subsets with values of x_j, then the gain is proportional to:

$$J_{Gini}(Y, X) = \sum_j \mathcal{P}(x_j) \sum_i \mathcal{P}(y_i|x_j)^2 \in [0,1]. \tag{3.20}$$

Generally in CART, the first rules for the best split (splitting the input data) are determined. Then each node splits into 2 and it continues as a recursive procedure until no gain can be made or some predefined stopping rules are met. The best split value should separate the maximum number of feature vectors, which are from different classes and separate the minimum number of classes, which belong to the same class. One criterion for determining splits in decision trees is Separability Split Value (SSV):

$$\begin{aligned} SSV(s, f) = 2 \sum_{i=1}^{K} |LS(s, f, D_i)| \cdot |RS(s, f, D - D_i)| \\ - \sum_i \min(|LS(s, f, D_i)|, |RS(s, f, D_i)|) \end{aligned} \tag{3.21}$$

where s is a subset for all possible values of the feature; D is the given dataset (feature vectors), and the left side (LS) and the right side (RS) for s is determined by a test $f(X, s)$ as [48]:

$$LS(s, f, D) = \{x \in D : f(x, s) = T\} \tag{3.22}$$

and

$$RS(s, f, D) = D - LS(s, f, D). \tag{3.23}$$

3.1.2. Wrapper Methods

The primary definition that should be explained to understand wrapper method is induction. Induction is presented as a set of feature vectors, which are describing the main training dataset, and with their corresponding labels (targets). Wrapper methods use the induction algorithm itself for feature selection. The idea behind the wrapper feature selection is shown in Figure 3.1. In induction algorithms, the data is broken into training and test sets and the feature subsets with the highest estimated values are chosen as the selected features. Then the classifier is evaluated on the unseen test set. In other words, wrappers use a learning machine as a "*black box*" to score the feature vectors according to their predictive power. A common way of performance evaluating in the wrapper method is cross validation, which is shown in Figure 3.2 [60].

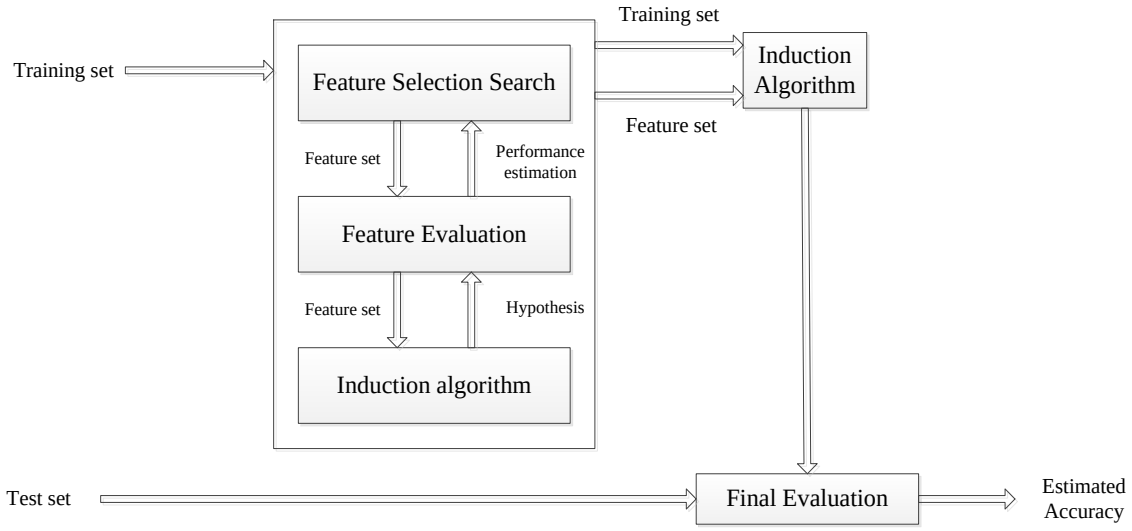

Figure 3.1. *The wrapper method for feature selection scheme* [60].

In the wrapper approach, the performance of a trained learning machine using the given feature vectors (training set) decides which feature sets should be kept, whereas in the filter method, the criteria which are not involved in any machine learning procedure rank the feature vectors. In essence, wrapper methods evaluate the quality of feature vectors using a learning machine. This is why this method can be combined by any machine learning method. Among many wrapper methods, Genetic algorithm, Sequential Forward Selection (SFS), and Oscillating Selection can be mentioned.

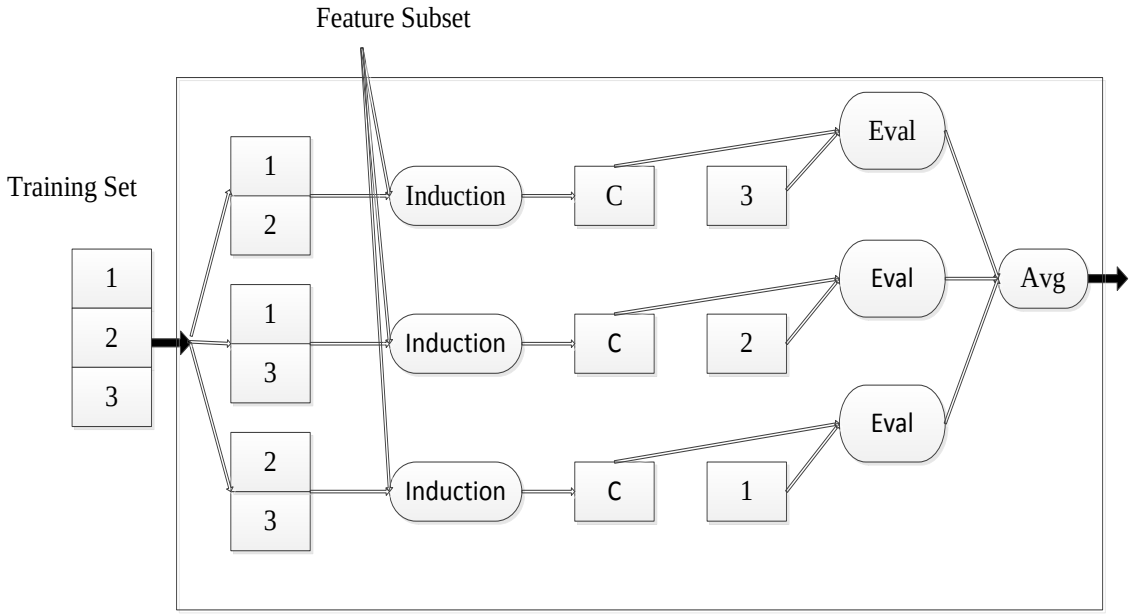

Figure 3.2. *The cross-validation method for accuracy estimation. 3-fold cross-validation is shown here* [60].

3.1.3. Embedded Methods

Embedded methods select the features in a classifier as part of a learning task. This is similar to the wrapper method, with the difference that in the embedded method the hypothesis and search space are combined.

In feature selection, it is desired to minimize the expected risk [48], which is defined as:

$$R(\alpha,\sigma) = \int L[f(\alpha,\sigma \odot x\,),y]\;\; dP(x,y), \tag{3.24}$$

where α is the set of parameters of a classification or regression function, $\sigma \in \{0,1\}^n$ is a vector that model the feature subsets. $\sigma := 0$ shows that the feature is absent in a subset and vice versa. $\odot$ denotes the Hadamard product. L is a loss function and P is a measure in the domain of the training data. Embedded methods use Equation (3.25) to minimize the $R(\alpha,\sigma)$:

$$\min_{\sigma\in\{0,1\}^n} G(\alpha^*,\tilde{T},\sigma,X,Y)\;\; s.t.\;\; \begin{cases} s(\sigma) \le \sigma_0 \\ \alpha^* = \tilde{T}(\sigma,X,Y)' \end{cases} \tag{3.25}$$

where $s : [0,1]^n \to \mathbb{R}^+$ measures the sparsity of the indicator σ, $\tilde{T}$ is a learner, and it could be any common classification algorithm. The function G measures the performance of trained classifier $f^*(\sigma)$ on the training data (X,Y) for a given σ. There are different methods, such as forward-backward methods and optimization of scaling factors to find a solution for Equation (3.25). Two examples of embedded methods are recursive portioning methods or Recursive Feature Elimination (RFE).

The filter, wrapper and embedded methods are schematically summarized in Figure 3.3. Filter and wrapper methods are different in their evaluation criterion. Filters mostly use criteria which are not involved in any learning machine, while wrappers use the performance of a classifier for feature selection. Also wrapper methods do not use information about any specific structure of the classification or regression function for feature selection purpose, whereas embedded methods do not separate the learning from the feature selection part. The wrapper method, in contrast to the other 2 methods, has more computational burden. The summary of these methods is shown in Table 3.1.

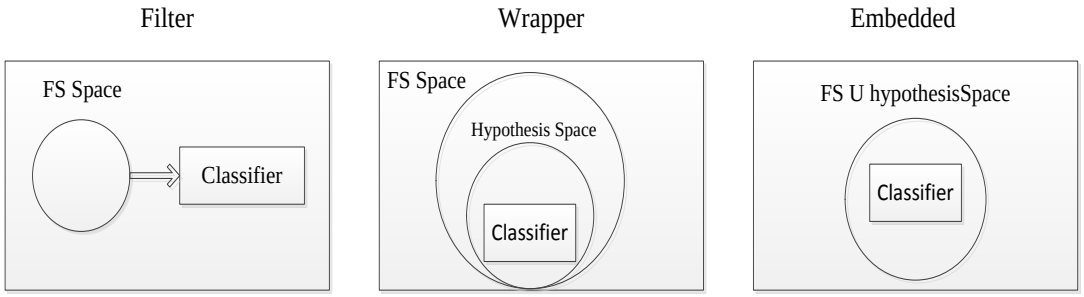

Figure 3.3. *Architecture of three different types of feature selection* [61].

Table 3.1. *Filter, wrapper and embedded methods comparison* [62].

	Filter	**Wrapper**	**Embedded**
Criterion:	Measure feature/feature subset "relevance"	Measure feature subset "usefulness"	Measure feature subset "usefulness"
Search:	Usually rank features	Search the space of all feature subsets	Search guided by learning process
Assessment:	Use statistical tests	Use cross-validation	Use cross-validation
Results:	- (relatively)Robust against over-fitting - May fail to select the most "useful" features	Can in principal find the most "useful" features, but - Are prone to over-fitting	Similar to wrappers, but - Less computationally expensive - Less prone to over-fitting

3.2. Feature selection Methods

In this section, the technical backgrounds of Conditional Mutual Information (CMIM), Fast Correlation Based Filter (FCBF), Mutual Information Feature Selection (MIFS), Mutual Information Maximization (MIM), Max-Relevance Min-Redundancy (MRMR), Joint Mutual information (JMI), and Double Input Symmetrical Relevance (DISR) as 7 different feature selection methods are described. The feature selection is performed using MATLAB Version 7.13 using the feature selection MATLAB toolbox (FEAST) [63].

3.2.1. Conditional Mutual Information Maximization

Fleuret [64] addressed two problems in feature selection. The first problem is that minimizing $\widehat{H}(Y|X_{V(1)}, \dots, X_{V(K)})$, which means minimizing the H function using K features selected during the feature selection process, by choosing $V(1), \dots, V(K)$ cannot be estimated by a training set with a realistic size since it needs to estimate 2^{K+1} probabilities, which leads to a heavy computational burden. One approach to solve this problem is to randomly sample the features in a way that no dependency can be found between features. Even with this approach, redundant features may exist and not be taken into account. Fleuret proposed Conditional Mutual Information Maximization (CMIM) as an approach to overcome these two problems. It can expressed as the following iterative scheme:

$$\begin{aligned} V(1) &= \underset{n}{arg\max}\ \hat{I}(Y\ ;\ X_n), \\ V(k+1) &= \underset{n}{arg max}\left\{\underset{l\le k}{min}\,\hat{I}\left(Y\ ;\ X_n\middle|X_{V(l)}\right)\right\}, \end{aligned} \tag{3.26}$$

where X is the feature vector and Y is the class. In essence, this means that feature X' is selected only if $\hat{I}(Y\ ; X'|X)$ is large for every already selected X. In other words, X' is selected if it carries information about Y (relevancy), and if this information has not been the same in any of the chosen X (redundancy). In Equation (3.26), the term $min_{l\le k}\,\hat{I}\left(Y\ ;\ X_n\middle|X_{V(l)}\right)$ is called score $s(n,k)$, where $I\left(Y\ ;\ X_n\middle|X_{V(l)}\right)$ is defined as:

$$I\left(Y\ ;\ X_n\middle|X_{V(l)}\right) = H\left(Y\middle|X_{V(l)}\right) - H\ (Y|X_{V(l)}, X_n), \tag{3.27}$$

where $H\left(Y\middle|X_{V(l)}\right)$ is the conditional entropy, which is the entropy of Y after observing the values of variable X (see Equations (3.13) and (3.14)). If score $s(n,k)$ is low either the feature is not relevant (i.e., X_n does not bring information about Y) or the feature is redundant (i.e., the information is already exist in $X_{V(l)}$).

The standard algorithm of CMIM creates a score vector s, which contains the score of every feature X_n, and is defined as:

$$\mathrm{s(n)} = \underset{\mathrm{l\le k}}{\min}\,\hat{\mathrm{I}}\left(\mathrm{Y}\ ;\ \mathrm{X_n}\middle|\mathrm{X_{V(l)}}\right) \tag{3.28}$$

where this vector is initialized with the values in Equation (3.26). At each iteration CMIM chooses the $V(k)$ with the highest score and then refreshes $s(n)$ by the minimum value of $s(n)$.

From an implementation point of view, it is not cost-effective to store all the features except the selected one during the rest of computation. Thus, the algorithm shown in Figure 3.4 is designed by Fleuret, where $ps[n]$ stores the paritial score X_n and is equal to $ps[n] = min_{l\le m[n]}\,\hat{I}(Y;\ X_n|X_{v(l)})$, and vector $m[n]$ contains the index of the last peaked feature. It should be noted that in this implementation, if the score of a can-

didate is below the best updated score (s^*) in that iteration, then the conditional mutual information between that candidate and the class will not be computed.

Fast Algorithm of CMIM

for $n = 1 \ldots N$ **do**
  $ps[n] \leftarrow mut\_inf\ (n)$
  $m[n] \leftarrow 0$
for $k = 1 \ldots K$ **do**
  $s^* \leftarrow 0$
  for $n = 1 \ldots N$ **do**
    while $ps[n] > s^*$ and $m[n] < k - 1$ **do**
      $m[n] \leftarrow m[n] + 1$
      $ps[n] \leftarrow min(ps[n], cond\_mut\_inf\ (n, nu[m[n]]))$
    if $ps[n] > s^*$ **then**
      $s^* \leftarrow ps[n]$
      $nu[k] \leftarrow n$

Figure 3.4. *The fast algorithm of CMIM*

3.2.2. Fast Correlation Based Filter

In the Fast Correlation Based Filter (FCBF) method [65], the relevant and redundant features are determined according to the correlation between features and the class, and the correlation between the relevant features, respectively. The proposed definition for the correlation in the FCBF method is based on the information-theoretical concept of entropy. Thus, Symmetrical Uncertainty (SU) is proposed as a criterion to measure the dependency shown in Equation (3.29). If $SU(X,Y)$ is 1, then both X and Y variables are dependent (correlated) and if it has the value of 0, then X and Y are independent (uncorrelated).

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right], \tag{3.29}$$

where $IG(X|Y)$ is the information gain and defined as:

$$IG(X\,|\,Y) = H(X) - H(X|Y), \tag{3.30}$$

where $H(X|Y)$ is the entropy of X after observing the values of variable Y. FCBF starts with determining a value of δ as a threshold. If $SU(X,Y) \geq \delta$ then there is a correlation between variables of X and Y and vice versa.

The used algorithm for selecting relevant and removing redundant features using FCBF can be summarized as selecting the relevant features, which have $SU(X,C) \geq \delta$

where X is the feature and C is the corresponding class. After selecting relevant features, among these features $SU(X,Y)$, where X and Y are both relevant features, is calculated. If $SU(X,Y) \geq \delta$, the X and Y are redundant (correlated) and if $SU(X,Y) < \delta$, then X and Y are irredundant (uncorrelated). In our study δ is equal to 0.00005.

3.2.3. Mutual Information Feature Selection

Battiti in [66] proposed the Mutual Information Feature Selection (MIFS) method. In MIFS all the mutual information (Equation (3.17)) values between each feature $f \in F$ and class C are calculated, and the first feature f that maximizes the mutual information ($I(C;f)$) is stored in vector S. In the next step, the mutual information between $f \in F^*$, where F^* is the set of feature vectors, excludes the first selected features. The next feature is then selected and stored in vector S that maximizes $I(C;f) - \beta \sum_{s \in S} I(f;s)$, where the coefficient β sets the relative importance of the mutual information between the next feature and the already selected features with respect to the $I(C;f)$. This continues until vector S has K components, which is the number of interested features. The algorithm is shown in Figure 3.5.

Algorithm of MIFS

1) Initialization:
   a. Set $F \leftarrow$ "initial set of n features;" $S \leftarrow$ "empty set."
2) Computation of the MI with the output class:
   a. **for** each feature $f \in F$
      i. Compute $I(C; f)$.
3) Choice of the first feature:
   a. Find the feature f that maximizes $I(C;f)$;
   b. Set $F \leftarrow F \backslash \{f\}$; set $S \leftarrow \{f\}$
4) Greedy selection: repeat until $|S| = K$:
   a. Computation of the MI between variables:
      i. for all couples of variables (f,s) with $f \in F, s \in S$
         Compute $I(f; s)$, if it is not already available.
   b. Selection of the next feature:,
      i. choose feature f as the one that maximizes
         $$I(C; f) - \beta \sum_{s \in S} f; s)$$
      ii. set $F + F \setminus \{f\}$; set $S \leftarrow S \cup \{f\}$
5) Output the set S containing the selected features

Figure 3.5. *The algorithm of MIFS method*

3.2.4. Mutual Information Maximization

A feature selection method based on mutual information is designed by D. Lewis [67] with the purpose of text categorization. He computed the mutual information between each feature and class individually and scored the features. After ranking the features, first K features were chosen. The value of K is a predefined number of features or stopping criteria. The main drawback of this method is that features are ranked individually and independently from other features, which causes the redundant features to be retained. In the MIFS method $I(C;f) - \beta \sum_{s \in S} I(f;s)$ is used as a criterion in order to overcome this problem.

3.2.5. Max-relevance and Min-redundancy

Peng et al. [68] presented a feature selection method in which both filter and wrapper methods are used. The two-stage feature selection algorithm first finds a candidate feature set and in the next stage selects a compact feature subset.

In the first stage, in order to select the candidate feature set, cross-validation classification error for the whole features is used to find a relatively stable range of error, which is called Ω. This stage is composed of three steps, as follows:

1) Mutual information is used as a criterion to rank the features (similar to MIM).
2) All the ranked features are evaluated using cross-validation classification error e_k in order to find the range of k with small e_k (i.e., small mean and small variance).
3) Finding the smallest e_k in the range of Ω. The smallest number of candidate features is chosen as the smallest k that corresponds to $e^* = \min e_k$.

In the second stage, the wrapper technique is applied on the small set features that are already selected in the first stage. Peng proposed 2 selection schemes of backward and forward selection for wrapper, as follows:

1) Backward: removes the redundant features within the candidate selection. In the backward selection the rule is to remove the feature so that its classification error e_{k-1} is no worse than the error of the other selected feature (e_k). For all possible configurations, all e_{k-1} are calculated, and if for every configuration the corresponding e_{k-1} has a higher value than e_k, then the backward selection stops.
2) Forward: selects the compact feature subset. Up to this step the selected feature set within the candidate features is stored in S_{n^*} vector. The wrapper with a forward approach selects one feature that leads to minimum error and removes that feature from S_{n^*} (i.e., ${S_{n^*}}^* \leftarrow S_{n^*}$). This selection is repeated until the classification error begins to increase ($e_{k-1} > e_k$).

3.2.6. Joint Mutual Information

H. Yang et al. [69] discuss joint mutual information as a criterion for feature selection. If the Kullback-Leibler formula (Equation (3.16)) is rewritten as follows:

$$I(X_i, y) = K[\mathcal{P}(x_i, y)||\mathcal{P}(x_i)\mathcal{P}(y)], \tag{3.31}$$

then the joint mutual information is defined as:

$$I(X_i, \dots, X_k; Y) = K[\mathcal{P}(i, \dots, k, y)||\mathcal{P}(i, \dots, k)\mathcal{P}(y)], \tag{3.32}$$

where $\mathcal{P}(i,k) = \mathcal{P}(i, \dots, k)$ and $\mathcal{P}(i,k,y) = \mathcal{P}(x_i, \dots, x_k, y)$.
Yang proposed selecting the first 2 or 3 most relevant features for visualization using the ones with the maximum joint mutual information or using Principal Component Analysis (PCA) to find new coordinates to display the data. However, these approaches can be used only for a small amount of data.

3.2.7. Double Input Symmetrical Relevance

Meyer et al [70] based their method on the theorem which expresses that "the mutual information of a subset S and a target variable Y is lower bounded by the average of the same quantity for all the sub-subsets X_{S-i} of X_S":

$$I(X_S; Y) \geq \frac{1}{d}\sum_{i \in S} I(X_{S-i};\ Y), \tag{3.33}$$

where X_S is a subset of X and X_{S-i} is the subset X_S that does not contain the variable X_i. In their method, instead of maximization of the mutual information ($I(X_S; y)$), they maximize its lower bound and replace the right-hand term again with its lower band until the subsets only have two variables:

$$\max_{S:\,|S|=d} I(X_S; Y) \geq \max_{S:|S|=d} \sum_{i \in S} I(X_{S-i}, Y)\ , \tag{3.34}$$

$$\geq \max_S \sum_{i \in S} \sum_{j \in S} I\left(X_{S-(i,j)};\ Y\right) \geq \sum_{i \in S} \sum_{j \in S} I(X_{i,j}, Y)\ , \tag{3.35}$$

where d is equal to the number of variables in X set. In other words, the subset which has the highest sum of mutual information in all possible communications of two variables is selected. For further improvement in this method Meyer used a normalized measure of mutual information called symmetrical relevance $SR(X;Y)$:

$$SR(X;Y) = \frac{I(X,Y)}{H(X,Y)}, \tag{3.36}$$

where $H(X,Y)$ is equal to Equation (3.14) and $I(X,Y)$ is equal to Equation (3.15). Thus, the resulting criterion can be written as follows:

$$X_{DISR} = \arg\max_{X_i \in X_{-S}} \{\sum_{X_j \in X_s} SR(X_{i,j}; Y)\}, \tag{3.37}$$

4. PATIENT-SPECIFIC EEG CLASSIFICATION SYSTEM

4.1. Overview of the Proposed System

The goal is to design a seizure detection system which is not only able to detect the onset of seizure occurrences but also able to extract the entire seizure section. In seizure detection using EEG signal, the main difficulty is that there are various seizure patterns through different subjects. This problem can be more serious if the subject suffers from different seizure types, especially in case of intractable seizures. In consequence, a feature's discrimination efficiency between seizure and non-seizure patterns can be different through different subjects and channels. For instance in Figure 4.1, the feature of coefficient of variation of approximation coefficients in fourth level decomposition using db4 is able to detect the seizure in subject 17 (seizure starts at 3025 seconds and ends at 3140 seconds), while the same feature on the same channel in subject 13 (where seizure starts at 934 seconds and ends at 1004 seconds) is not able to indicate the seizure event. Also in Figure 4.2 and Figure 4.3 different discrimination efficiency of two other features in subjects 17 and 13 can be seen.

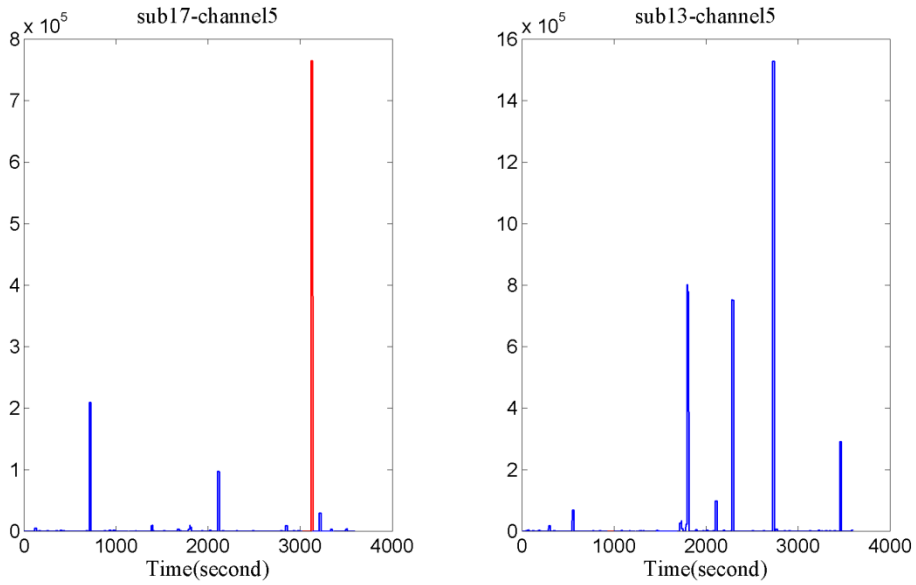

Figure 4.1. *The coefficient of variation of approximation coefficients in fourth level decomposition using db4.This feature is shown in two subjects, 17 and 13, and is different in showing seizure (the red colored parts show the seizure events).*

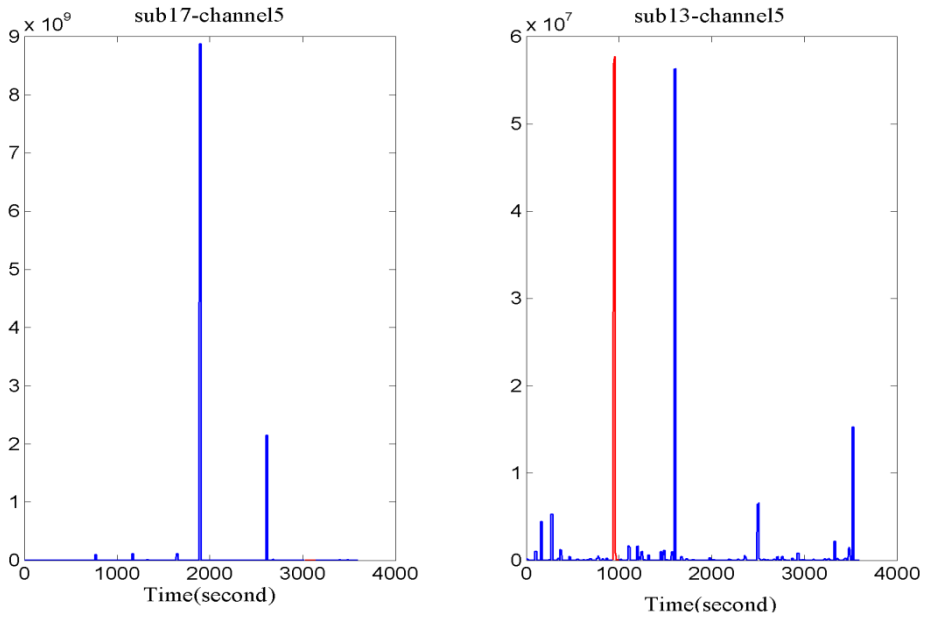

Figure 4.2. *The coefficient of variation of approximation coefficients in fourth level decomposition using db2.This feature is shown in two subjects, 17 and 13, and is different in showing seizure (the red colored parts show the seizure events).*

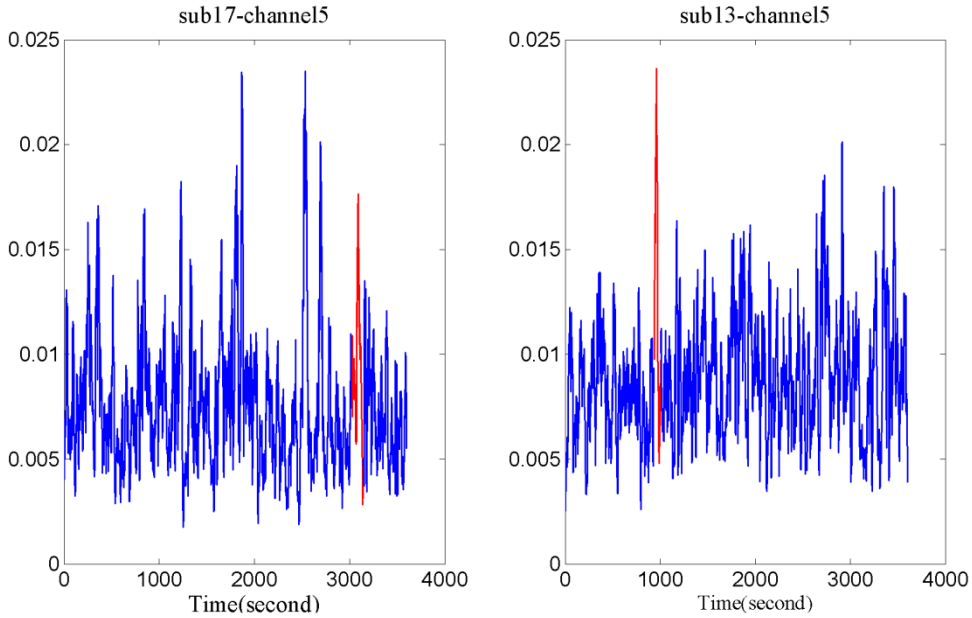

Figure 4.3. *The relative scale energy of approximation coefficients in fourth level decomposition using db4.This feature is shown in two subjects, 17 and 13, and is different in showing seizure (the red colored parts show the seizure events).*

All the features presented in Section 2.2 are extracted for each patient in order to exploit the characteristics of a patient's EEG pattern as much as possible. Then all extracted features are normalized and smoothed using moving average filter
tion 2.2.7). When all the features are extracted, the relevant and irredundant features are selected for each subject individually using the feature selection method (Section 3.2.1). This operation is done for each patient because different patients may suffer from dif-

ferent seizure types. In the last step a Support Vector Machine (SVM) with a linear kernel is used as the classifier. After all variables such as features and SVM parameters are set, the test set of same subject is applied to the system for seizure detection. In essence, the main idea is to train the framework with an earlier EEG signal and then use the system for detecting seizures for the future EEG recordings of the same patient. The block-diagram of the proposed framework is shown in Figure 4.4.

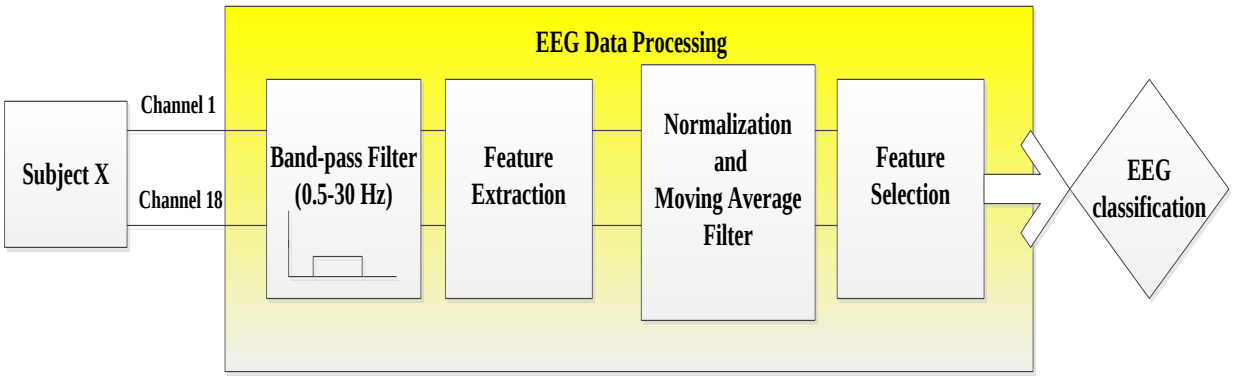

Figure 4.4. *Architecture of patient-specific EEG classification framework.*

4.2. Support Vector Machines

In the two-class case where the classes of ω_1 and ω_2 are linearly separable, the goal is to find a decision hyperplane, as follows:

$$g(x) = \omega^T x + \omega_0 \ , \tag{4.1}$$

where $\omega = [\omega_1, \omega_2, \dots, \omega_N]^T$ is called "*weight vector*" and ω_0 is known as the "threshold". x_i, $i = 1, 2, \dots, N$, is the feature vectors in the training set, X.

If it is assumed that x_1 and x_2 are two points on the decision hyperplane, then:

$$0 = \omega^T x_1 + \omega_0 = \omega^T x_2 + \omega_0 = 0 \Rightarrow \omega^T (x_1 - x_2) = 0, \tag{4.2}$$

from Equation (4.2) it is obvious that the vector ω is orthogonal to the decision hyperplane. In essence, $|g(x)|$ is a measure of the Euclidean distance of the point x from the decision hyperplane. The main task is to compute the unknown ω_j, $j = 0, \dots, l$. However, there is more than one hyperplane that can separate the two linearly separable classes. The question may arise as to which hyperplane is optimized. The answer to this question is related to the "generalization performance of the classifier", which refers to the capability of the classifier, which is designed by the training set, to operate satisfactorily with test data. The sensible choice for the hyperplane would be one that leaves the maximum margin from both classes.

Every hyperplane is quantified by its direction, ω, and its exact position in space, ω_0. Referring to the choice of hyperplane, it is reasonable for each direction to select the hyperplane which has the same distance from the nearest points in classes (ω_1, ω_2). The shortest distance, z, between point (x_0, y_0) and plane $ax + by + c = 0$ in Euclidean geometry can be calculated as follows:

$$z = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}. \tag{4.3}$$

Thus, the distance between a point and the hyperplane in the classification problem can be rewritten as follows:

$$z = \frac{|g(x)|}{||\omega||}. \tag{4.4}$$

The ω and ω_0 can be scaled in the way that the value of $g(x)$, at the nearest points in ω_1 is equal to 1 and for ω_2 is equal to -1, which is equivalent to

$$\begin{cases} \omega^T x + \omega_0 \geq 1, & \forall x \in \omega_1 \\ \omega^T x + \omega_0 \leq -1, & \forall x \in \omega_2 \end{cases} \tag{4.5}$$

Finding the hyperplane can be summarized as:

$$minimize\ J(\omega, \omega_0) = \frac{1}{2}||\omega||^2 \tag{4.6}$$

$$subject\ to\ \ y_i(\omega^T x_i + \omega_0) \geq 1, \qquad i = 1, 2, \dots, N \tag{4.7}$$

Minimizing Equation (4.6) is a nonlinear (quadratic) optimization task. The Karush-Kuhn-Tucker (KKT) conditions that have to be satisfied by the minimizer of Equation (4.6) are:

$$\frac{\partial}{\partial \omega}\mathcal{L}(\omega, \omega_0, \lambda) = 0 \tag{4.8}$$

$$\frac{\partial}{\partial \omega_0}\mathcal{L}(\omega, \omega_0, \lambda) = 0 \tag{4.9}$$

$$\lambda_i \geq 0, i = 1, 2, \dots, N \tag{4.10}$$

$$\lambda_i[y_i(\omega^T x_i + \omega_0) - 1] = 0, \qquad i = 1, 2, \dots, N \tag{4.11}$$

where λ is the vector of the Lagrange multipliers, λ_i, and Lagrangian function $\mathcal{L}(\omega, \omega_0, \lambda)$ is defined as:

$$\mathcal{L}(\omega, \omega_0, \lambda) = \frac{1}{2}\omega^T\omega - \sum_{i=1}^{N}\lambda_i[y_i(\omega^T x_i + \omega_0) - 1]. \tag{4.12}$$

Equation (4.8) can be extended by using Equation (4.12) and thus:

$$\omega = \sum_{i=1}^{N}\lambda_i\, y_i x_i. \tag{4.13}$$

The vector parameter ω of the optimal solution is a linear combination of $N_s \leq N$ feature vectors. These are called support vectors and the optimum hyperplane classifier is called a Support Vector Machine (SVM) [71].

In case that the problem is not linearly separable, *kernel trick* extends the application of SVM to nonlinear class borders. The main idea of kernel trick is to map the non separable data set $X = \{x_1, \dots, x_n\} \in \mathbb{R}^p$ to a higher dimensional data set $X' =$

$\{x'_1, \dots, x'_n\} \in \mathbb{R}^q$, where $q > p$, so that the structure of data in X' is more suitable than X. In essence, kernel trick is used to map the class borders, which are nonlinear, to class with linear class borders so that the linear classification approach presented above can be used. According to Mercer's theorem that states that for any data set X and any kernel function $k: \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ there is amapping $\varphi: \mathbb{R}^p \rightarrow \mathbb{R}^q$ so that $k(x_j, x_k) = \varphi(x'_j).\varphi(x'_k)^T$ [72].

5. EXPERIMENTAL RESULTS

5.1. Performance Evaluation Metrics

In this study, in order to evaluate the classification task, standard performance measures of *sensitivity*, *specificity*, and *accuracy* are used and are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.3}$$

where TP (True Positive) is the number of correctly segments detected as seizure, FN (False Negative) is the number of incorrectly detected segments as non-seizure, TN (True Negative) is the number of correctly detected segments as non-seizure, and FP (False Positive) is the number of incorrectly detected segments as seizures. In other words, sensitivity shows the ability of the classifier to identify positive results (i.e., seizures) and specificity shows the ability of a test to identify negative results (i.e., non-seizures). Sensitivity and specificity are mostly used in clinical tests and both are independent of the population being tested.

It should be noted that sensitivity, specificity and accuracy must be used at the same time, especially in the case of highly unbalanced data (e.g., detecting seizure segments in long-term EEG recording). For instance, in a one hour (3600 seconds) EEG signal the duration of seizure and non-seizure segments are 40 and 3560 seconds, respectively. If the classifier achieves 0% sensitivity and 100% specificity, then the accuracy will be 98.88%. Therefore using only accuracy rate is not a proper evaluation of classification performance and all the three sensitivity, specificity and accuracy are needed to be used at the same time.

5.2. Results

In this study, the EEG signal of each subject is divided into train and test sets. Train and test sets contain 50% of the seizure and 50% non-seizure segments contiguously while the former precedes the latter in time. As for the classifier the *linear* kernel is used for SVM as the classifier since it gives the best results on the test set as compared to other kernel alternatives such as RBF, polynomial or Gaussian. All the processing is done by MATLAB software version 7.13.

The classification results over the EEG recordings of 4 subjects with a duration of 21 hours, in terms of sensitivity, specificity and accuracy are obtained with and without (named original in Table 5.1) using the proposed feature selection method (CMIM). In addition, the classification results using the 6 other feature selection methods are compared against the performance obtained by CMIM. All classification results are shown in Table 5.1. In each column, the highest scores are highlighted in bold.

As can be seen, the highest average sensitivity is obtained by the CMIM method over the test set. Almost all the feature selection methods (except FCBF and MIFS) obtained acceptable results. However, it should be noted that these methods should be applied on more subjects in order to ensure the stability of the proposed framework.

Also as it was expected the specificity of all methods are almost 100%. Seizure detection problem is an imbalanced classification problem since the duration of seizure events in contrast to the nonseizure segment duration is short. Thus, the value of specificity rate, which is the ability of system in detecting nonseizure segments, is always high. This is also the reason why in many classification problems in clinical area the sensitivity rate is the main performance evaluation metric.

According to the sensitivity rates shown in Table 5.1, the sensitivity rate obtained using original features is not reliable in contrast with other feature selection methods (e.g., the standard variation values of original and CMIM are 0.0554 and 0.0366, respectively), however the average sensitivity rate of original features is 90.62%. In order to justify this the standard variation of the sensitivity rates of each method is calculated and shown in *Figure 5.1*. As can be seen in *Figure 5.1* using FCBF and MIFS are not reliable since they have relatively low average sensitivity (66.93% and 64.64%, respectively) and the standard variation of the sensitivity rates for these two methods are 0.2258 and 0.1654.

Table 5.1. *Classification results from using 7 different feature selection methods*

Subject	1			2			3			4			**Average**		
	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.
Original	87.35	9.89	99.69	85.39	99.61	**99.37**	**97.87**	99.96	**99.89**	91.88	97.82	97.51	90.62	99.32	**99.11**
CMIM	88.50	99.89	99.71	**94.38**	98.70	98.62	95.39	**99.97**	99.83	**96.85**	97.65	97.61	**93.78**	99.05	98.94
MRMR	86.20	99.86	99.64	92.13	99.30	99.18	89.36	99.85	99.52	87.43	98.26	97.69	88.78	93.31	99.00
MIM	89.65	99.92	99.75	92.13	99.30	99.18	93.61	99.89	99.69	92.93	97.68	97.43	92.08	99.19	99.01
DISR	**91.37**	**99.94**	99.80	91.01	98.79	98.66	87.94	99.90	99.53	95.02	97.40	97.27	91.33	99.00	98.81
JMI	90.22	99.90	99.74	93.25	98.47	98.38	87.94	99.95	98.53	90.31	98.53	**98.09**	90.43	99.21	98.68
FCBF	62.64	99.84	99.24	38.76	**99.76**	98.75	73.40	99.85	99.02	92.93	97.19	96.97	66.93	99.16	98.49
MIFS	63.79	99.78	99.20	41.57	99.67	98.70	77.30	99.94	99.23	75.91	**99.19**	97.95	64.64	**99.64**	98.77

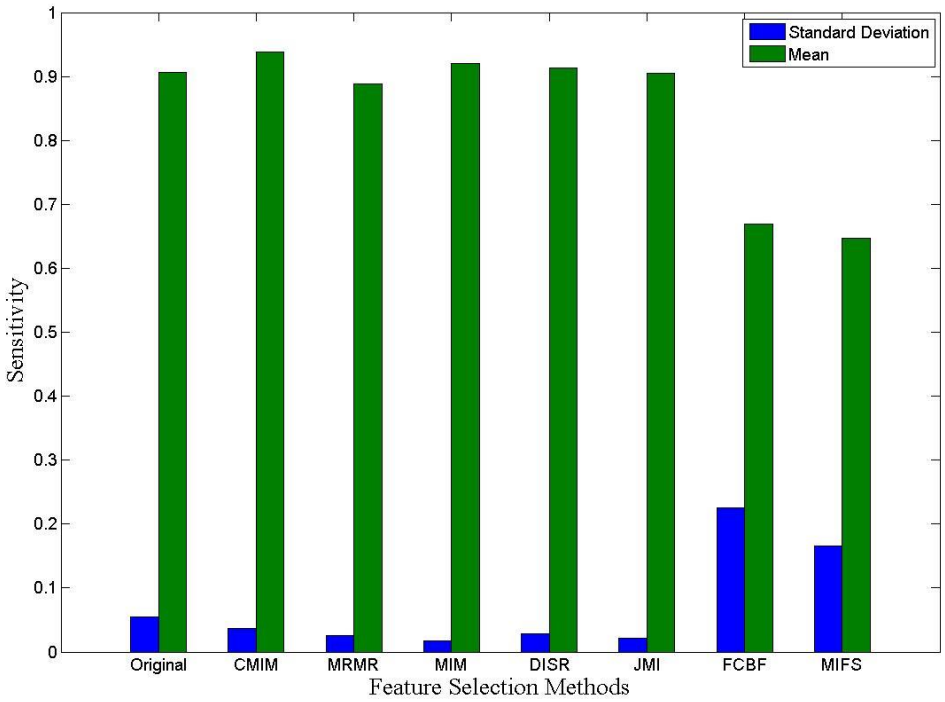

Figure 5.1. *The standard deviation (blue) and mean (green) of sensitivity rates for each feature selection methods*

In our MATLAB implementation, the average time spent over a one-second frame for each operation is listed in Table 5.2, which is also insignificant and can be significantly reduced with a dedicated and optimized implementation. As can be seen in Table 5.2, the spent time for classification using original features is approximately 1.5 times more than classification using only selected features. It should be noted that feature selection method is only applied on train set and once for each patient. In contrary, using original features, all the features are extracted from test set as well as train set. In addition, according to Table 5.2 MIM and DISR have the lowest and highest computational complexity, respectively.

Table 5.2. *The average time over a one-second frame for different operations*

Operation	**Method**	**Time (Second)**
Feature Extraction	-	2.51
Normalization	-	1.02×10^{-3}
Moving Average	-	4×10^{-5}
Feature Selection	CMIM	4.08×10^{-5}
	MRMR	6.93×10^{-5}
	MIM	4.41×10^{-6}
	DISR	1.88×10^{-4}
	JMI	1.26×10^{-4}
	FCBF	1.31×10^{-5}
	MIFS	5.31×10^{-5}
Classification	SVM (using selected features)	8.5×10^{-6}
	SVM (using original features)	1.3×10^{-5}

6. CONCLUSIONS

Seizure detection using EEG signal in long-term monitoring by neurologists is a time consuming task. In this work, we proposed patient-specific seizure detector and applied the proposed method over the seizure records of the 4 subjects with a total duration of 21 hours.

In this thesis, several feature extraction methods, which are the state-of-the-art features in this domain, are described and used for EEG classification. Also different feature selection techniques with their theoretical background are provided as the main part of the thesis. Each EEG frame is classified as a seizure or nonseizure. The linear kernel is used for SVM since it gives the best results on the test set as compared to other kernel alternatives such as RBF, polynomial or Gaussian.

Two main difficulties in seizure detection problem are addressed in this thesis, namely seizure segment detection and efficiency of features over different individuals. In many of previous works [32] - [33] the onsets of seizure occurances are detected while in this thesis the total duration of each seizure event is diagnosed. To address the second problem, instead of defining fixed number of features, a large set of state-of-the-art features are extracted, which in turn makes the proposed method independent from different seizure types. Then in order to avoid the "*Curse of Dimensionality*" problem for the classifier, several feature selection methods are used to extract the most relevant features specifically for each patient. This also yields a significantly less computational complexity.

The experimental results demonstrated that we achieved a delicate seizure classification accuracy. The proposed feature selection method, CMIM, selects such features that yield the best average sensitivity rate. We aim to test the proposed method over larger EEG datasets with more patients in order to ensure from its feasibility in real clinical settings.

REFERENCES

[1] T. L. Bennett, The Neuropsychology of Epilepsy, New York: Plenum Press, 1992.

[2] E. Niedermeyer and F. L. Da Silva, Electroencephalography. Basic Principles, Clinical Applications, and Related Fields, Philadelphia, PA: LIPPINCOTT WILLIAMS & WILKINS, 2004.

[3] W. J. Friedlander, The History of Modern Epilepsy. The Beginning, 1865-1914, Westport: Greenwood Press, 2001.

[4] R. S. Fisher, W. E. Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, "Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)", *Epilepsia*, vol. 46, no. 4, pp. 470-472, 2005.

[5] A. T. Berg and I. E. Scheffer, "New concepts in classification of the epilepsies: Entering the 21st century", *Epilepsia*, vol. 52, no. 6, pp. 1058-1062, 2011.

[6] National Institute of Neurological Disorders and Stroke, "National Institutes of Health", March 2004. [Online]. Available: http://www.ninds.nih.gov/disorders/epilepsy/detail_epilepsy.htm.mtlab

[7] A. T. Berg, S. F. Berkovic, M. J. Brodie, J. Buchhalter, J. H. Cross, W. Boas, J. Engel, J. French, T. A. Glauser, G. W. Mathern, S. L. Moshe, D. Nordli, P. Plouin, and I. Scheffer, "Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE commission on classification and terminilogy, 2005-2009", *Epilepsia*, vol. 51, no. 4, pp. 676-685, 2010.

[8] W. M. Burnham, P. L. Carlen, and P. A. Hwang, Intractable Seizures: Diagnosis, Treatment, and Prevention, New York, N.Y.: Kluwer Academic/ Plenum, 2001.

[9] World Health Organization, "World Health Organization", [Online]. Available: http://www.who.int/mental_health/neurology/epilepsy/en/.

[10] S. Shorvon, "The Surgical Therapy of Epilepsy", *Handbook of Epilepsy Treatment*, WILEY-BLACKWELL, 2010, p. 314.

[11] R. I. Kuzneicky, "MRI in cerebral developmental malformations and epilepsy", *Magnetic Resonance Imaging*, vol. 13, pp. 1137-1145, 1995.

[12] H. C. C. Juhasz, "Imaging the epileptic bratin with positron emission tomography", *Neuroimaging*, vol. 4, pp. 5-16, 2003.

[13] T. M.E. Nijsen, J. B. Arends, P. Griep, and P. J. Cluitmans, "The potential value of three-dimensional accelerometry for detection of motor seizures in severe epilepsy", *Epilepsy & Behaviour*, vol. 7, pp. 74-84, 2005.

[14] E. N. Marieb and K. Hoehn, Human Anatomy & Physiology, San Francisco: Pearson Education, 2004.

[15] R. P. J. Malmivuo, Bioelectromagnetism: Principals and applications of bioelectric and biomagnetic fields, New York: Oxford University Press, 1995.

[16] P. L. L. Sornmo, Bioelectrical signal processing in cardiac and neurological applications, California: Elsevier Academic Press, 2005.

[17] D. L. Schomer and F. L. Da Silva, Niedermeyer's electroencephalography: Basic principles, clinical applications, and related fields, Philadelphia: Lippincot Williams & Wilkins, 2012.

[18] J. K. U. K. Misra, Clinical Electroencephalography, Uttar Pradesh: Reed Elsevier India, 2009.

[19] T. F. Collura, "History and Evolution of Electroencephalographic Instruments and Techniques", *Journal of Clinical Neurophysiology*, vol. 10, no. 4, pp. 476-504, 1993.

[20] E. M. T. Yamada, Practical guide for clinical neurophysiology testing: EEG, Iowa: Lippincott Williams & Wilkins.

[21] N. Boutros, S. Galderisi, O. Pogarell, and S. Riggio, Standard Electroencephalography in Clinical

Psychiatry, WILEY-BLACKWELL, 2011.

[22] N. Kannathal, M. L. Choo, U. R. Acharya, and P.K. Sadasivan, "Entropies for detection of epilepsy in EEG", *Computer Methods and Programs in Biomedicine,* vol. 80, pp. 187-194, 2005.

[23] H. R. Mohseni, A. Maghsoudi, and M. B. Shamsollahi, "Seizure detection in EEG signals: A comparison of different approaches", *IEEE International Conference on Engineering in Medicine and Biology Society (EMBS),* New York, NY, 2006.

[24] N. F. Guler, E. D. Ubeyli, and I. Guler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification", *Expert Systems with Applications,* vol. 29, pp. 506-514, 2005.

[25] A. Subasi and E. Ercelebi, "Classification of EEG signals using neural network and logistic regression", *Computer Methods and Programs in Biomedicine,* vol. 78, no. 2, pp. 87-99, 2005.

[26] H. R. Mohseni, A. Maghsoudi, M. H. Kadbi, J. Hashemi, and A. Ashourvan, "Automatc Detection of Epileptic Seizure using Time-Frequency Distributions", *IET 3rd International Conference On Advances in Medical, Signal and Information Processing (MEDSIP 2006),* Glasgow, UK, 2006.

[27] B.R. Greene, S. Faul, W.P. Marnane, G. Lightbody, I. Korotchikova, and G.B. Boylan, "A comparison of quantitative EEG features for neonatal seizure detection", *Clinical Neurophysiology,* vol. 119, pp. 1248-1261, 2008.

[28] L. Kuhlmann, A. N. Burkitt, M. J. Cook, K. Fuller, D. B. Grayden, L. Seiderer, and I. M. Y. Mareels, "Seizure detection using seizure probability estimation: Comparison of features used to detect seizures", *Annals of Biomedical Engineering,* vol. 37, no. 10, pp. 2129-2145, 2009.

[29] M.E. Saab and J. Gotman, "A system to detect the onset of epileptic seizures in scalp EEG", *Clinical Neurophysiology,* vol. 116, no. 2, pp. 427-442, 2005.

[30] Y. Pan, C. Guan, K. K. Ang, K. S. Phua, H. Yang, D. Huang, and S. Lim, "Seizure Detection based on Spatiotemporal Correlation and Frequency Regularity of Scalp EEG", *IEEE World Congress on Computational Intelligence (WCCI),* Brisbane, 2012.

[31] "CHB-MIT Scalp EEG Database", 2010. [Online]. Available: http://www.physionet.org/pn6/chbmit/.

[32] A. Shoeb, and J. Guttag, "Application of machine learning to epiliptic seizure onset detection", *27th International Conference on Machine Learning (ICML),* Haifa, 2010.

[33] Y. U. Khan, O. Foroogh, and P. Sharma, "Automatic detection of seizure onset in pediatric EEG", *International Journal of Embedded Systems and Applications (IJESA),* vol. 2, pp. 81-89, 2012.

[34] J. Gotman, "Automatic recognition of epileptic seizures in the EEG", *Electroencephalography and Clinical Neurophysiology,* vol. 54, pp. 530-540, 1982.

[35] I. Kalatzis, N. Piliouras, E. Ventouras, C.C. Papageorgiou, A.D. Rabavilas, D. Cavouras, "Design and Implementation of and SVM-based computer classification system for discriminating depressive patients from healthy controls using P600 component of ERP signals", *Computer Methods and Programs in Biomedicine,* vol. 75, no. 1, pp. 11-22, 2004.

[36] I. Kalatzis, N. Piliouras, E. Ventouras, C.C. Papageorgiou, A.D. Rabavilas, and D. Cavouras, "Design and implementation of an SVM-based computer classification system for discriminating depressive patients from healthy controls using the P600 component of ERP signals", *Computer Methods and Programs in Biomedicine,* vol. 75, no. 1, pp. 11-22, 2004.

[37] M. D. V. M. S. J. A. K. S. W. V. P. a. S. V. H. B. Hunyadi, "Automatic Seizure Detection Incorporating", *Computer Science,* vol. 6791, pp. 233-240, 2011.

[38] L. Logesparan, A. J. Casson, and E. Rodriguez-Villegas, "Optimal features for online seizure detection", *Medical & Biological Engineering & Computing,* vol. 50, no. 7, pp. 659-669, 2012.

[39] B.R. Greene, S. Faul, W.P. Marnane, G. Lightbody, I. Korotchikova, and G.B. Boylan, "A comparison of quantitative EEG features", *Clinical Neurophysiology,* vol. 6, p. 1248–1261, 2008.

[40] C. H. Seng, R. Demirli, L. Khuon, and D. Bolger, "Seizure detection in EEG signals using support

vector machines", *38th Annual NortheastBioengineering Conference (NEBEC)*, Wollongong, NSW, 2012.

[41] R. Kumari and J. P. Jose, "Seizure detection in EEG using time frequency analysis and SVM", *International Conference onEmerging Trends in Electrical and Computer Technology (ICETECT)*, Tamil Nadu, 2011.

[42] R. Ferenets, T. Lipping, A. Anier, V. Jantti, S. Melto, and Hovilehto, "Comparison of entropy and complexity measures for the assessment of depth of sedation", *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 1067-1077, 2006.

[43] S. A. Fulop, Speech Spectrum Analysis, New York: Springer, 2011.

[44] P. Zarjam, M. Mesbah, and B. Boashash, "An optimal feature set for seizure detection systems for newborn EEG signals", *International Symposium on Circuits and Systems (ISCAS)*, 2003.

[45] L. Kuhlmann, M. J. Cook, K. Fulle, D. B. Grayden, A. N. Burkitt, and I.M.Y. Mareels, "Correlation analysis of seizure detection features", *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2008)*, Sydney, NSW, 2008.

[46] J. B. Dingwell, "Lyapunov exponents", *Wiley Encyclopedia of Biomedical Engineering*, John Wiley & Sons, 2006.

[47] M. Othman, A. Wahab, and R. Khosrowabadi, "MFCC for robust emotion detection using EEG", *IEEE 9th Malaysia International Conference on Communications*, Kuala Lumpur, Malaysia, 2009.

[48] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Feature extraction: Foundations and applications, New York: Springer Berlin Heidelberg, 2006.

[49] K. Pearson, "Mathematical contributions to the theory of evolution.—III. Regression, heredity and panmixia," London, Philos. Trans. R. Soc, 1895, pp. 253-318.

[50] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, vol. 7, no. 2, p. 179–188, 1936.

[51] E. E. Ghiselli, Theory of Psychological Measurement, New York: McGrawHill, 1964.

[52] I. Vajda, Theory of statistical inference and information, London: Kluwer Academic Press, 1979.

[53] D. R. Wilson and T. R. Martinez, "Value difference metrics for continuously valued attributes", *International Conference on Artificial Intelligence, Expert Systems and Neural Networks (AIE'69)*, Honolulu, Hawaii, 1996.

[54] P. Smyth and R. M. Goodman, "An information theoretic approach to rule induction from databases", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 4, no. 4, pp. 301-316, 1992.

[55] D. Michie, "Personal models of rationality", *Journal of statistical Planning and Inference, 21*, pp. 381-399, 1989.

[56] J. Rissanen, "Modeling by shortest data description", *Automatica*, vol. 14, pp. 465-471, 1978.

[57] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets", *Machine Learning*, vol. 11, pp. 63-91, 1993.

[58] J. R. Quinlan, C4.5 programs for machine learning, San Mateo: Morgan Kaufmann Publishers, 1993.

[59] L. Breiman, J. H. Friedman, R.A. Olshen, and C. J. Stone, Classification and regression trees, CA: Wadsworth and Brooks, 1984.

[60] R. Kohavi and J. H. George, "Wrappers for feature subset selection", *Artificial Intelligence*, pp. 273-324, 1997.

[61] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[62] I. Guyon and A. Elisseeff, “ClopiNet”, 2006. [Online]. Available: http://clopinet.com/isabelle/Projects/ETH/lecture9.pdf.

[63] G. Brown, A. Poccock, M. J. Zhao, and M. Lujan, “Conditional likelihood maximisation: A unifying framework for information theoretic feature selection”, *Machine Learning Research*, vol. 13, pp. 27-66, 2012.

[64] F. Fleuret, “Fast binary feature selection with conditional mutual information”, *Journal of Machine Learning Research*, pp. 1531-1555, 2004.

[65] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution”, *20th International Conference on Machine Learning*, Washington, D.C., 2003.

[66] R. Battiti, “Using mutual information for selecting features in supervised neural net learning”, *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 5, pp. 537-550, 1994.

[67] D. D. Lewis, “Feature Selection and feature extraction for text categorization”, *Proceeding of the workshop on Speech and Natural Language*, pp. 212-217, 1992.

[68] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-dedundancy”, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 27, pp. 1226-1238, 2005.

[69] H. Yang and J. Moody, “Feature selection based on joint mutual information”, *International Computer Science*, Rochester, 1999.

[70] P. E. Meyer and G. Bontempi, “On the use of variable complementarity for feature selection in cancer classification”, *Applications of Evolutionary Computing*, vol. 3907, pp. 91-102, 2006.

[71] S. Theoderidis and K. Koutroumbas, Pattern Recognition, California: Elsevier, 2009.

[72] T. Runkler, Data Analytics: Models and Algorithms for Intelligent Analysis, Springer, 2012.