Feature selection and artifact removal in sleep stage classification

by

Pasan Hapuarachchi

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2006

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The use of Electroencephalograms (EEG) are essential to the analysis of sleep disorders in patients. With the use of electroencephalograms, electro-oculograms (EOG), and electromyograms (EMG), doctors and EEG technician can make conclusions about the sleep patterns of patients. In particular, the classification of the sleep data into various stages, such as NREM I-IV, REM, Awake, is extremely important.

The EEG signal itself is highly sensitive to physiological and non-physiological artifacts. Trained human experts can accommodate for these artifacts while they are analyzing the EEG signal. However, if some of these artifacts are removed prior to analysis, their job will be become easier. Furthermore, one of the biggest motivations, of our team's research is the construction of a portable device that can analyze the sleep data as they are being collected. For this task, the sleep data must be analyzed completely automatically in order to make the classifications.

The research presented in this thesis concerns itself with the *denoising* and the *feature selection* aspects of the teams' goals. Since humans are able to process artifacts and ignore them prior to classification, an automated system should have the same capabilities or close to them. As such, the denoising step is performed to condition the data prior to any other stages of the sleep stage neoclassicisms. As mentioned before, the denoising step, by itself, is useful to human EEG technicians as well.

The denoising step in this research mainly looks at EOG artifacts and artifacts isolated to a single EEG channel, such as electrode pop artifacts. The first two algorithms uses Wavelets exclusively (BWDA and WDA), while the third algorithm is a mixture of Wavelets and Independent Component Analysis (IDA). With the BWDA algorithm, determining *consistent* thresholds proved to be a difficult task. With the WDA algorithm, the performance was better, since the selection of the thresholds was more straight-forward and since there was more control over defining the duration of the artifacts. The IDA algorithm performed inferior to the WDA algorithm. This could have been due to the small number of measurement channels or the automated sub-classifier used to select the *denoised EEG signal* from the set of ICA *demixed* signals.

The feature selection stage is extremely important as it selects the most pertinent features to make a particular classification. Without such a step, the classifier will have to process useless data, which might result in a poorer classification. Furthermore, unnecessary features will take up valuable computer cycles as well. In a portable device, due to battery consumption, wasting computer cycles is not an option. The research presented in this thesis shows the importance of a systematic feature selection step in EEG classification. The feature selection step produced excellent results with a maximum use of just 5 features. During automated classification, this is extremely important as the automated classifier will only have to calculate 5 features for each given epoch.

Acknowledgements

I would like to express my deep gratitude to Dr. Magdy Salama, Dr. Charles George, and Dr. George Freeman for the guidance they provided me with during my research and for allowing me to persue my graduate studies through the research grant from NSERC.

I would also like to thank Dr. Mohamed-Yahia Dabbagh for being one of my thesis readers and for providing me with excellent feedback. Finally, I would like to thank Dr. Rami Mangoubi, Dr. Paul Fieguth, Dr. Hamid Tizhoosh, and Dr. Tarek Abdel-Galil for the very helpful advice they provided me with during my research and Wendy Boles for all the administrative help she provided me with during the thesis writing process.

Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Introduction and Motivation

Patients with sleep disorders need to be monitored carefully by doctors, so that they can understand the underlying problem. In order to provide an accurate diagnosis, many different types of data can be collected from the patients. However, it has been established that the voltage activity that occurs in the brain of the patient is crucial for such an analysis. It is also highly desirable that any data collected be done in an as non-intrusive manner as possible.

To this end, electroencephalogram (EEG) signals can be measured from the scalp of the patient. With the use of this data, as well as other data channels, various features can be extracted in order to better understand the patient's situation. Medical experts typically divide the obtained EEG signal into 30-second intervals known as *epochs*. Based on the various features that can be extracted from such epochs and any pertinent contextual information, each such epoch can be classified into five different stages that will be discussed later. Medical experts are able to use this information, as well as other trends observed from the data, to assist in the diagnosis and management of patients.

As I found out during my interviews with members of the London Health Sciences Centre, the classification of such epochs is an involved task. Currently, doctors or trained technicians need to go through each epoch and make a classification manually, based on the features contained in the epoch, information contained in other channels such as electro-oculogram (EOG), and any pertinent contextual information. Considering a patient's one night sleep record could be approximately 8 hours, analyzing one sleep session could involve the analysis of approximately 960 epochs. Also, sometimes it is difficult to be consistent when making such classifications. For example, in most laboratories, an inter-human expert agreement of 90% is considered to be quite good. The inaccuracies that exist could be due to many factors. The ambiguity of the rules themselves, the difficulties present when extracting the features, fatigue of the technicians are to name a few of them.

One of the purposes of the teams' research is to eliminate some of that variability, in order to provide a more predictable classification. Given the same data, the goal is to develop an automated system that can perform the classification of the epochs with good accuracy. If successful, such a system will be extremely fast compared to a human scorer, and can help a human EEG technician immensely in analyzing the data. Furthermore, it is guaranteed that the automated system will never miss any of the features that were deemed pertinent when programming the system. As such, it can be a very important tool for a human scorer to check his or her work. Whenever, the human scorer's decision deviates from that of the expert system, the error can be analyzed and a decision can be made about whether to modify the system or the analysis of human scorer. This can assist in making the inter-human expert agreement higher as well.

A top level diagram of the proposed system is given in Figure 1.1. The first part of the thesis is to investigate the removal of artifacts from the raw signal. Traditionally, human experts only use high/low/band-pass filters before inspecting the signals. However, such filters are not directly useful in removing artifacts whose frequency content overlaps with those of desired components. For the purposes of removing the artifacts from the signals, there are two major tools used in this thesis; namely, Wavelets and Independent Component Analysis (ICA).

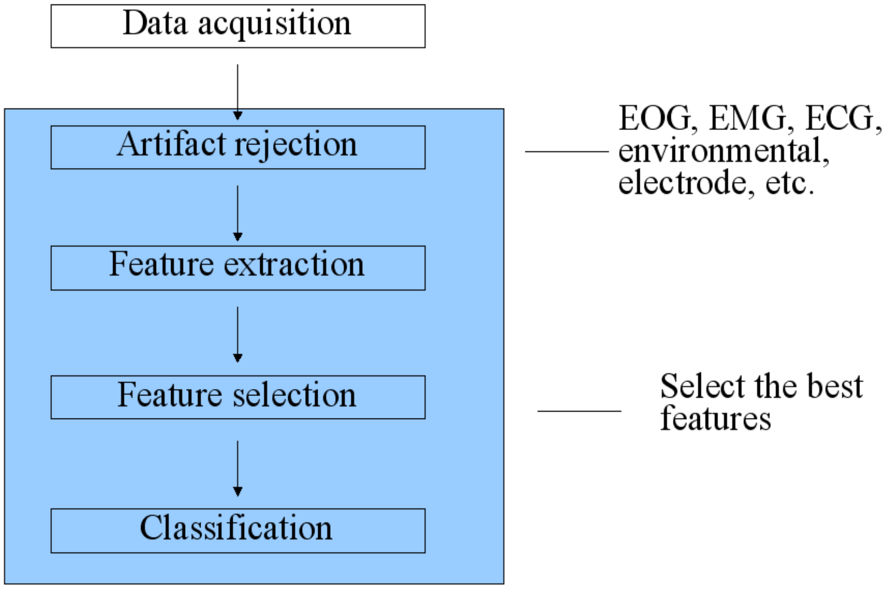

Figure 1.1: A top level view of the system

The second part of this research deals with feature extraction. In order to carry out the classification, it is necessary to extract the relevant features from the signals. To do this, it is important to establish *which* features are the best for separating the epochs into various sleep stages. If good features are left out, the classification will be poorer, and if useless features are included, it will waste computational cycles as well as potentially confuse the classifier.

When humans classify the signals to belong to the various sleep stages, they are able to ignore artifacts that they perceive to be extraneous before making the classification. The same ability needs to be given to an automated system before it attempts to perform the same task. The denoising step described in this research attempts to identify artifacts and remove them from the signal before further processing.

Once the raw signal is cleaned up, features need to be extracted to be used by the classifier. However, having irrelevant features or lacking important features will hurt the final classification. As a result, it is important to identify the most important features that should be used in the classification. This feature selection, is another aspect that is investigated in this research.

Obviously, the research also involves a use of a classifier. The details of the classifier will be discussed later in the thesis. It must be noted that the classifier itself is not the focus of this research and is not investigated in depth.

It is the hope of this research that a good classification rate can be achieved from a small subset of the initial feature pool. Selecting an appropriate feature subset has the potential to both improve the results and to improve the speed of the classification. Furthermore, identifying the artifacts and removing them from the raw signal, before feature extraction, is hoped to further improve the results.

1.2 Flow of the thesis

In Chapters 1 and 2 the necessary background information is given so that the terminology and the information needed for the rest of the thesis can be understood by the reader. After that, in Chapter 3 the research conducted by researchers in the industry is presented and discussed. At the end of the chapter, the motivation for the research presented in this thesis is discussed again and the connection to existing research is stated.

Chapter 4 discusses the methodology used in this thesis to identify and remove artifacts that occur in the EEG channel. Chapter 5 contains a discussion about feature extraction and the features in the feature pool. Chapter 6 discusses the role of the classifier and presents the pertinent information about the classifier used in this research. Since another member of the research team did her research on classification, the focus of this research is *not* the classifier.

Chapters 7 and 8 describe the experimental setup and present the results of the methodologies described in previous chapters. Chapter 9 summarizes the conclusions observed in this research and chapter 10 discusses how the research should be improved upon in the future.

Chapter 2

EEG and Artifacts

Before delving into the methods of removing artifacts, it is important to discuss the EEG signal and the different types of artifacts that are present in an EEG signal. They could be either physiological or nonphysiological in origin.

2.1 EEG Activity

The activity occurring in the brain can be measured in a variety of ways such as with EEG, magnetoencephalogram (MEG), and optical images. However, with MEG, the large magnetic sensors make it impractical to monitor freely moving subjects, such as moving patients. Also, optical imaging are constrained to surface events. The more invasive approach of deep brain wire electrodes, regardless of local accuracy, is not always suitable and might not be desired by some patients. EEG measurements, even though it works on a more macro scale spatially, is still quite effective. Furthermore, EEG allows for the assessment of cooperative neuronal activity at high temporal resolutions. [2]

Neurons within the brain produce currents that pass from the intracellular to extracellular space. Using EEG, the voltage generated as a result of this current can be measured. Many neurons contribute to the measured values and such, it gives us a macroscopic view of that local area.

The cerebral sources of EEG potentials are three-dimensional volumes of the cortex. The three-dimensional potentials produced by these sources are mapped to the two-dimensional surface of the scalp. Since EEG measurements are taken from the scalp, there is an inherent loss of information. In order to localize the actual underlying sources, it is important to appreciate and understand the physical and functional factors that produce the signals observed. [2]

The measured field potential is due to a variety of sources. Other than the expected synaptic activity, these include calcium spikes, voltage-dependent oscillations, and spike after-potentials observed in various neurons. However, the *principle generators* of EEG fields measured on the scalp are grade synaptic potentials: namely, excitatory and inhibitory postsynaptic potentials of pyramidal neurons. [2]

The field potential around an individual neuron is too small to be measured at the scalp. However, pyramidal cells are all aligned perpendicular to the surface of the cortex. If the activity of these cells are synchronous, the combined field produced by these cells would be large enough to be measured. The summation of potential fields resulting from synaptic currents can occur more readily than with other sources, due to the relatively long duration of the events causing the synaptic currents. [2]

However, it must be said that not all activity occurring in the cortex can be measured by scalp electrodes. The physical factors such as the source location, area, and orientation as well as functional factors such as the amplitude and frequency, determine the quality of the recording of a particular phenomenon. For example, if adjacent regions of the cortex have opposite orientations, the signals will cancel and no voltage field can be observed at the scalp. [2]

Another difficulty has to do with the localization of the source that generated the phenomenon of interest. Unfortunately, the original source of the phenomenon that is observed at the scalp might not be directly *underneath* the electrode used during the detection. Thus, attempting to explain certain phenomena as a result of the *brain segment* directly underneath the electrode might lead to incorrect analysis.

2.2 Measured signals

In order to perform sleep staging in any patient, it is recommended that signals from 2 electroencephalogram (EEG) channels, an electromyogram (EMG) channel, and an electro-oculogram (EOG) be used. These channels not only contain useful data, but noise elements as well. Having multiple EEG channels are essential in identifying sporadic activities in the signals [2]. The amplitude of normal EEG phenomena tends to be in the order of 20-50 uV. However, artifacts such as eye-movements and eye-blinks tend to be in the range of mV. The procedure in which these signals are collected is called a Polysomnogram (PSG) [3].

2.3 EEG and sleep staging

The human sleep cycle consists of five different sleep stages, and the awake state. The five sleep stages are Non-Rapid Eye Movements I-IV (NREM I-IV) and the Rapid Eye Movement (REM) stage. Examples of these sleep stages can be found in Appendix D. An EEG technician inspects the signal that is generated by the patient, and classifies each epoch (30 second interval) as belonging to one of the six stages in the sleep cycle. In this research, the data as well as the classified epochs were obtained from the Sleep Medicine Laboratory at the London Health Sciences Center.

The EEG technicians at the sleep lab generate the data from the sleeping patients during their stay at the Sleep Laboratory. Once the data is collected, the EEG technicians inspect the signals on their computers and classify the epochs accordingly. The computers have simple filters built in, such as high-pass and low-pass filters, to eliminate the most common forms of noise.

Features such as delta waves (0.5-4Hz), theta waves (4-8Hz), alpha waves (8-12Hz), beta waves (12-45Hz), and K-complexes are observed from the EEG recordings for the purposes of classification. The delta range was defined from 0.5Hz and upwards, in order to ignore potential sweat artifacts. On top of being used for classification, the EOG, ECG, and EMG channels can be used to identify artifacts as well. For example, with respect to classification, as a patient's sleep deepens from Stage I to Stage IV, his or her EMG activity lowers in amplitude and disappears completely during REM sleep [4].

In this data, whenever a proper diagnosis cannot be made for an epoch, it is automatically classified as being *awake*. In reality, this might involve non-awake situations such as movement of the patient while he or she is sleeping. The classification process is heavily based on the rules that Rechtschaffen and Kales have described in their manual. The actual rules will not be discussed in this thesis, as it was not explicitly used in the classification algorithms used in this thesis. However, there are researchers who model their classifier as rule based agents that use these rules explicitly.

It is very important to note that in terms of performance, the goal of laboratories is to have an inter-human expert agreement of 90%. Thus, a performance level of around 90% is a quite respectable score for an automated system.

2.4 Nonphysiological artifacts

Nonphysiological artifacts can occur from a variety of sources. They arise due to activity outside of the body and typically involves the electrode sites and environmental factors [2]. It is imperative that the EEG technician does everything in his or her ability to reduce such artifacts before the recording is performed.

2.4.1 Motion artifacts

Any movements of the patient can generate phenomena that can be observed by the scalp electrodes. The nature and the localization of the artifact observed on the scalp electrodes is dependent on the movement of the body part involved, the strength of the movement, and the relative location of the electrode wires. Parkinson's disease, myoclonic limb movements, nocturnal leg movements, and hypnic jerks are some examples of movements that can cause movement artifacts to appear on scalp electrodes. [2]

As I found out during my interviews in London Health Sciences Centre, the motion artifacts generated can be a significant source of noise. The noise generated can be as large as 14 mV and is usually contained within the 1 to 10 Hz range. These artifacts are readily visible in ECG, EEG, EMG, and impedance pneumography recordings.

The motion artifacts have two primary causes; namely, movements in the electrode metal-to-solution interface and skin-stretch. It has been demonstrated that with paste-filled recessed Ag-AgCl electrodes, motion artifacts due to electrode metal-to-solution interface are negligible. Thus, the majority of motion artifacts that occur are due to skin stretch.

To alleviate these artifacts, a number of approaches can be used. Abrading the skin at the electrode site and the use of electrodes that puncture the skin are a couple of notable solutions. Skin abrasion in particular require some experience on the part of the technician as too much abrasion can lead to skin irritation and too little abrasion will not reduce the noise significantly. [5]

2.4.2 Electrode pop

The *electrode pop* artifact is a nonphysiological artifact that occurs at the electrode-scalp junction with a slight electrode movement relative to the scalp. This movement causes a momentary change in the electrode-paste-skin interface that will produce the slight deflection in the recording. Proper scalp cleaning and electrode application can reduce the occurrence of this artifact [6]. Ideally, the electrolyte gel would absorb such motions, without changing the interface. In fact, [5] indicated that artifacts generated by electrode metal-to-solution interface movement is negligible when paste-filled recessed Ag-AgCl electrodes are used.

2.4.3 Sweat artifacts

Sweat artifacts manifest themselves as extremely low frequency signals typically in the range of 0.25 to 0.5Hz. While their amplitude can certainly affect the EEG recording, they can be easily removed with a high-pass filter due to their extremely low frequency range.

2.4.4 50/60 Hz noise

The 50/60Hz interference is a major cause of artifacts in the EEG signal. These induced voltages are due to the activity of nearby electrical equipment that operate at 50/60Hz. It is quite possible to measure a several volt difference from the human to earth-ground. Connecting a reference ground to a patient using an EEG electrode can significantly reduce the potential difference between the patient and the earth-ground [6]. While the differential amplifier used in EEG recordings *should* cancel out any uniform interference present in the human body from the two corresponding electrodes, impedance mismatches in the electrodes and the electrode-electrolyte-scalp interfaces prevents it from doing so perfectly. As a result, any reduction of the interference from within the human body is quite useful. Finally, it must be ensured that a patient is connected to only one ground at a time. Connecting a patient to more than one ground could be absolutely lethal to the patient due to the potential voltage difference [6].

That being said, the presence of 50/60 Hz does not create any significant problems with respect to EEG recordings. The frequency range of the signals of interest is typically well below the 50/60 Hz range. Therefore, applying a low-pass filter to the signal can get rid of these unwanted interferences quite easily.

2.5 Physiological artifacts

Physiological artifacts originate from sources inside the body, but not necessarily from within the brain [2]. The most notable physiological artifacts are due to the normal electrical activity of the heart, muscles, and the eyes. Of these, the ocular artifacts are the most relevant. [7]

2.5.1 Ocular artifacts

Eye movements that are recorded by a standard 10-20 montage are generated by the corneoretinal potential and the phenomenon created has an amplitude of approximately 50-100 mV [2]. In current data acquisition, ocular artifacts tend to be more dominant than other physiological artifacts (cardiac and muscle artifacts) and external interferences. [8]

The electrodes that detect the ocular phenomenon most prominently are the ones that are closest to the eyeballs; namely, Fp1, Fp2, F7, and F8. This is because ocular artifacts decrease rapidly as a function of the distance from the eyes [8]. The location of these electrodes can be seen in Figure 2.1. The ocular phenomenon is best regarded as a dipole where the positive pole is localized to the cornea and the negative pole is localized to the retina. [2]

The phenomena observed on the different channels vary significantly with the type of motion of the eyeballs. For example, when the eyes close, the movement of the eyeballs is in an upward direction. This is recorded as a positive potential with respect to the electrodes placed at

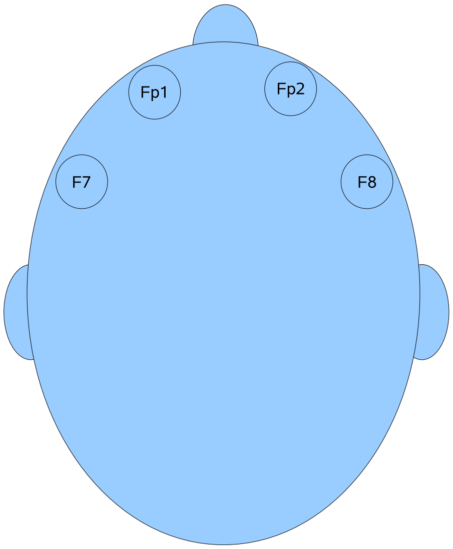

Figure 2.1: The locations of Fp1, Fp2, F7, and F8 on the scalp.

Fp1 and Fp2. However, when the eyes move to the left, the activity recorded at Fp1 and FP2 remain steady with no change in potential. On the other hand, the F7 electrode shows a positive deflection while the F8 electrode shows a negative deflection.

While the subject is awake, asking the individual to refrain from making eye-movements is obviously unrealistic. Even if a subject manages to consciously stop making any eye-movements, the mere fact that he or she is concentrating to do this will affect the eventual signal.

It should be noted that the EEG signal might contain pathological phenomena that might resemble ocular activity. Such activity should not be removed from the signal as they might be medically significant. It is important to identify such phenomena before removing potential artifacts [8]. To this end, one approach would be to verify that the unusual phenomenon is actually present in the primary channel meant to measure the artifact source in question; in this case the EOG channels.

2.5.2 Cardiac artifacts

The cardiac activity of a patient is easily monitored and can be accomplished by placing electrodes on the chest of the patient. The typical bipolar arrangement requires two electrodes to be attached to the left chest and the right chest, respectively. With respect to ECG artifacts, they usually occur in the EEG in referential montages, especially when using the ear electrodes as a reference. The field of the heart is oriented so that a negative polarity signal is produced on one side of the head and a positive polarity artifact is detected on the other side. ECG artifacts are more prominent in obese patients, patients with short necks, and babies; all these subjects have their heads close to the thorax.

Pulse artifacts are typically confined to a single electrode and usually occurs when placed over a surface artery. Such artifacts become most prominent when the electrode is loosely applied. The pulse artifact takes the form of a slow-wave potential and is time locked to the phenomenon on the ECG channel.

Artifacts generated from pacemakers take the form of high-voltage, short-duration spike activity and typically precedes the cardiac signal. Depending on the type of pacemaker, this type of artifact can be either continuous or intermittent. For further information regarding the topic of cardiac artifacts, please consult [2].

2.5.3 Muscle artifacts

When speaking of electromyographic artifacts, a number of different types of artifacts must be discussed. Lateral rectus artifacts are typically recorded from the F7 and F8 surface electrodes and has the form of a sharp positive deflection of very short duration followed by a slow falloff as the muscle relaxes. This type of artifact mimics the appearance of a calibration signal.

Single motor units can also be recorded by placing an electrode over one of the scalp muscles.

The appearance of the resulting artifact usually takes the form of a repetitive negative or positive deflection that takes a comb-like appearance. It is also possible for this type of artifact to occur transiently as single deflections that look random.

The frontalis electromyogram is recorded from the frontal electrodes and becomes present in patients who are contracting these muscles, such as when closing their eyes. These muscles are typically activated by photic stimulation and the amplitude of the phenomenon can be quite large; and as such, they can sometimes obscure EEG activity.

The temporalis EMG is recorded by placing the electrodes over the temporal lobe and usually occurs when patients tightly close their jaws or make chewing movements. Many of these artifacts can be reduced by ensuring the patient is relaxed. [2]

2.5.4 Glossokinetic

This form of artifact is produced by the movement of the tongue. The manifestation of this type of artifact is broad and can be recorded over the entire face or from frontal and temporal scalp areas. The artifact itself has a higher amplitude than the activity recorded on standard scalp electrodes, and is of low frequency. [2]

2.5.5 Respiratory artifacts

Respiration artifacts can also affect EEG measurements. Such artifacts can contain slow waves consistent with inhalation and exhalation or higher frequency activity due to snoring.

2.6 Summary

There are two main types of artifacts to be considered; namely, physiological and non-physiological artifacts. Non-physiological contain artifacts such as movement artifacts, electrode pop artifacts, sweat artifacts, and 50/60Hz noise. Typically, these artifacts are not explicitly monitored, and as such they need to be filtered out by their characteristics alone. For example, sweat artifacts tend to be of really low frequency, 50/60 Hz noise is contained within a narrow frequency band, and electrode pop artifacts are not necessarily time-aligned in two corresponding electrodes on the two sides of the scalp.

Physiological artifacts take the form of ocular artifacts, cardiac artifacts, muscle artifacts, glossokinetic artifacts, and respiratory artifacts. Most of these artifacts can be monitored with another channel, which in turn can be used during the denoising process of the EEG.

Chapter 3

Current research

There has been a significant interest into the areas of Sleep Stage classification and the removal of artifacts. This section describes some of the more relevant research to the work described in this thesis.

3.1 Sleep stage classification

To perform sleep staging, researchers have used a variety of techniques ranging from Neural Networks, Probabilistic models, Rule-based systems, and Fuzzy systems.

The research done in [9] describes a finite state machine that indicates the sleep stage with the use of Dempster-Shafer (D-S) theory. With the use of D-S theory, each of the hypothesis (sleep stages) are assigned a value of [0, 1]. This essentially indicates the belief in the hypothesis. D-S theory works by combining *evidence* to form the final belief in the hypothesis. In this system, there were a total of 130 rules to make the final belief set.

The *evidence* is dependent on the input characteristics of the type that will be discussed in Section 5.1. In [9], not all the features were used. Only the features, that were relevant to the corresponding sleep stage, were used when calculating the belief in the hypothesis. The relevant features were taken from literature and the researchers' knowledge. The features were not simply given to the D-S algorithm. Fuzzy Logic was employed to give a probabilistic weight to the input to describe *how well* it supports a particular sleep stage. The actual numbers in the fuzzification process was also based on the researchers knowledge of the sleep process and the accuracy of the detected waveforms.

Finally, contextual correction step was included to handle the nuances in the process. For example, some sleep stages don't exhibit a particular features *all the time*. However, if the previous epoch is classified to be of a certain type, the lack of evidence in the *current epoch* may not matter.

So, in essence, there were three stages: Fuzzification, D-S theory, and the application of Contextual Information. The researchers ran their algorithm against five sleep records and found the accuracy to lie between 78.44% and 90.6%, with a mean of 84.74%.

The research done in [4] uses a decision tree learning system to do the classification. In that research, the *recognition* of waves such as alpha waves, delta waves, sleep spindles, and K-complexes, are based on the directions, peaks, bottoms (negative peak), and durations. If the measured characteristics are *within* the predetermined limits specified for each type of wave, it is classified as belonging to that type. For a more detailed description of the procedure to *select the wave type*, please see [4].

When the waves are identified, general statistics about the number or the ratio of occurrences can be made. These features are subsequently given to the decision tree learning classifier. For this research, data from only one test subject was used. The researchers divided the data into five group randomly, and used four of the groups for training and one for testing. The experiment was repeated five times so that each group could be the test group. In four out of the five cases, the accuracy of the classifier had exceeded 80%. The mean classification rate for the whole experiment was 81.4%. Sleep stages that occupied most of the data stream, had a high accuracy rate, while sleep stages that have a low presence, had a relatively unimpressive accuracy rate (40% to 53%).

The research described in [10] uses a Neural Networks system for the classification of sleep stages. The system has three tiers that perform very different tasks. The first layer is called a *Sleep EEG Recognition Neural Network* (SRNN) and is responsible for the detection of several important characteristics waves in EEG.

The SRNN can recognize amongst five different characteristic waves; namely, (I) spindle, (II) hump, (III) alpha wave, (IV) slow wave that occupies 20%-50% duration of data segment, and (V) slow wave that occupies over 50% of data segment. The second layer is called the *Sleep Stage Diagnosis Neural Network* (SSNN) and is responsible for the actual classification of the sleep stage. The final tier of the system is called the *Contextual Diagnosis Neural Network* CDNN and is used for post-contextual correction.

For the experiments, 20% of all data was used as training data and the remaining data as the test data. The accuracy of the experiments yielded 82% agreement with the human expert.

The research that was described in [1] used a combination of Neural Networks and Genetic Algorithms. The Neural Networks were used for the purposes of classification, and the genetic algorithms were used for the selection of the optimal features from the feature pool and to find the optimal structure and initial weights of the Neural Network.

The initial feature pool consisted of 120 features in which 110 were by spectrum analysis, 7 features by statistical measure, and 3 using chaotic characteristics. Some of the features in the feature pool can be seen in Table 3.1. Of the set of features, 32 were chosen to be given to the Neural Network.

The experimental results indicated that the best features were the maximum power density in the α, β, θ, and δ bands and the *frequency* of the maximum power density. The constructed Neural Network had 32 input nodes, and 5 output nodes, and 1 hidden layer with 15 to 30

Table 3.1: Feature pool used by [1]

Type	**Examples**
Spectrum analysis	maximum power density, the frequency at the maximum power density, accumulated and relative power density, and the standard deviation of power density in the α, β, θ, and δ bands
Statistical measure	average amplitude, difference between the maximum and minimum amplitude, ratio between maximum and mean amplitude, standard deviation, maximum and minimum amplitude
Chaotic characteristics	fractal dimension of horizontally projected signal, box-counting dimension, and the second-order central tendency

hidden nodes. Of the structural optimization, only the number of hidden layers and the number of hidden nodes were variable elements. Unfortunately, this research paper did not indicate any concrete classification accuracy numbers.

There have also been hybrid classification solutions described in literature that attempt to integrate the best of different approaches. The research done in [3] wanted to demonstrate that a hybrid Rule-Based Expert System and a Neural Network can work well in conjunction. The Neural Network was essentially used to handle situation which might be difficult to handle with *just rules*. In this research, a multilayer feedforward network with two hidden layers were used with the error back propagation algorithm as the learning algorithm. The input of the Neural Network had 58 features. The reasoning given for the need for a Rule-Based expert system is that Neural Networks are not ideal for smoothing-rules. For example, the *3 minute rule* in EEG classification is heavily dependent on the epochs in the *vicinity* of the current epoch. Sometimes, this context is more important than the features directly observed in the current epoch itself. Therefore, the Rule-Based Expert System contains both *Single Epoch Reasoning* and *Multi-epoch Adjusting*. The Rule-Based Expert System has a notion of an overall reliability measure for all the decisions it generates. If there are any conflicts in the final decision, or the reliability measure is too low, the Neural Network system is used.

The researchers of this paper also did signal denoising before feature extraction. The signal denoising step involved the removal of ECG interference from the EEG channel, removal of harmonic noise at 20 Hz and 60 Hz using notch filters, and the removal of low frequency voltage due to sweat. They were also concerned with the fact that the traditional use of band powers calculated over the whole epoch does not necessarily give all the information regarding the epoch. For example, when the power is averaged over the whole epoch, the temporal resolution is completely eliminated. To determine the power statistics, for example, they divided the EEG epoch into 30 segments, and calculated the hamming windowed FFT over each segment to determine the desired statistic.

The experiment for this research yielded an accuracy of 83.1% with the use of just the Rule-Based expert system and an agreement of 85.9% with the hybrid system. These numbers are rather impressive considering the setup of the experiment. Of the 4 test subjects in the experiment, 2 were used exclusively for training purposes and the other 2 were used exclusively for testing purposes. Therefore, the generalization factor was relatively high. Interestingly, the researchers were only able to get an accuracy of just 55.1% when only the Neural Networks are used. They admitted that the use of various Neural Network architectures yielded similarly poor results and concluded that Neural Networks by themselves are not appropriate for sleep stage scoring.

3.2 Removal of artifacts

There has been a fair bit of research towards investigating how to remove artifacts in the EEG signal. The authors of the paper [7] concerned themselves with the removal of ocular, cardiac, and muscle artifacts from the EEG signal. The context of an artifact is sometimes dependent on the sleep stage. For example, awake and REM stages of sleep usually involve the contamination of the EEG signal with ocular artifacts. On the other hand, in some other sleep stages, EEG phenomena, such as K-Complexes, interferes with the EOG channels more noticeably. Such bidirectional mixing makes methods based on regression analysis difficult to utilize effectively with respect to ocular artifacts.

In their work, the EEG signal was recorded from 19 electrodes on the scalp. The goal was to apply Independent Component Analysis (ICA), discussed in B, with the end-goal of eliminating unwanted artifacts. In their work, they used the ICA variant *Algorithm for Multiple Unknown Signal Extraction* (AMUSE) for the separation of the mixtures into their independent components. This algorithm uses the time-structure of the signals instead of just assuming that the signals are generated by random variables. When the independent components that represent an ocular, cardiac, or muscle artifact are found, with the use of time, frequency, and scalp topography details of the independent components, they can be eliminated prior to the reconstruction of the denoised EEG signal. With the prior knowledge of the artifacts being investigated, as well as expected corticle activity, such comparisons between the templates and the separated independent components can be made.

The experimental analysis claims good results of separation. However, the authors did not publish any results as to how the denoising step affected sleep stage classification. Also, the large number of channels that were available to these researchers essentially means that the denoising problem they worked on is significantly different from the denoising problem analyzed in this thesis.

In order to remove EOG artifacts, time-domain and frequency-domain regression methods have been used [8]. Time-domain regression assumes that the propagation of ocular potential is volume conducted, frequency independent, and without any time delay. However, it has been argued that the scalp is not a perfect volume conductor and that some frequencies are attenuated more than others. Neither techniques, however, take into consideration the propagation of brain signals into the EOG channels. Also, the correction coefficients used are typically different for eye-blinks and eye-movements. [8]

A method based on Principle Component Analysis (PCA), when applied to the same problem, has outperformed the above mentioned regression methods. Unfortunately, it has performed poorly when the amplitudes are of comparable size [8]. In essence, PCA attempts to uncorrelate a set of given signals by using 2^{nd} order statistics. It should be understood that uncorrelating the mixtures is not as strong as making them statistically independent from each other. A more detailed description about PCA can be found in B.3.

Wavelet based techniques have also had success in removing ocular artifacts. Since ocular artifacts reside in the low frequency bands and is large in amplitude, thresholding the coefficients of the wavelet decomposition that are above a certain value would hopefully remove the artifact while keeping the original EEG signal relatively undisturbed [8]. However, the authors of [8] didn't quite say whether the results obtained were verified with EEG experts to judge the quality of the denoising process.

A paper by Brown et al described the possibility of statistical wavelet thresholding. In this approach, assuming EEG activity follows a somewhat normal distribution, coefficients that deviate from the normal distribution is rejected. Unfortunately, this approach failed to improve baseline drift, eye movements, and step artifacts. [11]

Haas et al published a paper which attempted to remove EOG artifacts by using an ARMAX (AutoRegressive Moving Average with eXogenous inputs) model. This model is used to model the recordings as a linear combination of EEG and EOG activity. By estimating the parameters of the model, it was the intention to locate the EOG artifacts and then to remove them. While this method was successful in removing some EOG artifacts not removable by standard EOG techniques, it is more computationally expensive and might introduce new EOG artifacts. [11]

Extended Kalman filters have also been used in an Adaptive Autoregressive (AAR) setting to filter out the artifacts. Once the parameters of the model are identified with the use of an Extended Kalman Filter, adaptive inverse filtering is applied to filter out the artifacts. The results indicated that the method performed better with muscle and movement artifacts than EOG or ECG artifacts. [11]

Independent Component Analysis (ICA) has been used successfully to separate a multi-channel scalp recordings into physiologically plausible independent components [8]. For example, in [11] the authors successfully decomposed an artificial mixture of EEG, EOG, and EMG signals into their independent components. However, it should be noted that these mixtures were artificially created by the authors. The performance against natural mixtures originating from the human body was not performed. Also, the obtained results does not seem to have received expert evaluation to verify the quality of separation.

Another research team has done work on using ICA to reject artifacts as well. They have used both simulated and real data to evaluate their method. The simulated data were obtained by artificially mixing channels recorded from the corticle surface of the human going through presurgical evaluation. Overall, the artifacts due to ocular activity was removed from the signal-set. Unfortunately, the quality of the decomposition does not seem to have been evaluated by domain-experts. Also, the channel-set that was used during the decomposition had 20 channels. So, while ICA can be successfully used when there are a large number of channels, its performance for a low number of channels is still not known conclusively. [12]

There have also been work done to simply identify artifact sources. Reference [13] described such a system that achieved approximately 90% accuracy rate with respect to domain-experts in identifying the presence of artifacts. Conceivably, such a system can be used by an expert system that performs sleep stage classification.

Most importantly, in [14], wavelets were used as a visualization tool by the researchers, to visualize the decomposed levels as a set of *time-series* in order to locate artifacts in Partial Discharge (PD) signals. This time-series reconstruction of the decomposed levels forms the basis of the wavelet solution presented in this paper. In this thesis this method will be modified to tackle the problem of locating artifacts from EEG.

3.3 Motivation

One can conclude from the above survey, with the exception of [1], there hasn't been much work on feature selection. Even in [1], 32 features were given to the Neural Network. It would be interesting to determine the performance of a Neural Network classifier when the number of features are significantly less.

Also, [3] stated that Neural Networks by themselves are not effective classifiers. Even though the system they proposed was very powerful indeed, their conclusion regarding Neural Networks seems premature.

As was seen from the above survey, there have been some work done to denoise EEG signals with ICA. However, they always seemed to use much more channels than used in this research. Therefore, it would be interesting to determine if the implementation of ICA used in this research is able to denoise EEG signals effectively when the number of channels is small.

As mentioned before, the Wavelet based solution described in this thesis is an improvement over the algorithm presented in [14]. Naturally, such an extension should be investigated.

Chapter 4

Removing artifacts

Once the artifacts are identified, it is necessary to remove them while keeping the effect on the desired signal to a minimum. For this purpose, there were three main techniques investigated in this paper: namely, the basic Wavelet denoising algorithm (BWDA), the Wavelet denoising algorithm (WDA), and the Independent Component Analysis denoising algorithm (IDA). It should be noted that IDA actually uses Wavelets in the initial stages of the algorithm. This research mainly looked at EOG artifacts that occur in the EEG channels and artifacts isolated to a single EEG channel, such as electrode pop artifacts.

4.1 Tools used

For the purposes of denoising the EEG signals, Wavelets are used within all the algorithms. When denoising artifacts, localization in the time-frequency axis is essential. The artifacts have particular frequency properties and they only occur *some of the time*. As a result, it is highly desirable to inspect a signal as a function of *both* time and frequency. Wavelets are excellent for this purpose. Since a *scale* in a wavelet decomposition can be mapped to a particular frequency, localizing in the time-scale axis is equivalent to localizing in the time-frequency axis. As a result, the algorithms used in this thesis use wavelets to great effect in localizing various artifacts.

Wavelets are also excellent at selectively suppressing artifacts. When an artifact is located in the time-frequency axis, it can be easily suppressed through a process called *Wavelet Thresholding.* Due to the localization in time, only that local area is affected; and due to the localization in frequency (through scales), waveforms of different frequencies can be suppressed independently.

However, the physiological model of the mixing that takes place at the scalp describes the mixing process as being linear. Thus, each *scale* separated by the wavelet decomposition is a linear mixture of the corresponding scales of the wavelet decompositions of the original sources. By only suppressing a few select scales by using wavelet thresholding, the other scales are essentially ignored. However, the mixing model indicates that those scales are mixtures of the original sources as well. Due to this scenario, instead of using Wavelet thresholding to denoise the artifacts, Independent Component Analysis (ICA) can be used. ICA assumes a linear mixing model of the sources and attempts to *demix* them to the original components. Even with the use of ICA, Wavelets are still used to locate potential artifacts.

More details about Wavelets can be found in Appendix A and about Independent Component Analysis in Appendix B.

4.2 Denoising with Wavelets (BWDA)

The types of artifacts that are considered in this research are EOG artifacts and *sporadic* artifacts that occur in the EEG channel, mainly due to nonphysiological issues. Both forms of artifacts are localized in time and are of low frequency. As a result, the capability of Wavelets to inspect the signal on the time-scale, and in turn on the time-frequency resolution is quite desirable.

The most straightforward approach to using Wavelets to denoise the EEG signals is to inspect the corresponding coefficients in the time-scale axis of different channels and to suppress them when necessary. The algorithms used for this purpose are seen in Tables 4.1 and 4.2.

For example, when the EOG channel and the EEG channel are compared, time-locked and large Wavelet coefficients in one of the higher scales, in both the EEG and EOG channel could potentially signal EOG contamination. When such a time-scale aligned coefficients are detected, the coefficient in the EEG signal can be suppressed. The threshold values that are needed to detect potential contamination can be established experimentally. The Wavelet algorithm based on this approach described within this section is named the Basic Wavelet Denoising Algorithm (BWDA).

4.2.1 Removing mixed biological artifacts using Wavelets

If the artifact in question is a biological artifact, a channel meant to measure the *source* of the artifact could be quite useful. For example, in order to remove EOG artifacts, the EOG channels can be used to select the location of potential contaminations in the EEG time-series. When any ocular activity is observed in the EOG channel, the EEG channel can be observed for a similar phenomenon.

The BWDA algorithm described in Table 4.1 compares the time-scale aligned wavelet coefficients of the EEG and the artifact channels before suppressing the necessary coefficients in the EEG signal. It is quite likely that such a correlation between the EEG and the artifact channel occur due to a contamination from the artifact channel into the EEG channel.

4.2.2 Removing sporadic artifacts using Wavelets

If the artifact that needs to be removed is an electrode/site related artifact or some sporadic waveform, then it is highly likely that it is present in only one channel. In this research, there are two EEG channels available that should be extremely correlated. When these two

Table 4.1: Algorithm for mixed artifact removal using Wavelets (BWDA)

1	Decompose EEG channel T_{eeg} and artifact channel(s) into N levels using an appropriate mother-wavelet.
2	Select the first/next artifact channel as the *current artifact* $T_{artifact}$ to be used in all the remaining steps in the algorithm. Each artifact is processed individually.
3	By inspecting the EEG signal and artifact signal, determine which levels contribute noticeably to the current artifact - L.
4	For each level of the decomposition listed in L, determine an appropriate wavelet coefficient profile to indicate the presence of artifacts. In each profile, obtain the amplitude values H_{eeg} and $H_{artifact}$, experimentally, such that when a coefficient value in T_{eeg} is greater than H_{eeg} and the corresponding coefficient in $T_{artifact}$ is greater than $H_{artifact}$, an artifact is said to have occurred.
5	Using the wavelet coefficient profiles generated in the previous step, compare each set of corresponding coefficients in the time-scale axis, for the scales in set L. If the EEG coefficient and the artifact coefficient satisfy the threshold profile, set the coefficient in the EEG decomposition to zero.
6	If more artifact types are present, armed with the updated wavelet coefficients for the EEG channel, goto Step 2 and process the next artifact.
7	Reconstruct the final set of wavelet coefficients for the EEG channel to generate the *denoised* EEG signal ($eeg_{current}$).

channels are compared, any *significant* discrepancy between them could potentially signal an unwanted artifact. It must be stated that there is an inherent difference in amplitude between the two hemispheres of the brain. Thus, when comparing the two EEG channels, it should be understood that their amplitudes won't be necessarily approximately equal. For example, difference in the skull thickness of the patient can account for voltage asymmetries of 20% to 70% and also mask or simulate abnormalities. Without having actual measurements about the skull thickness, asymmetries of less than 50% is usually diagnosed as being insignificant. Due to the interest of low frequency waveforms in this thesis, it should be noted that *transient* asymmetries of vertex sharp waves are common and normal. However, a significant asymmetry that *persists* is abnormal and suggests a cerebral disturbance lateralized to the side of lower voltage. [2]

Therefore, with consideration to the scaling differences and transient phenomena already discussed, the two corresponding channels should be approximately scaled versions of each other. Since it is quite difficult to tell the difference between valid transient phenomenon and sporadic artifacts, this thesis will only attempt to remove low frequency asymmetries that seem to have some *persistency* in an epoch. The algorithm for this purpose using wavelets is presented in Table 4.2.

Table 4.2: Algorithm for sporadic artifact removal using Wavelets (BWDA)

1	Decompose primary EEG channel ($T_{primary}$) and the secondary EEG channel ($T_{secondary}$) channel(s) into N levels using an appropriate mother-wavelet. For the automated analysis, the primary channel will be used solely to extract the features.
2	Using $T_{primary}$, $T_{secondary}$, and the nature of the sporadic artifacts, determine the levels that contribute noticeably to the artifact phenomenon - L.
3	For each level of the decomposition listed in L, determine an appropriate threshold profile to indicate the presence of artifacts. In each profile, obtain the amplitude values $H_{primary}$ and $H_{secondary}$, experimentally, such that when a coefficient value in $T_{primary}$ is greater than $H_{primary}$ and the corresponding coefficient in $T_{secondary}$ is less than $H_{secondary}$, a sporadic artifact is said to have occurred.
4	Using the wavelet coefficient profiles generated in the previous step, compare each set of corresponding coefficients in the time-scale axis, for the scales in set L. If the EEG coefficient and the artifact coefficient satisfy the threshold profile, enter them into the $L_{sporadic}$ list.
5	If $L_{sporadic}$ list contains sufficient entries, for a period of Y epochs, set the coefficient in the primary EEG decomposition of those entries, to zero.
6	Reconstruct the final set of wavelet coefficients for the EEG channel to generate the *denoised* EEG signal ($eeg_{current}$).

4.3 Denoising with Wavelets (WDA)

The approach used in Section 4.2 considers the coefficients, on the time-scale grid, individually. This approach might not reflect the actual phenomenon since the *duration* of certain artifacts might be longer than what is indicated by the implied frequency of using a particular scale. Taking higher levels in the decomposition, which averages out the neighbouring coefficients even more, *might* satisfy the length of the artifact. However, using these coefficients for comparison might not be ideal, since some of the detail present in the lower layers is lost if they are not used in the comparison explicitly.

To resolve these limitations, each scale can be *reconstructed* into a separate time-series; and then each resulting time-series can be segmented appropriately in the time axis. This method also makes it much easier to visualize potential artifacts by the naked eye than with the use of raw Wavelet coefficients directly. This forms the basis of the algorithm which is described in this section. The reconstructed time-series differs from the original time-series in the sense that most of the extraneous elements are discarded.

The method described in this section follows the work done by L. Satish and B. Nazneen [14]. In that paper, Wavelets were applied for the purposes of reducing noise and unwanted interference present in Partial Discharge (PD) signal measurements.

That research is relevant to the problem discussed in this thesis, since in that research, much like with my own research, the interference and the desired signal had overlapping spectral properties. In [14], the process is not completely automated. The reconstructed Wavelet decompositions were essentially used to allow the human researcher to better visualize the signal and its many components. The methodology essentially decomposes the signal into an appropriate number of scales, and then *reconstructs* a time-series from each scale. For example, if the PD signal was decomposed into eight levels, the resulting algorithm would produce a set of nine time-series corresponding to the coefficients at the eight detailed scales and the one approximate scale. From this set of time-series, the researcher can select which reconstructed scales contain artifacts, and which do not. This should be straight-forward to do as phenomenon resembling the artifact shape is easily recognized and can be eliminated within each scale. The details present in other scales, that overlap in the time-axis will be preserved.

In the present EEG research, it is possible to automate this selection procedure as well. In [14], the human's expertise was required to discriminated between desired phenomenon and artifact phenomenon. In the EEG research, the nature of the artifacts and their locations can be approximated through the use of *other channels* that are available to us.

4.3.1 Removing mixed biological artifacts using Wavelets

Since the mixing of other biological artifacts with the EEG is instantaneous, any contamination should essentially overlap in time on the EEG channel. Since the ocular activity on the EOG channel, and any related artifacts on the EEG channel are time-locked, simply checking for the degree correlation is sufficient to verify the presence of an artifact. The algorithm used to detect any potential mixed artifacts with the Wavelet method is presented in Table 4.3.

4.3.2 Removing sporadic artifacts using Wavelets

As described before, if the artifact that needs to be removed is an electrode/site related artifact, then it is highly likely that it is present in only one channel. Normal phenomenon that occurs in the two EEG channels used in this study should be extremely correlated. As discussed before, it is also possible for valid *transient* vertex sharp waves to occur without signaling any abnormalities [2]. When these two channels are compared, any *significant* and somewhat persistent discrepancy between them could potentially signal an unwanted artifact. As with Section 4.3.1 the signals are divided into segments prior to carrying out the denoising. The algorithm for the removal of sporadic artifacts using wavelets is presented in Table 4.4.

4.3.3 Issues with using wavelets for artifact removal

The selection of the amplitude thresholds requires some work as there are no classification data available for each *segment* of an epoch. Also, the selection of the mother-wavelet and the number of levels in the decomposition is an important issue. Within this research, only a limited number of mother-wavelets were considered. The work done by [8] indicates that the 'coif3' mother-wavelet is appropriate due to it's close resemblance to an eye-blink artifact. As

Table 4.3: Algorithm for mixed artifact removal using Wavelets (WDA)

1	Decompose EEG channel and artifact channel(s) into N levels using an appropriate mother-wavelet.
2	Reconstruct N+1 time-series (T_{eeg}) from the decomposed coefficients of the EEG signal - N from the detail coefficient levels and one from the remaining approximate coefficient level.
3	Reconstruct N+1 time-series ($T_{artifact}$) from the decomposed coefficients of the first/next artifact channel - N from the detail coefficient levels and one from the remaining approximate coefficient level.
4	Using T_{eeg} and $T_{artifact}$, determine which levels contribute noticeably to the artifact phenomenon - L.
5	Divide these levels into M segments each and let the i^{th} segment of channel *foo* be denoted by T_foo_S_i. Through trial and error, using the *training set* only, determine the minimum amount of *correlation* (C_{min}) present between T_eeg_S_i and T_artifact_S_i whenever the artifact is present in the i^{th} segment of the EEG channel. Also, determine an appropriate amplitude profile, (A), of the segment from the *EEG* channel. This could be simply the maximum height within the segment.
6	Using the correlation and amplitude profiles generated in the previous step, compare the i^{th} segment in the EEG channel and the artifact channel of the *testing set*. If the correlation is above the C_{min} value and the amplitude profile is a match, a artifact is assumed to be detected.
7	If an artifact was detected in the EEG segment, set that segment to zero. Else, retain the current segment without modification.
8	If more artifact types are present, armed with the updated T_{eeg}, goto Step 3 and process the next artifact.
9	*Add* all the levels of T_{eeg} to generate the final *denoised* EEG estimate ($eeg_{current}$).

a result, the same mother-wavelet was used in this research as well.

The number of levels within the decomposition depends on both the size of the data and the resolutions of interest. It was found in this research that setting $N = 5$ gave sufficient resolution to pin-point potential artifacts within only a single level. Increasing the number of levels to a large number has an effect of creating very low frequency DC-like waveforms in the highest scales and do not tell anything useful. On the other hand having a very low number of levels in the decomposition, would not give the decomposition sufficient frequency resolution. In such

Table 4.4: Algorithm for sporadic artifact removal using Wavelets (WDA)

1	Decompose primary EEG channel and the secondary EEG channel channel(s) into N levels using an appropriate mother-wavelet. For the automated analysis, the primary channel will be used solely to extract the features.
2	Reconstruct N+1 time-series x 2 (denoted by $T_{primary}$ and $T_{secondary}$) from the decomposed coefficients of both the primary and secondary EEG signals.
3	Using $T_{primary}$, $T_{secondary}$, and the nature of the sporadic artifacts, determine the levels that contribute noticeably to the artifact phenomenon - L.
4	Divide these levels into M segments each and let the i^{th} segment of channel *foo* be denoted by T_foo_S_i. Through trial and error, using the *training set* only, determine the *maximum* amount of *correlation* (C_{max}) present between T_primary_S_i and T_secondary_S_i whenever there is an artifact present (discrepancy between the two segments). Also, determine an appropriate amplitude profile, (A) of the segment from the *EEG* channel. This could be simply the maximum height within the segment.
5	Using the correlation and amplitude profiles generated in the previous step, compare the i^{th} segment of the primary EEG channel and the secondary EEG channel of the *testing set*, for all i. For each segment, if the correlation is below the C_{max} value and the amplitude profile is a match, an artifact is said to have occurred.
6	If an artifact was detected in the primary EEG segment, enter the segment into the $L_{sporadic}$ list.
7	If $L_{sporadic}$ list contains sufficient entries, for a period of Y epochs, set those segments to zero.
8	*Add* all the levels of $T_{primary}$ to generate the final *denoised* EEG estimate ($eeg_{current}$).

a case, the suppression of a level due to the presence of an artifact has the negative effect of suppressing useful detail as well. By trial and error, it was found that setting $N = 5$ gave the best resolution to isolate the artifact in the time-series reconstructed from the different scales.

4.4 Denoising with ICA (IDA)

In section 4.3, we discussed how wavelets can be used as a tool to denoise artifacts. Once the regions of interest were discovered, and analyzed, the appropriate regions were set to zero.

Now, instead of setting those regions to simply zero, ICA can be used to demix those *segments* of interest. Therefore, a system consisting of both wavelets and ICA can be constructed and its performance evaluated. It is important to note that only the segments that are *flagged* are demixed using ICA.

ICA essentially allows us to separate out an estimate of the artifact from the desired signal. If there are useful information in the same frequency range, setting the whole segment to zero would eliminate useful information as well. Given two signals, even if there are frequency overlap, ICA has the ability to distinguish the contribution of each source component to the observed signals, and thus identifying the desired signal.

4.4.1 Removing mixed biological artifacts using ICA

This algorithm is a modified version of the algorithm found in Section 4.3.1. Only the latter number of steps are different from the original algorithm. The modified partial listing of the algorithm can be seen in Table 4.5.

Table 4.5: Algorithm for mixed artifact removal using ICA (IDA)

7	If an artifact was detected in the EEG segment, demix the *full EEG segment* (that generated the wavelet decomposition) with the *full artifact channel* segment. Notice that these *full* segments are the original segments that generated the current set of wavelet decomposition. This is in contrast to a normal segment which refer to a segment of the *time-series* generated from a particular scale of the wavelet decomposition.
8	Once all the segments are evaluated, regenerate T_{eeg} from the fully assembled EEG epoch ($eeg_{current}$).
9	If more artifact types are present, armed with the updated T_{eeg}, goto Step 3 and process the next artifact.

4.4.2 Removing sporadic artifacts using ICA

Much like in Section 4.4.1, the ICA algorithm is built upon the wavelet counterpart. The algorithm presented in this section is a modified version of the algorithm in Section 4.3.2. The modifications can be seen in Table 4.6.

Table 4.6: Algorithm for sporadic artifact removal using ICA (IDA)

7	If $L_{sporadic}$ list contains sufficient entries, for a period of Y epochs, demix the *undecomposed* primary EEG segment (from the measured signal) with the *undecomposed* secondary EEG segment. From the demixed output, keep the one that does not have the artifact. Once all the segments are evaluated, the final EEG epoch is implicitly created.

4.4.3 Relevancy of data size

It is important to select an appropriate size of data when using the Independent Component Analysis algorithm. The phenomenon that needs to be removed only occurs *some* of the time. If the data segment that is given to the ICA algorithm is too long, the emphasis will not be put on the phenomenon in question. Any significant statistics that are observed surrounding the waveform will simply get averaged out over the length of the large dataset. Of course, having too little data is also poor and becomes a classic over-fitting problem. When this is the case, the observed statistics are not reliable and it will be difficult to calculate any meaningful statistics. To give an analogy, there would be too many unknown parameters with too little equations keeping them together. Thus, it is important to establish an appropriate window (segment) length before applying the ICA algorithm. The actual window length that is used in this research will be given later in the thesis.

4.4.4 Other important notes

It is important to note that the desired signal and the artifacts that needs to be removed can be thought of as *independent components* in a certain sense. Obviously, they are not purely independent in the strictest sense since various phenomenon at different parts of the body are related. For example, when a person sees an object that he or she really desires, EEG waveforms might contain certain characteristics patterns. At the same time, the heart might also beat faster. However, the activation of related phenomenon in the different parts of the body are not necessarily *time-locked* to each other. This aspect of these related biological signals can be used to classify them as independent components in the *time-locked* sense. Appendix B gives more details about ICA.

Also, in this research, the data-set that is available for analysis has only 6 channels. There are 2 EEG channels, 2 EOG channels, an ECG channel and an EMG channel. This is less than the number of distinct sources that are observed on the human brain. Traditionally, when Independent Component Analysis is applied to decompose brain signals, there are much more channels available to the researcher. As discussed in Appendix B, if the number of independent sources in the system is greater than the number of measurements, it is not possible to find an accurate decomposition without further constraints. This is analogous to having more variables than equations when trying to solve a math problem. As a result, each *variable* can take more than one legal value. Without sufficient constraints in the form of channels, the *decomposed* signal can take many *legal* forms. Naturally, this is not acceptable. Thus, to obtain good results, it is important to use as many quality channels as possible.

4.4.5 Choice of ICA method

The optimization algorithm that was chosen to perform Independent Component Analysis in this research was the *FastICA algorithm.* Details of this algorithm is discussed in detail in Appendix B. The *FastICA algorithm* was chosen due to it's ability to process batch-data effectively. Simulated experiments with the use of various random generators showed that the FastICA algorithm is capable of separating the mixtures into the original independent components. Also, it was shown in [15] that it was possible to decompose mixed EEG signals into their independent components when the number of channels available are approximately 15.

4.4.6 Choice of higher-order statistics

As described in Appendix B, the choice of the higher-order statistic is crucial for the proper separation of the independent components. The higher order statistics are in essence used as an *approximation* to a true measure of independence. Naturally, it is not an *exact* approximation, but given the application, hopefully a sufficient one. For this thesis, the higher-order statistic used was the Negentropy of the function as described in Appendix B.5.4. The G(y) function is given in Equation 4.1.

$$G(y) = -exp(-y^2/2) \tag{4.1}$$

4.4.7 Selecting the denoised EEG

After ICA is used to demix the raw signals, a set of *demixed* signals is produced. Unfortunately, deriving which *demixed* signal is the EEG signal is not straightforward. This is because, during the ICA algorithm, the notion of order is not preserved.

As a result, in order to select the EEG signal, a post-identification step must be performed. The algorithm that was developed in this research for this purpose is presented in Table 4.7.

The candidate demixed signal chosen essentially is the signal that is *most unlikely* to any of the artifact signals. In this algorithm, the notion of *most unlikely* is based on the summation

Table 4.7: Selecting the EEG signal from the set of demixed signals produced by ICA.

```
N=# of signals
FOR j=2 to N (raw signals - assume signal #1 is unclean EEG)
    FOR i=1 to N (ICA demixed signals)
        C = Find correlation of $j^{th}$ raw signal with $i^{th}$ ICA demixed signal
    END
END

Sort matrix C, such that C(j,i) tells you *how close* the $i^{th}$ ICA separated signal
is to the original $j^{th}$ signal.

FOR i=1 to N (ICA demixed signals)
    FOR j=2 to N
        L(j-1)=Find how close the $i^{th}$ ICA separated signal is to the
        $j^{th}$ original signal (artifact) with respect to the other ICA separated
        signals (rank based measure).
    END
    D(i) = sum of *closeness* of the $i^{th}$ demixed signal to the *raw signals* (artifacts).
END

Select the $k^{th}$ demixed signal, such that min(D) = D(k)
```

of the relative ranks as seen in Table 4.7. The quality of this cost function with respect to other possible cost functions needs to be investigated further in future work.

The ICA demixed signals could not simply be checked against the original EEG raw signal, since the original raw EEG signal might contain the artifact. And as such, an ICA demixed signal that *contains* some artifact might provide a high correlation to the original EEG signal, since the original EEG signal might contain that artifact. As a result, the candidate demixed waveforms were compared against the raw *artifact* channels, with the hope that the signal that resembles the artifact channels the *least*, with respect to the cost function defined in Table 4.7, is the denoised EEG signal.

In future work, instead of using the *summation* of the ranks, simply using the *highest* rank, when comparing ICA separated signals against the original artifact channels, *must* be considered. This is because, the ICA separated signals ideally should only contain a single artifact. And as such, considering the other ranks after the primary match-up, can affect the result negatively.

Chapter 5

Feature Selection

Feature Extraction is the process of obtaining certain descriptions of the data that might be more readily used for classification purposes. The types of the features as well as the number of features selected and given to the classifier affects the final outcome greatly. *Feature Selection* is a process that sits between Feature Extraction and Classification stages that attempts to *prune* and *select* the most relevant features from the initial feature pool before giving it to the classifier. The mathematics of the feature selection, will be discussed later in this chapter. In this work, features extracted from EEG, EOG, and EMG channels are considered.

The need for feature selection is two fold. Firstly, having unnecessary features can make the classification accuracy lower by confusing the classifier. This is analogous to information overload. When the classifier tries to tune the parameters, it's more difficult when there are useless information to process and integration into the system. More unknown parameters require more data to give a similar level of confidence in the classifier. If some of the features, that we know are not very useful, can be left out, the classifier will have an easier time tuning itself. Secondly, feature selection will also make the classification process go faster. When there are more features to extract during the classification stage, each epoch that needs to be classified needs to generate more features. If there are too many features that are needed, it might be the case that there isn't enough time to classify the data in a real-time fashion.

This section describes a systematic approach of determining the number and the types of features that should be selected for optimum classification. As a first task, the data is described and the types of appropriate statistical techniques used for feature selection are explained. Once the algorithms are performed, the nature of the optimal features can be determined.

The nature of the features selected would naturally depend on the classifier selected for the task of classification. In this thesis, the main classifier used was the *Conjugate gradient Back-Propagating Neural Network*.

The intention of this work is *not* to specify how the different features discussed in the R&K manual fit together. Rather, it is to *give* the *important* features to the classifier, and let it learn their relationships effectively based on the classified data. If features that are representative of the features that were used to do the classification using the R&K rules are extracted, it should be expected that a good classifier be able to find the proper connection amongst those features to satisfy the classified output.

5.1 Feature types considered

From the set of measured signals (time-series), many features can be extracted. For the purposes of this research, the features described in Table 5.1 are extracted. The features that are wavelets coefficients were retrieved with the use of the 'coif3' mother-wavelet.

5.2 K-complex detector

A K-Complex is a high amplitude, low frequency, diphasic wave that usually occurs during Stage 2 sleep. Since it is a prominent waveform in Stage 2 sleep, checking for its presence is useful when performing sleep stage classification. [2]

Table 5.1: Extracted features

	Channel	**Name**	**Type**
1	EEG	Delta band (0-4Hz) (EEG)	Wavelet coefficient
2	EEG	Theta band (4-8Hz) (EEG)	Wavelet coefficient
3	EEG	Alpha band (8-12Hz) (EEG)	Wavelet coefficient
4	EEG	Beta band (12-45Hz) (EEG)	Wavelet coefficient
5	EOG	Delta band (0-4Hz) (EOG)	Wavelet coefficient
6	EOG	Theta band (4-8Hz)(EOG)	Wavelet coefficient
7	EOG	Alpha band (8-12Hz) (EOG)	Wavelet coefficient
8	EOG	Beta band (12-45Hz) (EOG)	Wavelet coefficient
9	EMG	Delta band (0-4Hz) (EMG)	Wavelet coefficient
10	EMG	Theta band (4-8Hz) (EMG)	Wavelet coefficient
11	EMG	Alpha band (8-12Hz) (EMG)	Wavelet coefficient
12	EMG	Beta band (12-45Hz) (EMG)	Wavelet coefficient
13	EEG	# K-complex	Estimates the presence of K-complexes
14	EEG	Most prominent frequency	Number indexing the most prominent frequency band
15	EEG	Total power	Power of all the frequencies of interest

Unfortunately, there aren't any concrete amplitude and frequency guidelines in determining a K-Complex. Simply finding the low-frequency parts of the signal is not sufficient as there are many phenomenon that would have low frequencies, but with significantly different shapes from those of a K-Complex. In order to detect their presence, the *shape* of the waveform is critical. As a result, it was decided to apply a *Template Matching* algorithm in order to identify the general shape of the candidate waveform.

The *features* to match between the candidate waveform and the templates were simply taken to be actual points of the respective curve. Since the width of a K-Complex is not set in stone, *Dynamic Time Warping* was used in the *Template Matching* algorithm as well. Using dynamic time warping in the template matching algorithm allows for some slack in the time-axis of the template. This way, the horizontal scaling factor of template matching will become relatively insignificant. This allows the candidate waveform to be more fluid in the time-axis and doesn't force it to take a more rigid shape. The significant downside to this algorithm is that it takes a noticeable amount of the computational cost of the *whole* algorithm.

5.3 Features and the Classifier

Since the discrimination of various sleep stages might be best suited from different feature sets, it was decided to create a set of binary classifiers that would be able to discriminate a particular class against all the rest. As a result, the optimum feature set would be calculated with respect to each of the six classes. The approach to do this will be discussed later.

5.4 Maximum Significant Difference and Independence

In order to select the optimal features, the idea of *Significant Difference* (SD) and *Feature Independence* (FI) will be used. Significant Difference is a statistical measure of the ability of a particular feature to discriminate between various classes [16]. When a candidate feature is applied to the training data, its effectiveness at separating the different classes can be estimated. The features that have high Significant Difference figures have a very good potential to be selected in the final feature set.

Feature Independence on the other hand checks for the interdependency of *different* features [16]. This is important since two features that have high Significant Difference, might be highly correlated to each other. If they make the same decisions, then having both of them is redundant. The idea is to select features that complement each other, that will work together to provide a *better* classification, and not simply the same classification provided with fewer features.

Since it was decided to have different classifiers to specialize in the identification of each sleep stage, the data was divided into two classes to train *each* classifier in question. This way, the optimum features can be extracted from the data to discriminate against that individual sleep stage.

There are numerous statistics that are especially catered to data that have normal distributions. The tests are quite powerful *if* the data satisfies the normality conditions. Otherwise, the tests are meaningless. If the normality conditions are not met, then more general statistical tests, such as rank based tests, can be performed.

If the data follows a normal distribution, the conclusions made by the tests that assume normality will be more *precise* than the tests that make no such assumptions. Of course, if the data used in this research do not satisfy the normality conditions, the statistics based on the normality assumptions cannot be used. In fact, when I applied the *Bera-Jarque parametric hypothesis test of composite normality* (JBTEST) to the data, I found that none of the training groups satisfied the normality conditions. As a result, it was decided that only rank based statistical tests be used for the analysis of the features.

In order to establish the *Significant Difference* of the features with respect to their ability to successfully discriminate the class in question, the *Mann-Whitney test* is used. Details about this statistical test can be found in Appendix C. The formulation of the Mann-Whitney test was obtained from [16] and is presented in Equation (5.1).

$$\begin{aligned} Z &= \frac{\Big|R_s - E\big(R_s\big)\Big| - 0.5}{\sqrt{Var\big(R_s\big)}} \\ \text{where } E\big(R_s\big) &= \frac{n_s(1+N)}{2} \\ \text{and } Var\big(R_s\big) &= \frac{n_s n_m}{12}(1+N) \end{aligned} \tag{5.1}$$

Here, Z is the level of significant difference, R_s is the sum of the ranks of the elements in the class with less elements, n_m is the number of elements in the class with more elements, and $N = n_s + n_m$ is the total number of elements.

To establish the level of *Feature Independence* between any two candidate features, the *Spearman Correlation* is used. Details of this statistical test is also found in Appendix C. It must be noted that the *Pearson Correlation* is not appropriate for the experiments, since as was shown by the JBTEST, the data, in any of the classes, does not follow a normal distribution. The Pearson Correlation inherently assumes the data to have a normal distribution.

The top-level Maximum Significant Difference and Independence (MSDI) algorithm used in [16] was also used here, and is shown in Table 5.2.

Table 5.2: MSDI algorithm

1	Create an empty set: selected-features
2	Compute the Significant Difference (SD) of each of the candidate features and insert into set {sd-set}
3	Select the features with the maximum SD from {sd-set} and insert it into the {selected-features} set. Delete the same entry from {sd-set}.
4	Calculate the Significant Level (SF) of *each of the features* in {sd-set} with respect to the features in the {selected-features} set.
5	Select the feature with the largest SF value from Step 4, and insert it into {selected-features} set. Delete the same entry from {sd-set}.
6	If the maximum number of features are selected: Exit. Else, goto Step 4.

In step 4 of the MSDI algorithm in this thesis, each feature in {sd-set} is compared with each of the features in {selected-features} individually to determine the *maximum correlation* of *each* feature in {sd-set} with *some* feature in {selected-features}. These *maximum correlation* (C_M) values are related to the FI values by the formula shown in Equation (5.2). For example, when the maximum correlation is closer to zero, the feature independence is closer to one and thus is extremely high.

$$FI = \sqrt{1 - C_M^2} \tag{5.2}$$

Also, SD is simply equal to Z in Equation 5.1. Once the FI and SD values are calculated, the SF value can be calculated with (5.3) [16]. This formula allows us to combine the Significant Difference and Feature Independence into one convenient number.

$$SF = SD\text{x}FI \tag{5.3}$$

5.5 Mann-Whitney approximation

As mentioned before in Section 5.4, the Mann-Whitney test is used to establish the *Significant Difference* of the features with respect to their ability to successfully discriminate the two classes in question. In each of the six classifiers, one of the two groups is a collection of data points from *multiple* classes. The other group is obviously the single class that needs to be discriminated successfully (C_{main}). It should be noted that even though the Mann-Whitney test is a rank-based test and does not assume normality, it does assume that the distributions are identical and only differ in the mean. In this work, the two distributions in each classifier are certainly not identical. But, since the test is rank based, a good result can still be obtained by intelligently separating the classes to more than two groups and testing the relevant pairs separately.

The group that contains five classes, will most likely contain a much larger spread and a more complex distribution than the group with a single class, C_{main}. If the Mann-Whitney approximation is used on these two groups as is, the performance should be expected to be quite poor. Figure 5.1 might illustrate this point visually. The two distributions shown in this figure illustrates the values of one of the candidate features used in this experiment for the two different groups.

Here, the green distribution shows the values of some feature, from class C_{main}. The blue distribution illustrates the values of the same feature from all the other classes. Since the rank

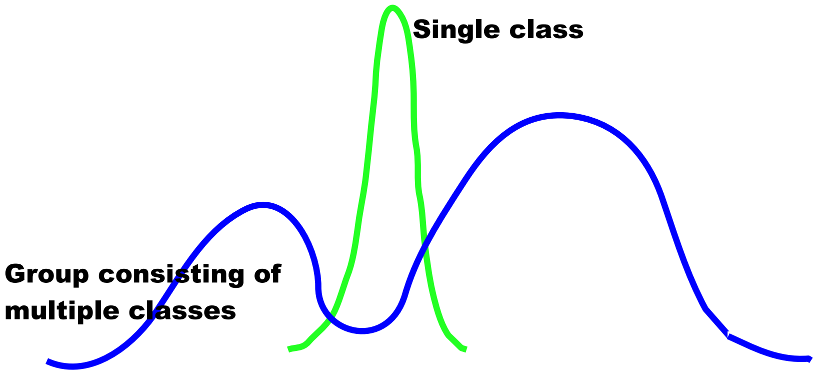

Figure 5.1: The need for the Mann-Whitney approximation

of points in the blue distribution is on both sides of the rank of points in the green distribution (C_{main}), the Mann-Whitney test will not obtain an accurate result as the sum of the ranks of the feature values in the blue distribution could very well *average out* to a sum that could be produced from the points from the green distribution, assuming the same number of points are obtained.

Therefore, for the purposes of this thesis, the feature values illustrated in the blue distribution were further subdivided by class, so that the Mann-Whitney test was performed *twice* for each of the candidate features within each of the six classifiers. Within each classifier, for each feature f_i, the first Mann-Whitney test compared the main class being discriminated against (C_{main}) with the group that consists of all the classes whose individual feature mean ($mean(f_i)$) is *less* than the feature mean ($mean(f_i)$) within C_{main}. The second Mann-Whitney test compared the main class being discriminated against (C_{main}) with the group that consists of all the classes whose individual feature mean ($mean(f_i)$) is *greater* than the feature mean ($mean(f_i)$) within C_{main}. The final Z measure is simply a weighted average of the two calculated Z measures. While this modification is not perfect, it is much better than using the default groups.

5.6 Monotonically Increasing Curve

A subsequent post-processing stage can be applied to the features selected by the MSDI step, to validate the improvement in performance. The algorithm described in [16], called Monotonically Increasing Curve (MIC), is used in this research and is described in Table 5.3.

Table 5.3: MIC algorithm

1	Sort the features selected by the MSDI step from the best feature to worst, using an appropriate sorting criterion (e.g. SF, SD, FI)
2	Plot the performance curve (classification) using the features selected by the MSDI step. The x-axis corresponds to the number of features used in the sorted array.
3	Delete the *left-most* feature that contributes negatively to the performance. i.e. the *first* feature that causes the performance curve to be *not* monotonically increasing.
4	Re-plot the performance curve with the updated feature list.
5	Goto step 3, until the curve is monotonically increasing or until the maximum number of iterations are reached.

Obviously, this is not an *optimal* solution, but it certainly has potential to yield some improvement over simply using the MSDI algorithm. It is not optimal since this algorithm does not take a global view of the features, and so there is no guarantee that it will find the *perfect* features to delete. However, due to computational limitations, this thesis will use the algorithm as described in Table 5.3.

Chapter 6

Classification

Among several numerical classification methods, it is believed that Artificial Neural Networks are one of the most attractive techniques for sleep stage classification [3]. Neural Networks in general are a wide-spread tools for the task of classifying patterns.

A Neural Network is a collection of processing units called *neurons* connected together to form a larger network. The identity of the Neural Network is defined by both the properties of the neurons themselves and the nature of the interconnections *between* the neurons.

In this thesis, the emphasis was feature extraction and denoising of artifacts. Therefore, different techniques of classification were not explicitly investigated. After straightforward trial and error, it was decided that a *Conjugate Gradient Back-Propagating Neural Network* (CGBNN) would be sufficient as the classifier, for the investigation of the effects of feature extraction and artifact removal.

The CGBNN used in this research is a *feed-forward* neural network and uses backpropagation to adjust the weights between its neurons. Thus, during the training phase, the errors calculated between the network output and the expected output is used to further adjust the weights within the network. Given such an error, the CGBNN uses a conjugate gradient formulation to determine the subsequent search direction for the parameters (weights) in question [17]. A diagram of a feed-forward neural network with one hidden layer is seen in Figure 6.1. The neural network in the figure has three input neurons, four neurons in the hidden layer, and one output neuron.

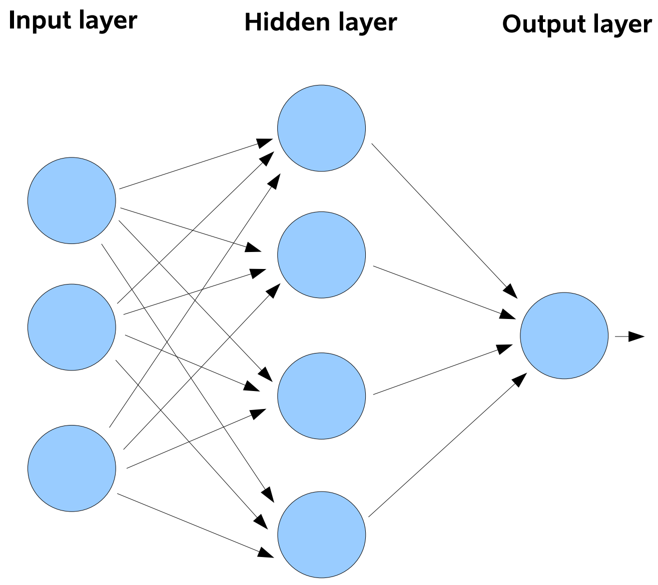

Figure 6.1: Feed-forward neural network

A conjugate gradient algorithm essentially uses a combination of the current gradient and previous search directions to determine the new search direction. In contrast, a basic gradient descent algorithm will simply use the current gradient. The CGBNN used in these experiments is the Powell-Beale version of the conjugate gradient algorithm. The Powell-Beale version of the algorithm has two important properties. First, it *resets* the search direction to the negative of the current gradient, whenever a particular condition becomes true. Second, whenever the condition is not true, it uses a combination of the current gradient, the previous search direction, and the last search direction before the previous reset, to calculate the new search direction. This algorithm is already defined in MATLAB as *traincgb*, and was used in the experimentations. [17]

Unless otherwise mentioned, any reference to a neural network made in this thesis refers to the Conjugate Gradient Back-Propagating Neural Network described in this section. Since a Neural Network is used as the classifier in this research, the aim is that given the features, the classifier will deduce the rules based on the data. In contrast, a rule based system would involve more direct involvement from the researcher in setting up the rules.

Using the research done in [1], it was decided that a single hidden layer with 15 to 30 hidden nodes would be a good starting point for the classifier used in the experiments. Of course, the number of input features in that research was significantly higher than the aimed number of features in this research. Thus, it is to be expected that a fewer amount of hidden nodes would also yield similar or even better results. Using the results of [1] as a basis, it was seen that 20 hidden nodes in the single hidden layer provided good results. Also, the transfer functions used in the two layers were the Tan-Sigmoid Transfer function (tansig) and the Linear Transfer Function (purelin), respectively. As mentioned before, the classifier itself was not investigated extensively, as it is not the focus of this research.

Chapter 7

Experiments

In this section, the experiments that uses the above mentioned techniques are described. The comparisons between the effects of the various techniques showcase the various strengths and weaknesses of each approach. The denoising step in this research mainly looked at EOG artifacts and artifacts isolated to a single EEG channel, such as electrode pop artifacts. Artifacts such as sweat artifacts were easily eliminated since their frequency ranges were well defined and could be eliminated easily.

7.1 Experimental Setup

For the experiments, data from two subjects are used. The information about the two subjects are seen in Table 7.1.

Table 7.1: Subjection information

	Sex	**Age**	**# of epochs**
Subject 1	Female	23	604
Subject 2	Female	33	900

The experiments were performed to gauge the effectiveness of applying a *Feature Selection* stage into the feature extraction step and applying denoising techniques prior to feature extraction. Unless indicated otherwise, each *experiment* with a fixed number of selected features were run for 10 iterations and the results were averaged. The experiments that investigated the effects of varying the number of features were run a number of times until a clear trend could be observed. For illustration purposes, from all the test-runs for each experiment, a single candidate that is representative of each experiment was chosen and included in the thesis.

7.2 Measuring quality of denoising

For the purposes of this thesis, the *quality* of the results were judged by comparing the results generated by the automatic classifier with the results generated by the EEG technician. As mentioned before, it is important to note that in terms of performance, generally, the goal of sleep laboratories is to have an inter-human expert agreement of around 90%.

Chapter 8

Results

This chapter describes each experiment and discusses the results obtained.

8.1 Without feature selection nor denoising

In this section, the separation of all six classes was attempted. As described before, each class was discriminated against all other by a *dedicated* Neural Network. This essentially allows feature selection optimizations to be performed on the classifiers *individually*, instead of being forced to apply any global optimizations. The details about the Neural Network are given in Chapter 6.

The results of this experiment are given in Table 8.1. As we can see, the performance is quite poor.

Table 8.1: Classification results: no feature selection, no denoising

	Accuracy						
	NREM I	NREM II	NREM III	NREM IV	REM	Awake	Total
F23	10.8%	54.8%	14.1%	18.8%	25.2%	7.8%	36.1%
F33	1.6%	51.4%	1.3%	63.9%	34.4%	21.0%	39.4%

There is an inherent difference between the training set and the test set. If there is a feature that does not discriminate between the classes, but is considered by the classifier, a poor result will be obtained. This can be seen more clearly in Figures 8.1 and 8.2 by observing the performance level as a function of the number of features selected. It seems that the performance level, when all the features are selected, is extremely low. Some of the poor features that do a bad job at classifying the data quite possibly do not capture the trends very well.

The performance seems to improve drastically as only the most relevant features from the feature pool are selected through the feature selection step described earlier in the thesis. This experiment demonstrates the importance of the feature selection step in EEG classification.

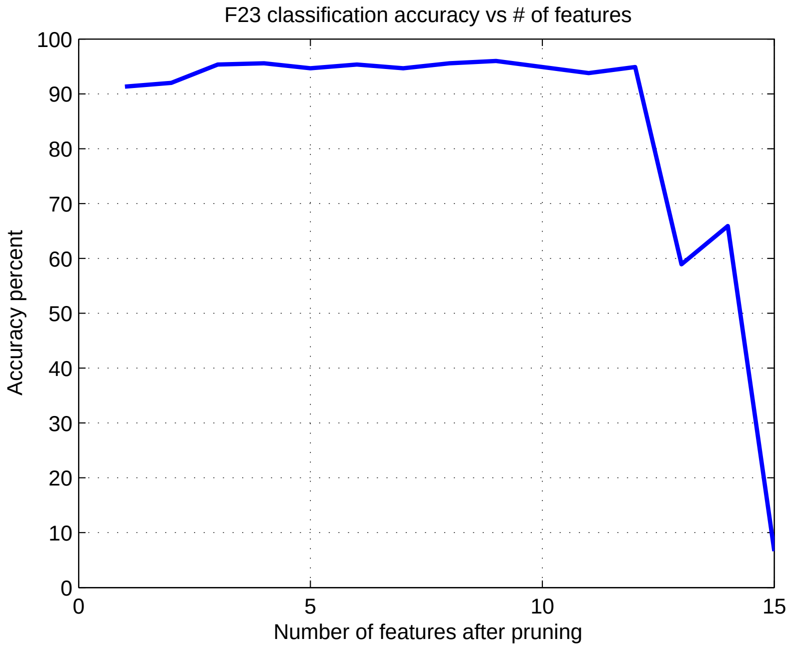

Figure 8.1: Accuracy vs number of selected features for F23

From these graphs, it was determined that using approximately 5 features would yield good results. Therefore, for the purposes of this thesis, 5 features were given to the classifier for both subjects. As discussed before, many prior research done by other researchers simply used the initial feature pool in their classifiers without any appropriate feature selection. In the feature

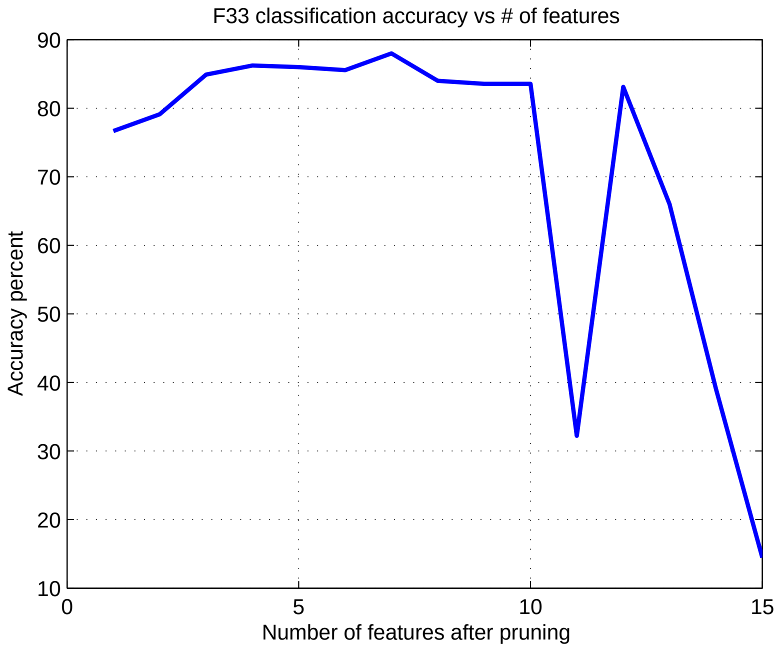

Figure 8.2: Accuracy vs number of selected features for F33

pool used in this experiment, it can be seen that there are some features that are extremely poor. However, in larger features pools, it can be the case that many more features yield poor results. So, it is greatly desired that for EEG classification, a feature selection step is included.

8.2 With feature selection but without denoising

After setting the number of features selected to be 5, the original experiment with 10 iterations was repeated. The performance of that subsequent experiment is seen in Table 8.2. Clearly, the performance was *significantly* improved.

Table 8.2: Classification results: feature selection, no denoising

	Accuracy						
	NREM I	NREM II	NREM III	NREM IV	REM	Awake	Total
F23	74.6%	94.3%	61.1%	89.4%	91.1%	82.8%	87.7%
F33	29.1%	93.8%	22.5%	99.1%	90.5%	83.0%	84.8%

8.3 Denoising with BWDA

Now that the importance of the feature selection step has been investigated and established, the effect of introducing the *denoising* step can be investigated. For this task, the wavelet decomposition used the 'coif3' mother-wavelet and used 5 levels in the decomposition. This resulted in 5 detailed levels and 1 approximate level.

As discussed before, EOG artifacts are the most prominent type of artifact that affect EEG signals. In this section, the effect of removing EOG artifacts and sporadic artifacts with the use of BWDA was investigated. Performance of this algorithm can be seen more clearly in Figures 8.3 and 8.4. The thresholds were selected such that the artifact in Figure 8.3 would be denoised optimally. As we can see, the epoch in Figure 8.4 was not denoised at all. The original *eeg* signal and the post-denoised *modified eeg* signal look identical. The artifact present in the 6^{th} segment is still present. With this algorithm, tuning the thresholds proved to be a difficult task. The thresholds that work for some segments don't easily work for others. With the WDA algorithm, selecting a successful threshold proved to be an easier task. The methodology used in this experiment (BWDA) was discussed in detail in Section 4.2.

Now, the effectiveness of the BWDA denoising step with respect to automatic classification was investigated. The performance numbers can be seen in Table 8.3.

Inspecting the results shows that the performance of the automated classifier is marginally worse for the F23, but slightly better for the F33 patient. The decrease in performance for the F23 patient is most likely well within the expected error in this experiment. It seems

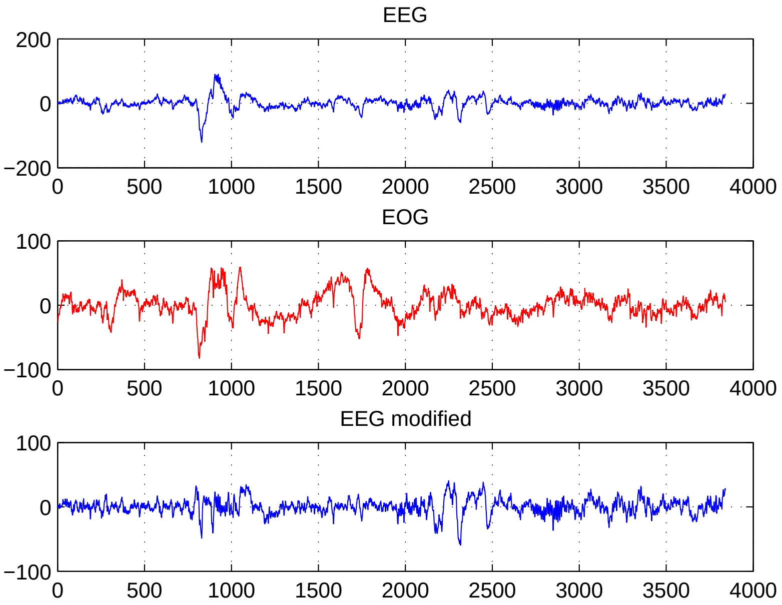

Figure 8.3: BWDA denoising example (1). Artifact present in the second segment is suppressed in EEG modified.

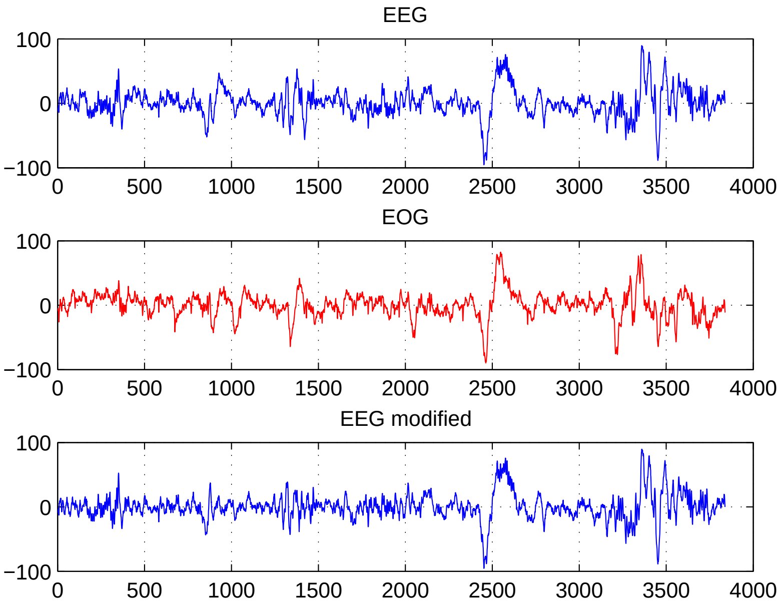

Figure 8.4: BWDA denoising example (2). Artifact present in the sixth segment is not suppressed in EEG modified.

Table 8.3: Classification results: feature selection, BWDA denoising

	Accuracy						
	NREM I	NREM II	NREM III	NREM IV	REM	Awake	Total
F23	76.3%	95.2%	63.7%	84.6%	88.6%	79.7%	87.5%
F33	33.7%	95.2%	15.0%	98.9%	90.8%	84.6%	85.9%

that denoising the signal to eliminate low-frequency artifacts does not have much impact on the automated classification, especially considering the BWDA algorithm missed a number of denoising opportunities as was seen in Figure 8.4. Out of the 15 features, only a few features are affected by the denoising step. With respect to assisting the human technician to identify artifacts, the BWDA algorithm does not perform reliably. The performance level of the other two algorithms that were described earlier in the thesis will now be discussed.

8.4 Denoising with WDA

The methodology used in this section was discussed in detail in Section 4.3. For this task, the wavelet decomposition used the 'coif3' mother-wavelet and used 5 levels in the decomposition. This resulted in 5 detailed levels and 1 approximate level.

The performance and a visualization of what the WDA denoising step produces can be more readily seen in Figures 8.5 - 8.7. As discussed before, the algorithm essentially decomposes the signal into an appropriate number of levels, and then *reconstructs* a time-series from each level. For example, since the EEG signal was decomposed into five levels, the resulting algorithm would produce a set of six time-series corresponding to the coefficients at the five detailed scales and the coefficients of the one approximate scale. From these new time-series, it is easier to establish *which* scales are most representative of the contamination observed. If a contamination is found, it is simply suppressed in the relevant scale. By segmenting the signal into many smaller segments, the suppression can be made to be extremely local. Therefore, the suppression of a potential artifact will only affect the area surrounding the immediate artifact and not the rest of the epoch. The other details present in other scales, that overlap the artifact in the time-axis will be preserved. In the graphs shown,the magenta shows the actual signal while blue shows the *time-series* generated from the coefficients of the wavelet decomposition.

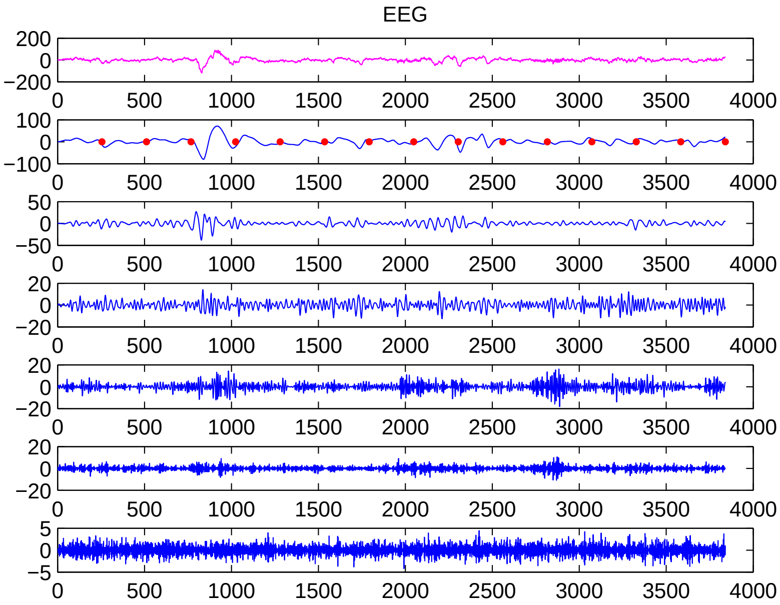

Figure 8.5: WDA denoising example (Set 1) - original EEG signal. From top to bottom [original EEG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

We can clearly see an EOG contamination in the EEG in segment 2 (delimited by the red dots). After running the algorithm, the *EEG modified* graph has eliminated that artifact by setting the segment within the approximate coefficients to zero. That activity in the *EEG modified* graph is simply the activity from the other scales. It should be noted that there is a difference in the *y-axis* resolution between the original EEG graph and the modified EEG graph. Another example of the Wavelet denoising procedure can be seen in Figures 8.8 - 8.10. It can be clearly seen that the EEG signal, indicated in magenta on Figure 8.8, contains an ocular artifact on segments 5 and 6. After the denoising step, that artifact is no longer present in the *modified EEG* graph, shown in magenta on Figure 8.10.

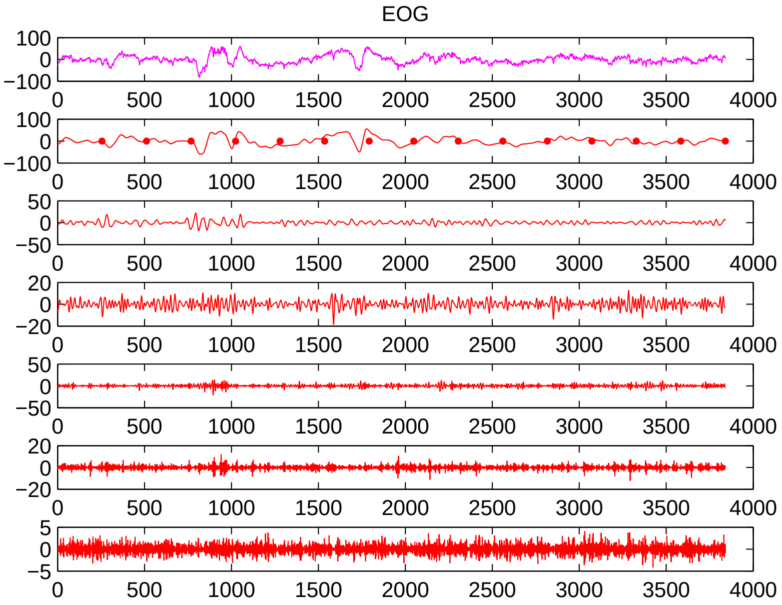

Figure 8.6: WDA denoising example (Set 1) - EOG signal. From top to bottom [EOG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

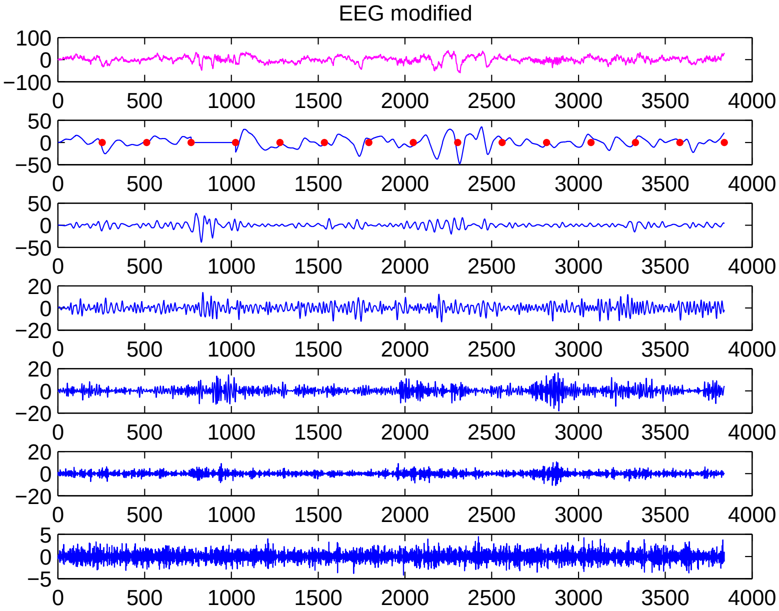

Figure 8.7: WDA denoising example (Set 1) - Denoised EEG signal. From top to bottom [Denoised EEG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

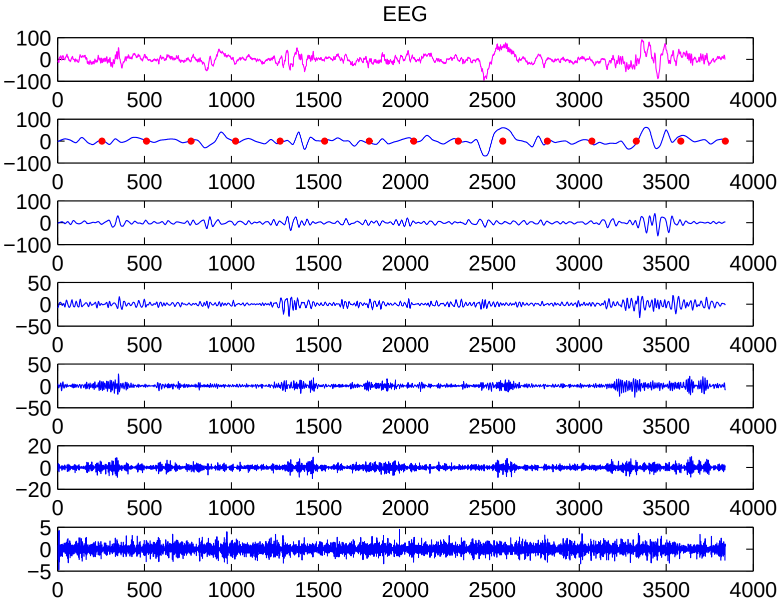

Figure 8.8: WDA denoising example (Set 2) - original EEG signal. From top to bottom [original EEG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

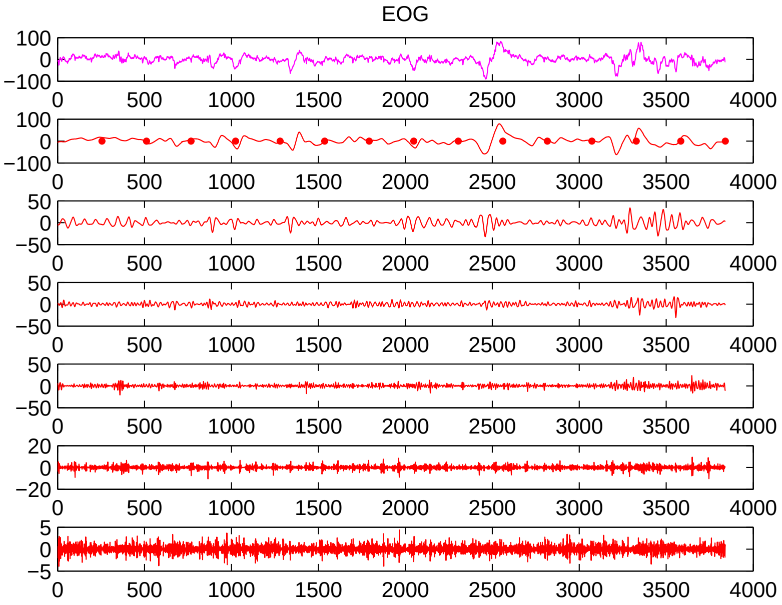

Figure 8.9: WDA denoising example (Set 2) - EOG signal. From top to bottom [EOG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

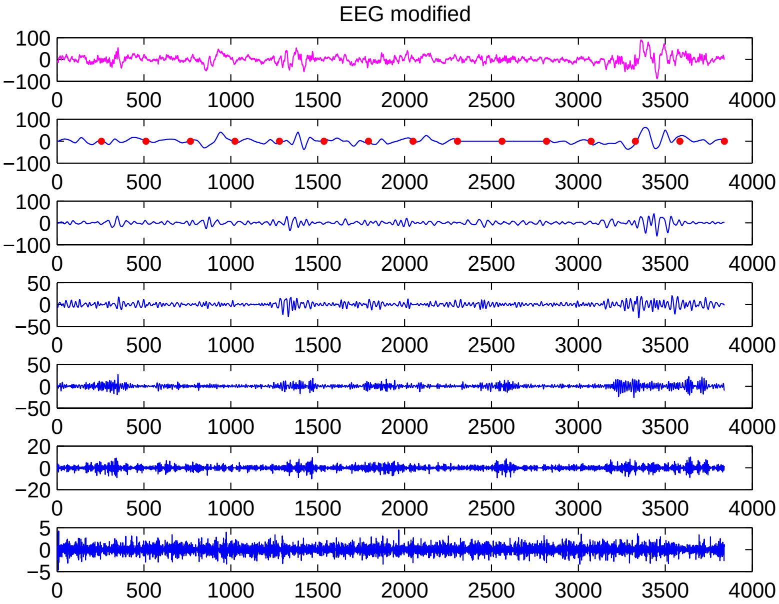

Figure 8.10: WDA denoising example (Set 2) - Denoised EEG signal. From top to bottom [Denoised EEG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

Clearly, this denoising step is quite effective in eliminating EOG artifacts. The EOG artifact that was identified in the original signal is no longer present in the modified signal. It is much easier to select the WDA thresholds over BWDA thresholds, since the actual artifact shape is more readily visible in WDA levels. With WDA, segmentation of the epoch had to be used since the time resolution within the *reconstructed* level was identical to the original epoch. As such, the duration of any potential artifacts had to be considered through the segmentation. With the BWDA method, segmentation was not used as the coefficients in the highest levels capture the nature of a whole *neighbourhood* of the area of interest. Lastly, varying the segment length and the number of levels of the Wavelet decomposition in the WDA algorithm gives much more control than just varying the number of levels in the Wavelet decomposition of the BWDA algorithm. Since a researcher can vary both the number of levels and the window (segment) length, the WDA offers more control during the denoising step. Furthermore, for a human, it just seems more natural to set the appropriate thresholds using the time-series reconstruction provided by the WDA algorithm.

Now, the effectiveness of this denoising step with respect to automatic classification was investigated. The performance numbers can be seen in Table 8.4.

Table 8.4: Classification results: feature selection, WDA denoising

	Accuracy						
	NREM I	NREM II	NREM III	NREM IV	REM	Awake	Total
F23	75.0%	93.6%	64.8%	88.5%	90.0%	85.3%	87.8%
F33	29.3%	94.4%	35.0%	99.8%	90.5%	84.4%	85.6%

The performance of the automated classifier, after denoising, improved slightly, but not significantly. This could be due to the fact that the *improvements* are limited by the occurrence rate of the artifacts and the limited impact the artifacts have on most of the features. While the effectiveness of the denoising can be clearly observed from the graphs, the statistical features that are extracted from the data does not seem to change as much when the automated classifier is concerned. However, the effectiveness of this denoising algorithm is still very valuable for an actual human EEG technician. With a cleaner signal, an EEG technician will have a much easier time at classifying the epochs.

8.5 Denoising with IDA

This section describes the performance level reached by using Independent Component Analysis in the denoising step. As described before, the ICA denoising algorithm used in this thesis is a modification of the Wavelet denoising algorithm, and was discussed in detail in Section 4.4.

The output of the ICA denoising step for one epoch can be seen in Figures 8.11 - 8.13. Unlike the WDA, with the ICA denoising algorithm, the wavelet decomposition is only used to *identify* segments that contain artifacts. The actual demixing with ICA is done to the original time-series (in magenta). When Figure 8.11 and 8.13 are compared, taking Figure 8.12 into account, it seems the phenomenon that is common to both Figure 8.11 and Figure 8.12 *is* suppressed in Figure 8.13. However, this is not necessarily desirable as some of those higher frequency phenomenon might have originated from the EEG source itself. Unfortunately, without any additional measurement channels, such uncertainty in the results do occur.

The EEG and the EOG data used in these figures are the same ones that were shown in Figure 8.5 and Figure 8.7.

The effectiveness of the IDA, with respect to automated classification is seen in Table 8.5. The performance level reached with the ICA implementation used in this thesis is poorer than with the WDA solution.

When prior information is not considered, ICA needs to have more measurement channels in order to ensure a reliable decomposition. Also, in this application, the number of channels observed was less than the number of macro sources in the human brain. In such scenarios, some phenomena could belong to one of many separated channels, and still satisfy the constraints.

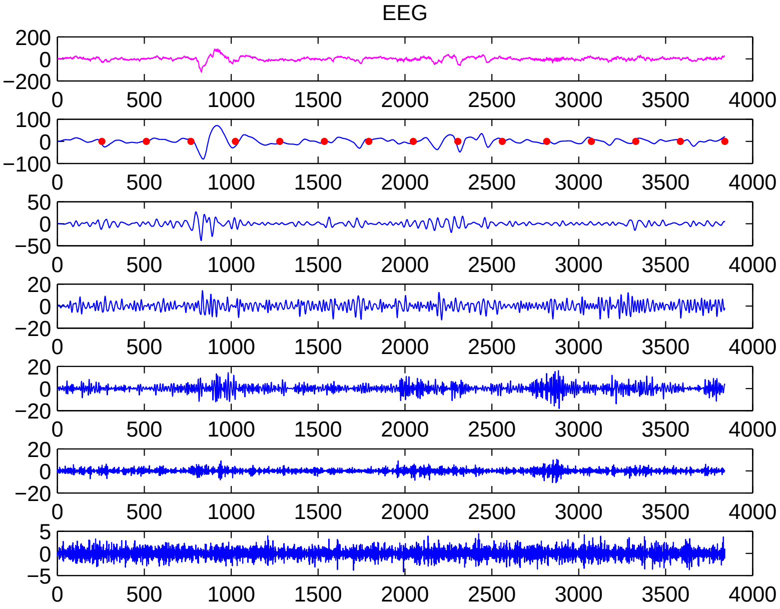

Figure 8.11: IDA denoising example - original EEG signal. From top to bottom [original EEG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

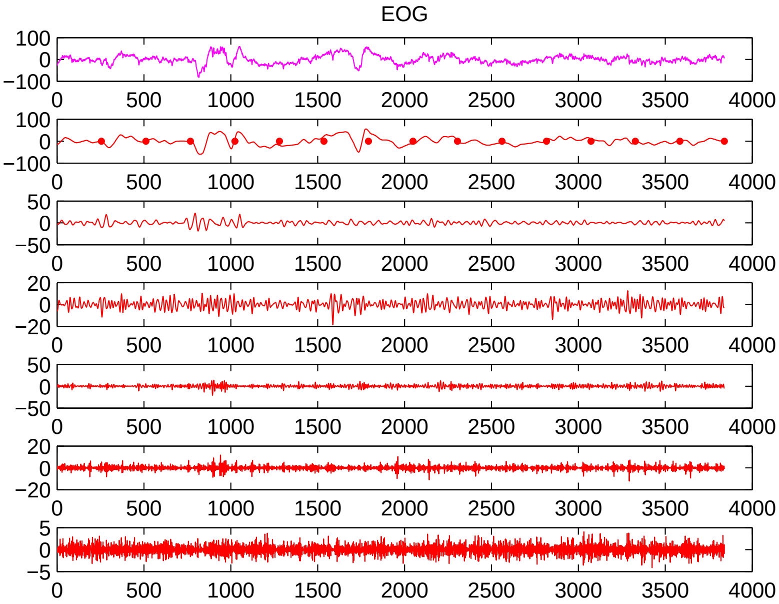

Figure 8.12: IDA denoising example - EOG signal. From top to bottom [EOG signal (magenta), reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

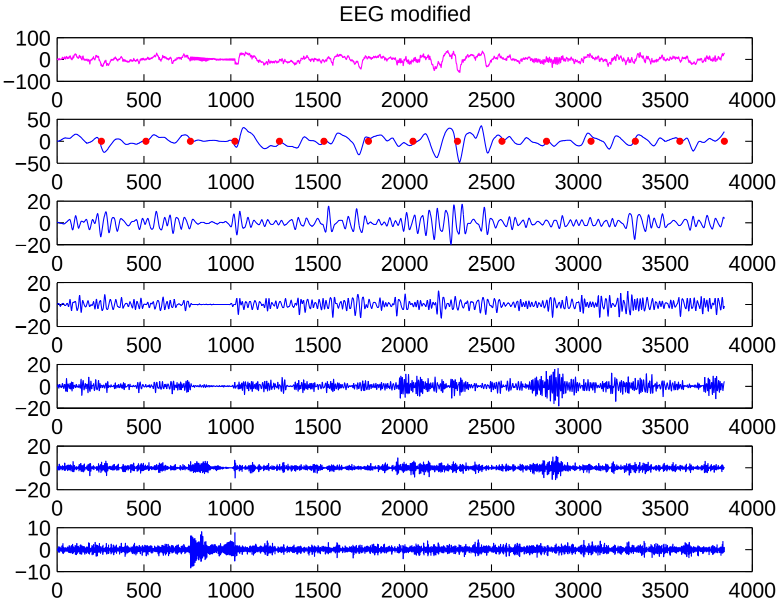

Figure 8.13: IDA denoising example - Denoised EEG signal. From top to bottom [Denoised EEG signal (magenta), [reconstructed signals from approximate coefficients and detailed levels 5 to 1, respectively.]

This experiment suggests that in order for ICA to be successful, prior information must be integrated into the system or more channels needs to be measured. This aspect of ICA, with respect to EEG, needs to be investigated in future work. Also, as discussed before in Section 4.4.7, during the selection of the modified EEG signal, instead of using the *summation* of the ranks, simply using the *highest* rank, when comparing ICA separated signals against the original artifact channels, *must* be considered.

Table 8.5: Classification results: feature selection, ICA denoising

	Accuracy						
	NREM I	NREM II	NREM III	NREM IV	REM	Awake	Total
F23	70.0%	93.5%	69.6%	85.5%	86.7%	85.0%	87.1%
F33	28.6%	95.3%	42.5%	99.4%	89.3%	81.7%	85.3%

8.6 Sleep/awake classification

For completion, the effectiveness of the algorithm at discriminating sleep vs awake stages was also measured. In order to do this, stages NREM I-IV and REM were grouped together into the *sleep* class, and was made distinct from the *awake* class. As mentioned before, the result of the experiment is an average over 10 iterations. The final performance level can be seen in Table 8.6.

Table 8.6: Classification results: Sleep/Awake

	Accuracy
Subject 1	97.1%
Subject 2	95.7%

As we can see, the average agreement with the human EEG expert is excellent. This demonstrates that the classifier that was selected is quite capable of discriminating between the *sleep* and *awake* stages. For this experiment, the same classifier described in Chapter 6, was used. In this case however, the classifier was single stand-alone classifier and was not a combination of multiple sub-classifiers. The reason for this was because, the discrimination was done between only two classes; namely, sleep and awake. However, the same classifier parameters listed in Chapter 6 were used.

Chapter 9

Conclusion

The purpose of this research was two-fold. Firstly, this research wanted to establish the importance of a theoretically sound feature selection step. Based on the results, the importance was clearly established. Many researchers select the features given to the classifier by trial and error. Instead, having a proper feature selection step will be more fundamentally sound, and will yield excellent results.

The MSDI algorithm and the MIC algorithm are not the most *optimal* by any means for the task of feature selection. The MSDI algorithm for example gives the same importance to Significant Difference and Feature Independence. The MIC algorithm does not consider a notion of *global relationships*. For example, two features that might compliment each other brilliantly, might not get selected since they were both eliminated in the absence of the other. However, even with these considerations being mentioned, the MSDI and MIC algorithms provide significant improvements in performance. The results obtained in this thesis clearly demonstrate that. Also, the work presented in this thesis had success with 5 features, compared to the 32 features selected in [1].

Secondly, this research wanted to investigate the effect of denoising the signals before classification. The benefits of this is two-fold. Firstly, if the artifacts present in the EEG signal are removed, the signal will become more useful to the EEG technicians as they don't have to worry about accommodating the presence of the artifacts. Secondly, the removal of the artifacts should make the automated classifier more accurate as well, since it now has a more cleaner signal to work with.

The parameters that were selected for the BWDA algorithm in this research did not provide consistent results. While successful denoising is certainly a possibility, selecting the proper thresholds is not as straightforward as with the WDA algorithm. The WDA algorithm was clearly successful in eliminating the ocular artifacts as was seen in the figures. Even though the performance improvement in automated classification was only marginal, it was better in both patients. The marginal improvement in performance could be due to the fact that the statistical features in the test were not significantly affected by the elimination of the various low frequency waveforms from the EEG channel. However, for the human EEG technician, the elimination of these artifacts can be more significant.

The IDA algorithm yielded poorer results than the WDA algorithm with respect to automated classification. After denoising with the IDA algorithm, the performance improvement over the non-denoised data was superior in one subject but was inferior in the other. This *relatively* unimpressive performance of the IDA algorithm, when compared to the WDA algorithm, could be explained by the fact that the number of measurement channels available are less than the number of sources that exist in the human brain. Also, the algorithm used to select the *denoised EEG* signal from the set of ICA demixed signals, described in Section 4.4.7, might need more improvement as well.

Chapter 10

Future Work

The research described in this thesis, did not consider any contextual information. All the epochs in this research were classified based on each epochs *own* features. As a result, the results should be even more encouraging in two important ways.

Firstly, the introduction of contextual information to the classification algorithm will undoubtedly improve the results even further. This step should be a post-processing step to the classified stages obtained from the system described in this thesis. When these context rules that the human experts use are explicitly written out, the performance should get better.

Secondly, it should be noted that these *contextually-classified* stages affect other epochs in the same sleep stage *negatively.* This is because, based on the *features themselves*, the epoch in question does not belong in that sleep stage. It was only classified to be of that sleep stage due to contextual information. When all these epochs are given to the automated classifier, including the contextually-classified epochs, the classifier needs to find a way, with respect to the available features, to integrate all of them to the same sleep stage. If these contextually-classified epochs were eliminated from the sleep stage, the classifier would be able to integrate the features better. After the epochs are classified based on their inherent features, they could be corrected as needed based on contextual information.

Also, Independent Component Analysis, in it's classical formulation, require more measurement channels than sources. When EEG demixing is attempted with ICA, many channels are usually available. Unfortunately, in this research only a very limited number of scalp electrodes were used. In [15], the *FastICA* algorithm was used throughout the whole book and they had success demixing a set of measurements from 15 channels. Given the same algorithm, the results obtained in this thesis looks to be of lesser quality. A solution to this would be to include *a priori* knowledge into the formulation, to compensate for the lack of channels. This aspect of ICA should be investigated in the future. Also, as discussed before in Section 4.4.7, during the selection of the modified EEG signal, instead of using the *summation* of the ranks, simply using the *highest* rank, when comparing ICA separated signals against the original artifact channels, *must* be considered. This might improve the result significantly.

It is also important to carry out the experiments with more subject data once they become available. For future work, data from an additional four subjects should be analyzed. During these future experiments, it might be also prudent to collapse NREM III and IV stages together and classify it as *slow wave sleep.* Since, EEG technicians don't necessarily pay as much attention to the distinction between NREM III and IV, this has the potential of improving the classification noticeably.

Appendix A

Wavelets

A wavelet can be thought of as a *little wave*, because it is short in duration, has finite energy, and integrates to zero. Due to it's unique characteristics, it is extremely suitable to represent transients. [18]

A.1 Motivation for its use

Given a signal, engineers can perform transformations on it in order to observe the frequency content of the signal. The most popular transform that was used in the past was the Fourier Transform. Unfortunately, the basis functions of the Fourier Transform, sines and cosines, are not localized in time. As a result, any frequency information calculated by the classical Fourier transform is a statistical average over the duration of the *whole* signal. If a *transient* exists in the signal, it's contribution to the Fourier transform will be small, and its location on the time-axis will be lost. Also, Fourier transforms are very poor at analyzing non-stationary signals. [18]

In many types of research, analysis of the transients of the signal are very important. Therefore, the use of the classical Fourier transform will yield undesirable results as its localization properties are quite poor. One solution to this has been the Windowed Fourier Transform or Short-Time Fourier Transform (STFT). With this approach, the signal is divided into multiple windows (segments) before applying the Fourier Transform to each segment. A narrow window would provide good time-resolution but would give poor frequency-resolution. On the other hand, a wide window would give good frequency-resolution but poor time-resolution. Typically, the window size is established *a priori* to analysis. Since it is fixed, a notion of a *combined* time-frequency resolution does not exist. Of course, a solution to this would be to have a sequence of windows of different widths. However, this solution becomes quite time consuming. [18]

The wavelet transform does not have such limitations as its base functions are local both in time and frequency. Due to wavelets ability to focus on short-time intervals and long-time intervals, it is inherently capable of discovering information about both high-frequency phenomena and low-frequency trends.

A.2 Description

Similar to sines and cosines in the Fourier transform, Wavelet Analysis uses a prototype function called the *mother wavelet*. This function has a mean of zero, fast decaying in an oscillatory fashion, has finite energy, and integrates to zero. The mathematical definition of the Continuous Wavelet Transform (CWT) of a given signal, $x(t)$, is seen in Equation A.1. In this equation, a is the scale factor, b is the translation factor, and $g(.)$ is the mother-wavelet function. [18]

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) g\left(\frac{t-b}{a}\right) dt \tag{A.1}$$

A wavelet coefficient, denoted by CWT(a,b), is a measure of how well the original signal, $x(t)$, and the mother-wavelet that has been scaled by a factor of a and translated by a factor of b, match with each other. So, in essence, the original one-dimensional signal has been mapped to a two-dimensional space across scale a and translation b. [18]

The mother wavelet can be thought of as a windowing function as well. A large scale factor allows wideband frequency components of the signal to be observed while a small scale factor allows narrow-band frequency components to be observed. There are many types of mother wavelets mentioned in literature and used in practice; such as, Haar, Daubechies, and Morlet. The shapes of some of these mother wavelets were generated from MATLAB and is shown in Figure A.1.

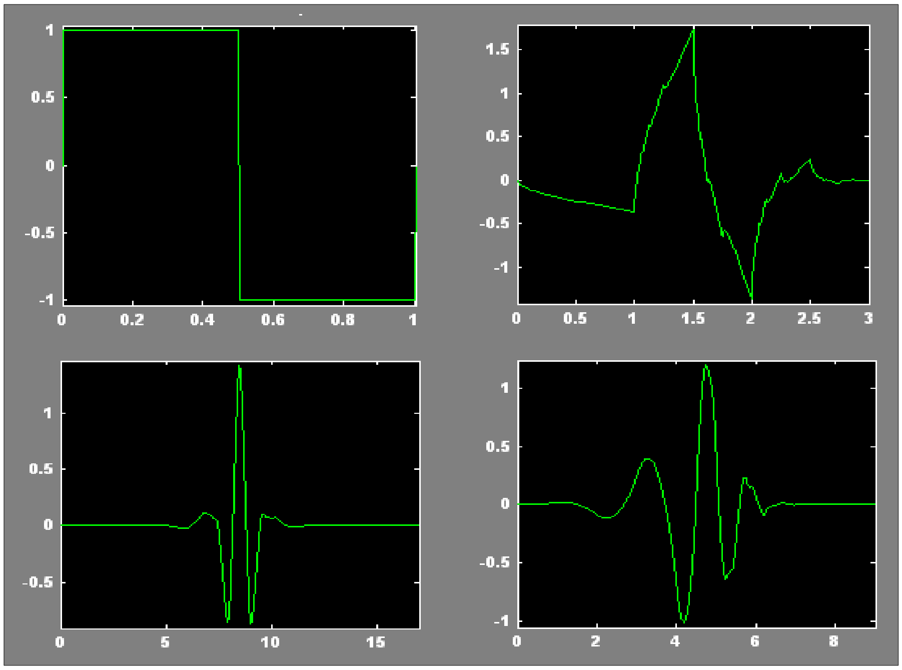

Figure A.1: Mother Wavelets

Unfortunately, the continuous wavelet transform cannot be implemented in a computer system. For this, the Discrete Wavelet Transform (DWT) must be used. The general equation for the discrete wavelet transform is given in Equation A.2. In this equation, the DWT is a function of the parameters m and k, and a_0 and b_0 are constants. A comparison of the resulting transforms generated by the CWT and the DWT is shown in Figure A.2. [18]

$$DWT(m,k) = \frac{1}{\sqrt{a_0}^{m}} \sum_{n} x(n) g\left(\frac{k - n b_0 a_0^m}{a_0^m}\right) \tag{A.2}$$

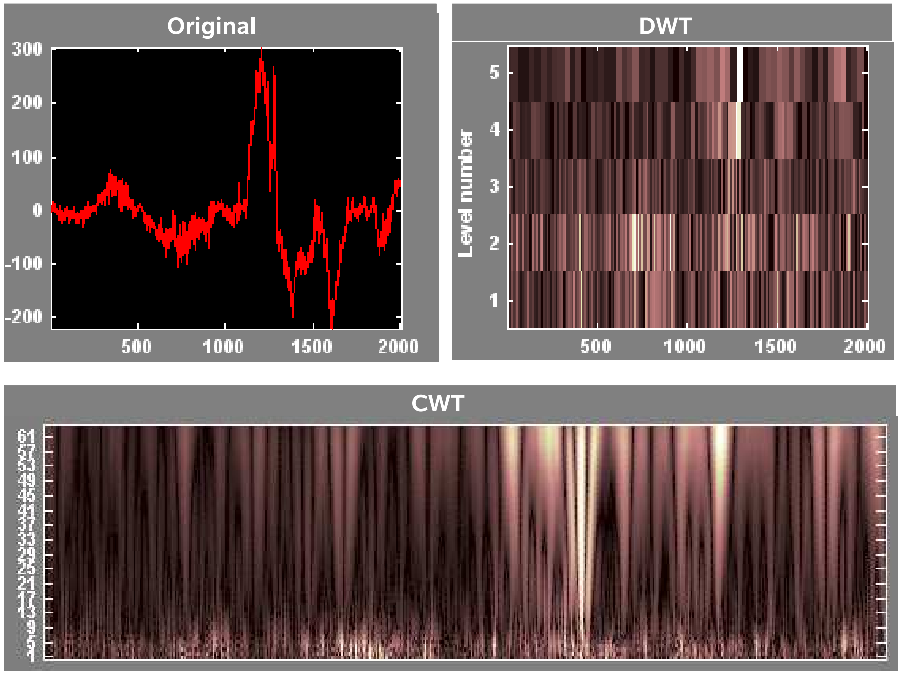

Figure A.2: CWT and DWT

A.3 Wavelet thresholding

Wavelet thresholding is a common technique which is used to eliminate noise from a signal. The *basic* algorithm for wavelet thresholding is shown in Table A.1. [8]

Table A.1: Wavelet thresholding

1	Decompose the signal and find the coefficients of the wavelet transform of the signal, S'.
2	Compare each wavelet coefficient against an appropriate threshold, and keeping only those coefficients larger than the threshold.
3	Applying the inverse wavelet transform to the result to obtain $\hat{S}$.

Obviously, this is not an *universal* algorithm, as the *noise* could be *larger* than the desired signal. If this is the case, in contrast to Step 2, coefficients *less* than a particular threshold should be *retained*. Also, it is not necessary to check *all* the coefficients separately. For example, the coefficient set can be divided into segments, and a statistical feature of each *segment* can be compared with a threshold.

So, in essence, the algorithm presented in Table A.1 is the thresholding performed for the most basic application. A researcher would have to modify it accordingly to suit his or her own research problem.

Appendix B

Independent Component Analysis

A fundamental reality in Signal Processing applications is the inability to take measurements *directly* from the most useful sources. The measurements that are taken are not necessarily *pure* source signals and are usually a mixture of the desired signals.

It is usually highly desirable to extract the independent source components, which are mixed to create the measurements, prior to any Feature Extraction and Classification stages. Discovering the fundamental independent components making up the measurements might provide more *readily available information* than the measured signals themselves.

Peoples' conversations in a large room is one such example. The sensors, instead of picking up the original voices, pick up the *mixtures* of the voices instead. When the mixed voice signals are received, they are usually *demixed* by the listener to understand the original dialog spoken. This is an example of the fundamental problem analyzed in this thesis.

The sources do not necessarily need to be physically separate components. It is quite possible for multiple independent signals to be generated from a single tangible source. In such a scenario, the independent source signals could be thought of as being generated from multiple *logical sources* than a single physical source. In this thesis, all references to a *source* should be thought of as a *logical source*.

B.1 Assumptions

For the purposes of discussion and analysis within this thesis, it will be assumed that the signals being generated from any independent sources are mixed *linearly* at the sensors. Since the separation problem is difficult enough as is, for the experiments within this thesis, it will be assumed that additive noise at the sensors is not present. As can be seen in System (B.1), the source signals s_i are mixed linearly to generate the measurement signals, x_i [15]. If the mixing matrix A is known, given X, obtaining a good estimation of the source signals should be straightforward. However, typically, knowledge of the mixing matrix is not present. Within this thesis, it is also assumed that the mixing matrix A is stationary in time and the mixing process is instantaneous. This can be seen clearly in the problem formulation defined in (B.1).

$$\begin{aligned}
x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1n}s_n \\
x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + \cdots + a_{2n}s_n \\
&\vdots \\
x_m(t) &= a_{m1}s_1(t) + a_{m2}s_2(t) + \cdots + a_{mn}s_n \\
\mathbf{X} &= \mathbf{As} \qquad \text{(B.1)}
\end{aligned}$$

It is also necessary to make an assumption about the variance of the source signals. The reason for this can be seen in (B.2).

$$x(t) = \sum_i \left(\frac{1}{\beta_i}\mathbf{a_i}\right)\left(\beta_i s_i(t)\right) \qquad \text{(B.2)}$$

Any scaling, β_i, performed on any of the sources s_i can be canceled out by dividing the corresponding column of the mixing matrix A by the same factor. As a result, within this thesis, it is assumed that all source signals are of unit variance. However, the ambiguity of the *sign* is still present [15]. However, this ambiguity will be ignored in this thesis.

B.2 Blind Source Separation

Knowing the mixing matrix A would allow the original source signals to be approximated quite easily. However, in reality, information about the mixing process is very limited. The Blind Source Separation (BSS) problem deals with the approximation of both the mixing matrix A and the estimation of the source signals, given only the measurements at the sensors, the nature of the mixing itself (e.g. linear), and perhaps the noise characteristics. Essentially, very little is known other than the measurements themselves; thus, the use of the term *blind* is quite appropriate. [15] [19]

Independent Component Analysis (ICA) is a widely used class of algorithms that is used to perform BSS. However, before delving into ICA, a less powerful method known as Principle Component Analysis (PCA) will be discussed. The need for ICA should become clear during the discussion of PCA.

Looking at the problem stated in (B.1), it looks rather difficult to solve at first glance. After all, assuming only the measurements are available, clearly there are more unknown variables than known variables. However, it turns out that the simple assumption that the original sources are independent of each other is sufficient to fill this void [15]. Just using this assumption of independence, along with the assumptions made with regards to the mixing process and other relevant technical assumptions discussed in the thesis, ICA is capable of estimating the independent source components. Once again, it should be noted that in this thesis, only problems that contain *linear* mixing at the sensors will be discussed.

Distinguishing between 2 *physically different* source components are only possible if they are independent of each other. ICA is limited in that sense. It is not possible to tell ICA to find the signal generated at a particular source. ICA will *implicitly* identify the sources that are independent to each other. So, if a single physical component (such as the heart), emits 2 different independent signals, they will be identified as 2 different independent components instead of as a single component. Thus, the notion of a *logical source* is quite appropriate.

B.3 Principle Component Analysis

For the remainder of the thesis, the measurements and the source signals will be considered to be a vector of random variables. This is done since the ideas and methodologies discussed in this thesis does not use the time structure of the signals. As a result, the time index t is dropped, and entities such as $\mathbf{x}$ (mixtures) and $\mathbf{s}$ (sources) are considered to be random vectors. Also, within this thesis, any *estimate* of the independent source components, by any method, will be known as a *Source Components Estimate* (SCE).

B.3.1 Introduction to PCA

One method that is used to reduce the redundancy in the measurement set and to increase the level of independence between the components, is Principle Component Analysis (PCA). PCA attempts to do this by uncorrelating the signals by using 2^{nd} order statistics. It should be understood that uncorrelating the mixtures is not as strong as making them be statistically independent from each other. Statistical independence is a much more richer and stringent concept than decorrelation. [15]

For 2 random variables x and y to be uncorrelated, (B.3) must be satisfied. [15]

$$
\begin{aligned}
E[xy] &= E[x]E[y] \\
&= 0, \text{ if zero mean}
\end{aligned}
\tag{B.3}
$$

However, in order for 2 signals, x and y, to be independent, they would have to satisfy (B.4), for *every* absolutely integrable functions g and h as well. Condition (B.3) can be derived from (B.4), by making g and h linear. Thus, (B.4) enforces much stricter constraints than (B.3). [15]

$$
E[g(x)h(y)] = E[g(x)]E[h(y)] \tag{B.4}
$$

In this thesis, the *PCA step* would result in a signal covariance matrix that is I (identity), and not simply a diagonal matrix. Therefore, within the discussion in this thesis, the traditional PCA procedure that simply diagonalizes the covariance matrix and the subsequent whitening procedure are combined into one step and will be known as the *PCA step.* The whitening step is important since, as described before, only source components of unit variance are being considered.

If X is the measurement matrix, the *PCA step* performs the transformation seen in equation (B.5). Here, D is the diagonal eigenvalue matrix and V is the eigenvector matrix of the sensor covariance matrix, estimated from the measured signals. The columns of V indicate the individual eigenvectors. [15]

$$
\hat{S} = D^{-\frac{1}{2}} V^T X \tag{B.5}
$$

B.3.2 Finding the correct rotation

As can be seen in Figure (B.1), uncorrelating and whitening the mixture is not sufficient. The obtained result is a *rotation* of the original source components. Even though the *PCA step* transformation is uncorrelated and whitened, it is obviously not independent; the knowledge of one variable tells a lot about the 2^{nd} variable.

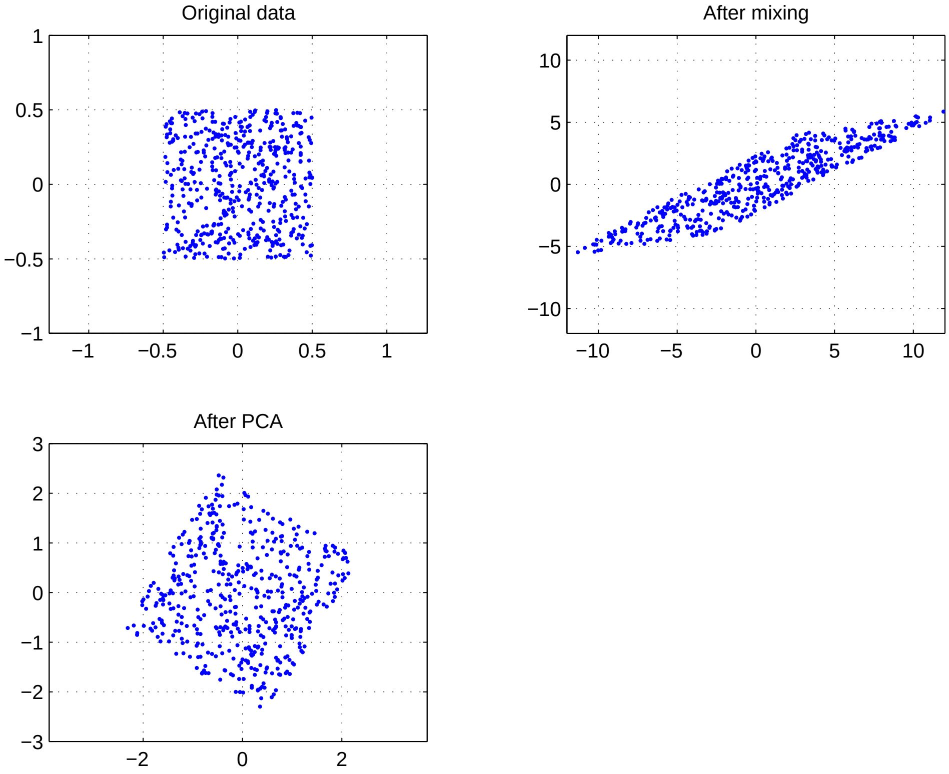

Figure B.1: Deficiency of PCA

In fact, there are many more solutions that will satisfy uncorrelatedness without satisfying independence. For example, consider Systems (B.6) and (B.7).

$$\begin{aligned} \hat{S} &= QX \\ \text{And, } C_{\hat{S}} &= QC_XQ^T = I \end{aligned} \qquad \text{(B.6)}$$

As can be seen in (B.6), the matrix Q found by the *PCA step* diagonalizes and whitens the covariance matrix of the estimate. Thus, the *PCA step* has achieved its main task.

However, if any orthogonal matrix U is applied to $\hat{S}$, as seen in Equation (B.7), the result would still be uncorrelated. [15]

$$\begin{aligned} \hat{S}' &= U\hat{S} \\ \text{So, } C_{\hat{S}'} &= UC_{\hat{S}}U^T = UIU^T = I \end{aligned} \tag{B.7}$$

Therefore, the original SCE found by the *PCA step* and all the other SCEs that were derived by applying an orthogonal transformation matrix, should have uncorrelated components. However, of all the possible solutions in this whitened space, the solution that is closest to the actual source signal set is the most desirable. The 2^{nd} order statistics cannot distinguish amongst the solutions in the whitened space; and thus, PCA is not capable of pinpointing the actual final solution. Even though PCA was instrumental in reducing the size of the solution space to the whitened solution set, higher order statistics must be used to make further distinctions. [15]

B.4 General idea behind higher order statistics

After the signal set has been decorrelated, other algorithms that take higher order statistics into consideration can be applied iteratively until a solution is obtained from within the whitened solution set. Usually, the algorithms that consider the higher-order statistics will attempt to find the *best* available result. Therefore, it is usually not a matter of finding *some* result; but a matter of the quality of the found result being poor. The relationship between the different subsets that have been discussed so far can be seen in Figure B.2.

A final point to discuss is whether the solution space that is eliminated due to the whitening process is detrimental to finding an appropriate solution. Due to the *assumption* that all source

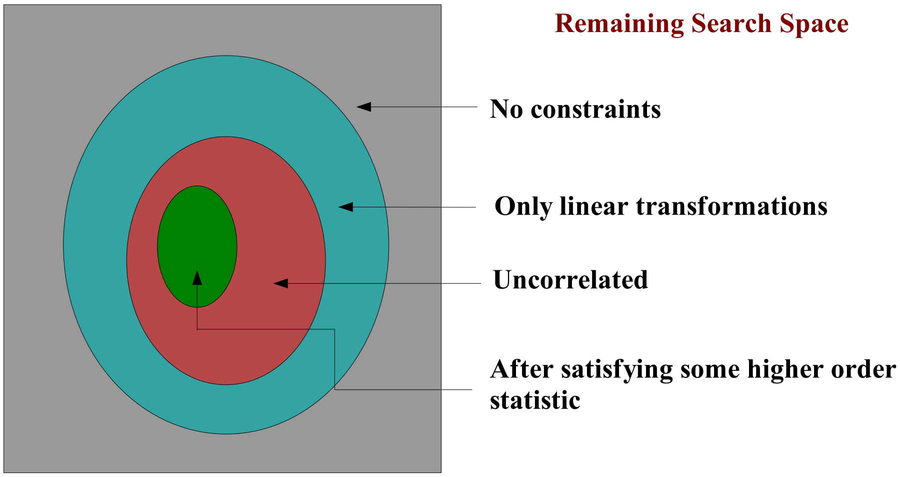
Remaining Search Space
No constraints
Only linear transformations
Uncorrelated
After satisfying some higher order statistic

Figure B.2: Relationship between the solution spaces after various transformations

components have a variance of 1, all the demixing matrices and the corresponding SCEs that would force the covariance to be $D \neq I$ will never be found as a solution. If the original source signals in fact *do* have a covariance matrix $D \neq I$, the exact signal will never be found due to the assumption made with regards to the variance of the source components. However, a *scaled version* of the source signal with a covariance of I should be available to be found by the proper application of discussed algorithms. This is considered to be sufficient in most applications and will be considered to be sufficient in this thesis as well. If the actual scaling factor is important, more information is required.

Since the original source components were independent, it must be ensured that only those SCEs whose member components satisfy (B.3) and (B.4) are considered. However, it is impossible to test all the higher-order statistics when finding a solution. Typically in practice, after the signal set is uncorrelated, a single algorithm that considers only a very limited amount of higher-order statistics is usually sufficient to arrive at a reasonable conclusion. Independent Component Analysis (ICA) is one such algorithm, and will be discussed in Section B.5. [15]

Equation (B.4) is not necessarily used directly to separate the mixtures; it was simply stated to demonstrate the importance of higher-order statistics with respect to independence. Even though the Taylor expansion of certain functions can have an infinitely many higher-order terms, the higher-order terms are usually insignificant due to the factorial in the denominator.

Even if PCA cannot find the final desired solution by itself, its importance cannot be stressed enough. From a large set of possible solutions, PCA allows the search to be reduced to a much smaller set, whose members contain components that are uncorrelated to each other. Due to this vast reduction in the solution space at very little computational cost, the *PCA step* is a powerful pre-processing step. [15]

Yet another important application of PCA occurs if the number of independent sources are less than the number of mixtures that are observed. Then, uncorrelating the components would cause some of the components to have insignificant information, indicated by their low variance values. This allows those components to be simply rejected in order to ease any subsequent analysis. [15]

It should also be noted that mixtures consisting of multiple Gaussian sources cannot be separated using the previously said methods. If the underlying sources were independent Gaussian random variables, any linear mixtures of them would also be Gaussian. Uncorrelating and whitening such a mixture would naturally produce a covariance matrix that is I. However, with Gaussian random variables, uncorrelatedness implies independence. This can be easily derived from the joint probability density function when the covariance matrix is diagonal. The implication of this is that uncorrelatedness *implicitly* implies independence and therefore *any* measure of independence, using higher-order statistics, would be satisfied by the current uncorrelated mixture itself. That is not to say that the solution obtained through PCA is the perfect solution. In fact, as before, any orthogonal transformation on the uncorrelated mixture would produce another uncorrelated mixture; and this mixture would be statistically independent as well.

It is impossible to distinguish between all these potential solutions. Thus, without knowledge of the mixing matrix itself, it is impossible to state which of them is the correct estimation. With another type of distribution, even if an orthogonal transformation maintains uncorrelatedness, the level of statistical independence will most likely be different, and can be detected with higher-order measures.

B.5 Independent Component Analysis

As the name implies, Independent Component Analysis assumes that the original source components are independent. This is an assumption that is absolutely needed to carry out the analysis, since the number of unknown variables in (B.1) is greater than the number of known variables. The identity of the sources are implicitly identified by ICA, such that the redundancy between them is minimized. The sources themselves are not known beforehand and cannot be assumed to be physically separate components. The assertion that the original sources are independent allows ICA to implicitly estimate the actual underlying logical sources.

As was seen in Section B.3, uncorrelating the measurements using 2^{nd} order statistics was not sufficient. Thus, ICA uses higher-order statistics and continues the search where PCA left off. As can be seen in (B.4), the use of 2^{nd} order statistics and higher-order statistics should provide a better measure of independence than using 2^{nd} order statistics alone. As the amount of higher order statistics used in the analysis increases, the accuracy should improve.

Realistically speaking, it is only possible to consider a limited amount of higher order information. Since, typically ICA does not restrict the user to a particular set of higher-order information, the user is free to select the appropriate functions as needed. [15]

ICA has 2 main components, the objective function and the optimization algorithm. Their connection is seen in (B.8). [15]

$$\text{ICA method} = \text{objective function} + \text{optimization algorithm} \qquad \text{(B.8)}$$

The objective function is a *measure of independence* using higher-order statistics. Obviously, in the strictest theoretical sense, it is not a *completely reliable* measure of independence. However, it has been shown in practice that a well-thought-out objective function, even if it only considers a limited amount of higher order information, can still give a sufficiently acceptable measure of independence in order to carry out the separation of the mixtures. [15]

Of course, having a measure of independence is not sufficient. An algorithm must exist to exploit this measure. Thus, the proper selection of the optimization algorithm is also crucial. The properties of the ICA method used is dependent on both the objective function and the optimization algorithm. The objective function determines statistical properties such as consistency, asymptotic variance, and robustness; and the algorithm determines properties such as the convergence speed, memory requirements, and numerical stability. [15]

Within this thesis, objective functions consisting of the Kurtosis and the Negentropy of the signal will be considered. For optimization purposes, the *FastICA* fixed-point algorithm will be used exclusively.

B.5.1 The role of Gaussianity

One of the implications of the Central Limit Theorem is that the linear combination of independent and identically distributed random variables will resemble the Gaussian distribution more so than the original random variables themselves. [15]

As a result, the distributions estimated from a *linear combination* of the source signals, should resemble the Gaussian distribution more so than the original source distributions themselves. Here, the source signals are assumed to be drawn from independent and identically distributed density functions. Thus, the level of *Gaussianity* in a signal can be used as a means of determining whether the signal is a mixture or an original source component. [15]

Now, all that is left is to determine a *quantitative measure* of Gaussianity. Due to the use of the *FastICA* algorithm, each component will be identified separately. As a result, all measures of Gaussianity is calculated separately for each *component* estimation.

B.5.2 Separation with Kurtosis

The *Kurtosis* of the signal can be used as a measure of Gaussianity. The Kurtosis of a random variable is the name given to its 4^{th} order cumulant, seen in (B.9). [15]

$$kurt(y) = E\big[y^4\big] - 3\Big(E\big(y^2\big)\Big)^2 \tag{B.9}$$

The Kurtosis for a Gaussian random variable is 0. Therefore, the Kurtosis of a signal that is generated from a Gaussian random variable, should tend to 0, as the number of data points increase. [15]

Since a linear mixture of independent and identically distributed random variables should be more Gaussian than the original source distributions, the maximization of the absolute value of the Kurtosis, as a function of the rotational vector, has the potential of finding the correct solution. [15]

The PCA step reduces the remainder of the search to the set of orthogonal matrices; an orthogonal rotation matrix is constructed from a set of rotational vectors. Therefore, using the absolute value of the Kurtosis as an objective function, the orthogonal matrix set can be explored to find the correct rotation matrix. [15]

B.5.3 Separation with Negentropy

One of the biggest problems associated with using the Kurtosis of a signal as a measure of Gaussianity is its sensitivity to outliers, when estimated from sampled values. Due to the 4^{th} power that is present in the calculation, even a moderately large outlier could skew the results significantly. For example, if 1000 data points are obtained from a channel with a data variance of 1, and one of them takes the value of 10, the Kurtosis will equal to at least 7 [15]. It would be beneficial to have a measure that not so sensitive to outliers but that still retains the speed of the Kurtosis method.

Based on this, a common measure of Gaussianity that can be made to be less sensitive to outliers is directly related to the *entropy* of the signal. The entropy essentially deals with the amount of randomness that is present in a signal; or alternatively, the lack of structure that is present in a signal. The entropy for discrete and continuous variables are seen in (B.10). [15]

$$\begin{aligned} \text{Discrete: } H(X) &= -\sum_i P(X = a_i)\, log\, P(X = a_i) \\ \text{Continuous: } H(X) &= -\int p_x(\epsilon)\, log\, p_x(\epsilon)\, d\epsilon \end{aligned} \tag{B.10}$$

It has been shown in literature that of all random variables of unit variance, a Gaussian random variable has the largest entropy. In fact, generally, the Gaussian distribution has the largest entropy of all distributions with a given covariance matrix. Thus, entropy can be used as a measure of Gaussianity if all of the signals under consideration have the same covariance. A larger entropy value implies that the signal under consideration is more Gaussian than the other signals and thus is probably a mixture. [15]

In order to use the entropy based measure of Gaussianity, it is important to establish that all signals that are analyzed by ICA have the same covariance. Otherwise, it is not appropriate to compare the signals with one another using the entropy measure. Now, due to the *PCA step*, we are only searching the *orthogonal matrix set* to obtain the final component of the demixing matrix; namely, the orthogonal rotation matrix. Before and after any such orthogonal transformation, all of the individual component estimates will continue to have a variance of 1. Therefore, estimating the level of Gaussianity between the estimates with the entropy measure is perfectly valid.

A larger entropy typically means that the signal under consideration is closer to a Gaussian distribution than a signal with a smaller entropy. Also, as discussed before, a higher measure of Gaussianity typically means that the signal under consideration is *more of a mixture* of source components than a signal with a lower measure of Gaussianity. Of course, this is with the assumption that the original source distributions are independent and identically distributed. The distributions being *identically distributed* is usually not the case in real world applications. Even though it really depends on the actual distributions, the experiments in the thesis show that it is possible to achieve good separation even if the distributions are not identical.

Now, to make the measure of Gaussianity be zero for a Gaussian variable and be nonnegative in general, an alternative measure called *Negentropy* can be defined. Negentropy is directly related to entropy. However, Negentropy is always a positive quantity and the negentropy of a gaussian random variable is 0. In the Negentropy equation defined in (B.11), $\mathbf{x_{gaussian}}$ is a gaussian random variable with the same variance as the random variable $\mathbf{x}$. Also, an important property of negentropy is that it is scale invariant. This can be seen in equation (B.12). [15]

$$J(\mathbf{x}) = H(\mathbf{x_{gaussian}}) - H(\mathbf{x}) \tag{B.11}$$

$$J(c\mathbf{x}) = J(\mathbf{x}), \text{ for a constant scalar } c \tag{B.12}$$

Since the measurements available are simply sampled values, some form of density estimation needs to be applied in order to estimate the Negentropy. It is important to note that calculating the Negentropy based on its definition, for an arbitrary density function, could be quite computationally expensive due to the presence of the integral [15].

Various algorithms that attempt to approximate the Negentropy are available due to this high computational cost. [15] These measures are by no means perfect; but as long as the measures are accurate in a *relative* sense, the approximation can be considered to be quite good. For example, if solution 1 has a true Negentropy measure that is higher than that of solution 2, it should be the case that the *approximation* of Negentropy should also provide a similar relationship, even if the absolute values are incorrect.

B.5.4 Approximating Negentropy

All the methods that approximate the Negentropy makes various assumptions. The goal is to approximate the Negentropy as accurately as possible, but at a reasonable computational cost. The high computational cost of the original definition is one of the motivational factors for the need for a suitable approximation. [15]

Cumulant based approximation

This method makes the fundamental assumption that each signal in question has a distribution *very close* to that of a Gaussian distribution. The derivation approximates the probability density function (pdf) of the signal with a combination of the *standard gaussian pdf* and *higher-order cumulants* that approximate the degree to which the actual pdf is *different* from the gaussian pdf (B.13). [15]

$$p_x(\epsilon) \approx \hat{p}_x(\epsilon) = \varphi(\epsilon)\left(1 + \kappa_3(x)\frac{H_3(\epsilon)}{3!} + \kappa_4(x)\frac{H_4(\epsilon)}{4!}\right) \tag{B.13}$$

Substituting this into the definition of negentropy, and assuming that the cumulants in (B.13) are very small, an approximation for the Negentropy can be obtained and is seen in (B.14).

[15]

$$J(x) \approx \frac{1}{12}E[x^3]^2 + \frac{1}{48}kurt(x)^2 \tag{B.14}$$

However, if the Negentropy method is used with the approximation, the same issues that arose with the use of the Kurtosis as a measure of Gaussianity is encountered. As with that method, this approximation is very sensitive to outliers. [15] Also, the assumption made with regards to the pdf being *close* to the gaussian pdf is also important. If this assumption is false, obviously, the approximation will not be very good. However, the approximation should improve as the distribution approaches the gaussian distribution. So this assumption is quite appropriate when dealing with signals that are heavily mixed. However, as the mixture starts to become less mixed, and thus less Gaussian, one has to be careful with making this type of assumption. Since this approximation suffers from the same problems as the Kurtosis measure, it is not quite useful.

Maximum Entropy based approximation

Given a set of samples from some distribution, it is impossible to estimate the original distribution since there are an infinitely many distributions that will satisfy the constraints implied by the sampled points; and most of these distributions will have different entropy values from each other. [15] As described previously, an estimation of the density function is needed for the calculation of Negentropy.

The *Maximum Entropy method* is interested in the distribution with the *maximum entropy* that satisfies the constraints implied by the data points of the potentially transformed signal. The entropy of the *actual* distribution will be something less than the *maximum entropy* quantity. The assumption is that minimizing the maximum value of the entropy, that is consistent with the data points, will also hopefully minimize the actual entropy as well. [15] While, this is not true in general, it should give decent results for the most part. Since the cost function is usually in terms of Negentropy, the optimization problems essentially deals with the *maximization* of the Negentropy value, as discussed below.

Similar to Section B.5.4, this derivation makes the assumption that the maximum entropy density that is consistent with the points in the transformed signal in question is not far from the gaussian density of the same mean and variance. [15] Obviously, the invalidity of this assumption leads to the issues raised previously.

The resulting approximation is shown in (B.15) [15].

$$J(x) \approx \frac{1}{2} \sum_{i=1}^{n} E[F^i(x)]^2 \tag{B.15}$$

The F^i functions seen in (B.15) form an orthonormal system as defined in (B.16). Here, $\varphi(\epsilon)$ is the *standard gaussian distribution.* For a complete derivation, please see [15].

$$\begin{aligned} \int \varphi(\epsilon) F^i(\epsilon) F^j(\epsilon)\, d\epsilon &= \begin{cases} 1 & \text{if i = j} \\ 0 & \text{if i} \neq \text{j} \end{cases} \\ \int \varphi(\epsilon) F^i(\epsilon) \epsilon^k\, d\epsilon &= 0, \text{ for k=0,1,2} \end{aligned} \tag{B.16}$$

It is not easy to simply select the F^i functions, and have them satisfy (B.16). Therefore, usually a set of *linearly independent* functions, G^i, are selected and the Gram-Schmidt orthogonalization scheme is performed on the set containing the G^i functions and $\epsilon^i, k = 0, 1, 2$. The resulting set of functions, F^i, will satisfy the orthonormality requirements stated in (B.16). [15] The procedure for doing this orthonormalization is found in [20].

When the distribution of x is Gaussian, the Negentropy specified in (B.15) will evaluate to 0. This is because it can be shown that $E[F^i(x)] = 0, \forall i$ when the random variable is Gaussian. This is easily seen by making $k = 0$ in the constraints specified in (B.16). [15] Therefore, as it should be, the negentropy approximation of a gaussian variable is 0.

Without going through the full derivation, the reason for the ability to use any *arbitrary* set of functions that conforms to the orthonormality requirements might not be conceptually clear. What is important to realize is that the F^i functions are essentially meant to measure *how far* a distribution is away from the Gaussian distribution. Since the actual distribution is not available, $E[F^i(x)]$ will be estimated from the data points themselves. The effectiveness of the F^i functions chosen will certainly depend on the underlying distribution that might have generated the signal in question, and therefore must be selected carefully.

For a nongaussian distribution, $E[F^i(x)] = c_i, \forall i$. These c_i values are not necessarily zero, and the sum of their squares give an approximate Negentropy value that indicates how far the distribution is from the gaussian distribution. A higher approximation of Negentropy should indicate a distribution further away from a Gaussian distribution.

Since the researcher selects the G^i functions, it is his or her responsibility to select them in an intelligent manner. In practice, G^i functions are selected to measure a characteristic of the underlying distributions of the signals, that might indicate a meaningful measure of distance to the gaussian distribution. For example, *odd* and *even* functions are popular choices as they measure skewness and peakiness, respectively [15]. With these functions, a nongaussian random variables should usually give a nonzero value for the approximated Negentropy and should assist the researcher in determining how far the distribution is from the Gaussian distribution. The goal is to increase the distance as much as possible, since the distributions of *mixtures* of source components resemble the Gaussian distribution more so than the distributions of the original source components. Also, it is important to select the G^i functions to be robust with respect to outliers.

Selecting $G^1(y) = y^3$ and $G^2(y) = y^4$ will result in the cumulant based approximation seen in (B.14). However, for reasons already discussed, this approximation is considered to be poor. Also, it is not required to have two G^i functions. Selecting only one such function can still give superior results to that of (B.14). The resulting simplified equation is seen in (B.17). Here, v is a zero mean, unit variance Gaussian random variable. Random variable y is also zero mean and of unit variance, but is not necessarily Gaussian. [15]

$$J(y) \propto (E[G(y)] - E[G(v)])^2 \tag{B.17}$$

A popular choice for $G^1 = G$ in the single function approximation is seen in (B.18). [15] The approximation based on this function is used in the experiments in this thesis.

$$G(y) = -exp(-y^2/2) \tag{B.18}$$

B.6 FastICA

Of course, having an objective function is only part of the problem. Given an objective function, it is important to investigate the possible solutions in order to arrive at a suitable solution. As expected, *Gradient* methods have been fairly popular for this purpose. [15]

If the goal is to maximize a function, $C(\mathbf{w^T z})$, as a function of $\mathbf{w}$, the Gradient ascent method in (B.19) can be used. [15] Here, $\mathbf{w}$ represents the ICA rotation vector that is applied to the vector of whitened random variables, $\mathbf{z}$, to obtain an estimate for *one* of the source signals. Once all n of the rotation vectors, $\mathbf{w_i}$, are obtained separately, a suitable orthogonalization scheme can be applied to orthogonalize them. These schemes will be discussed in Section B.7. Once again, it is important to note that due to the unavailability of any actual density functions, any expectations will be evaluated with the data points.

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} + \alpha \frac{\partial C(\mathbf{w^T z})}{\partial \mathbf{w}} \\ \mathbf{w} &\leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|} \end{aligned} \tag{B.19}$$

Gradient methods depend greatly on the selection of the *learning rate.* An improper learning rate can either make convergence extremely slow or destroy convergence completely. [15] Therefore, the authors of [15] have proposed a method known as *FastICA*, which is independent of such learning rates. FastICA actually has its roots in gradient based algorithms.

With the use of the Lagrangian method, it can be shown that at the maximum point of the function $C(\mathbf{w^T z})$, when $\mathbf{w}$ is constrained to be on the unit sphere, the gradient of $C(\mathbf{w^T z})$ is pointing in the same direction as $\mathbf{w}$. Therefore, if $\mathbf{w_f}$ is the solution to the maximization problem, constrained by $\|\mathbf{w_f}\| = 1$, the constraint shown in (B.20) is true. [15]

$$\frac{\partial C(\mathbf{w^T z})}{\partial \mathbf{w}}\Big|_{w=w_f} = \alpha \mathbf{w_f}, \text{ where } \alpha \text{ is some constant} \tag{B.20}$$

Generally, at a *local maximum or minimum* of a typical function, the gradient is 0. However, in the optimization problem we are interested in, the gradient at the maximum point, $\mathbf{w_f}$, when it is constrained to be on the unit sphere, is said to be a scalar multiple of $\mathbf{w_f}$. The reason for the gradient *not* being zero at $\mathbf{w_f}$ should be conceptually clear. None of the *local maximum or minimum points* of the $C(\mathbf{w^T z})$ function are necessarily at $\mathbf{w_f}$. $\mathbf{w_f}$ is only the maximum point when the solution space is *constrained* to the unit sphere ($\|\mathbf{w}\| = 1$).

For example, let the Kurtosis of the signal be the cost function. One of the properties of the kurtosis measure is that $|kurt(\alpha x)| = \alpha^4 |kurt(x)|$. Unless the underlying distribution of $\mathbf{w^T z}$ is mesokurtic, $|kurt([\alpha \mathbf{w^T}]\mathbf{z})|$ should always give a higher value than $|kurt(\mathbf{w^T z})|$, whenever $|\alpha| > 1$. Therefore, none of the local maximum values of $C(\mathbf{w^T z})$ will be at $\mathbf{w_f}$ or anywhere on the unit sphere, as any vector $\alpha \mathbf{w_f}$ when $\alpha > 1$, will give a higher kurtosis value.

While the condition in (B.20) must be true at a maximum, the converse is not necessarily implied by the statement. However, it is still used as the basis for the algorithm. Considering the *Gradient Ascent algorithm* would itself identify such a point as a solution, since any point satisfying (B.20) would force the Gradient solution to stabilize, it seems acceptable to use it as a basis for the *FastICA algorithm* as well [15].

Thus, satisfying (B.20) is assumed to imply the presence of a maximum at the point in question. This leads to the equation shown in (B.21) that can be solved to obtain a potential solution. [15]

$$\begin{aligned} \alpha \mathbf{w} &= \frac{\partial C(\mathbf{w^T z})}{\partial \mathbf{w}}, \text{ Alternatively} \\ \mathbf{w} &= \beta \frac{\partial C(\mathbf{w^T z})}{\partial \mathbf{w}} \end{aligned} \tag{B.21}$$

The FastICA algorithm based on (B.21) is seen in B.22. The algorithm is run indefinitely until the value of 2 consecutive **w** vectors converges, maybe except for the multiplicative sign.

$$\begin{aligned} \mathbf{w} &\leftarrow \frac{\partial C(\mathbf{w^T z})}{\partial \mathbf{w}} \\ \mathbf{w} &\leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|} \end{aligned} \tag{B.22}$$

This is essentially a *Fixed-Point* algorithm followed by a normalizing step. Looking at equation (B.22), it is clear that the scalar constant β was dropped from (B.21) by the designers of the algorithm. The motivation for dropping the scalar constant is based on the fact that the **w** vector is normalized at the end; thus, the effect of any scalar constant would vanish [15].

However, it should be noted that this approach cannot possibly be taken in the general case for an arbitrary equation $w = \beta H(w)$. Obviously, when we solve such an equation using a fixed-point algorithm, we are essentially trying to find the intersection between $y = w$ and $y = \beta H(w)$ [21]. The intersection points in this system, if any, clearly depends on β. With the FastICA algorithm, β is assumed to be 1. The solution found by the fixed-point algorithm, when $\beta = 1$, is clearly by definition, a scalar multiple of the gradient at that point; the word *point* is used since the only *variable* in the optimization problem is **w**. However, when

$\mathbf{w} = \mathbf{w}_{unnormalized}$ is normalized, $\mathbf{w_{norm}}$ is obtained. In general, there is no guarantee that the gradient at $\mathbf{w_{norm}}$ has the same direction as the gradient at $\mathbf{w}_{unnormalized}$. Therefore, in general, finding the solution by assuming $\beta = 1$ and then normalizing does not necessarily guarantee a proper solution.

However, when the cost function is either the *Kurtosis* or the *Negentropy*, it can be shown that the *gradient* when $\mathbf{w} = \mathbf{x}$ and the *gradient* at a scalar multiple of $\mathbf{x}$, $c\mathbf{x}$, only differ by a scalar constant dependent on c. Thus, with the cost functions discussed in this thesis, the *gradient* at $\mathbf{w} = \mathbf{w^T_{norm}}$ has the same direction as the *gradient* at $\mathbf{w} = \mathbf{w^T_{unnormalized}}$, which obviously has the same direction as $\mathbf{w_{unnormalized}}$, and which in turn has the same direction as $\mathbf{w_{norm}}$. Therefore, the *gradient* at $\mathbf{w} = \mathbf{w^T_{norm}}$ has the same direction as $\mathbf{w_{norm}}$. Therefore, $\mathbf{w_{norm}}$ satisfies (B.20) and is also on the unit sphere; and thus, satisfies all the criteria of the search.

The advantage of the FastICA algorithm is its independence of the learning rate. With the cost functions discussed in this thesis, the FastICA algorithm can find all the solutions that can be *potentially* found by the Gradient Ascent Algorithm, but is not dependent on the learning rate.

B.6.1 FastICA and the Kurtosis measure

When the cost function $C(\mathbf{w^T z})$ is made to be the Kurtosis of the signal, $K(\mathbf{w^T z})$, the ICA step is reduced to the algorithm shown in (B.23): [15]

$$\begin{aligned} \mathbf{w} &\leftarrow E\big[\mathbf{z}(\mathbf{w^T z})^3\big] - 3\mathbf{w} \\ \mathbf{w} &\leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|} \end{aligned} \tag{B.23}$$

B.6.2 FastICA and the Negentropy measure

When the cost function $C(\mathbf{w^T z})$ is made to be the Negentropy of the signal, $J(\mathbf{w^T z})$, and after some optimization, the ICA step is reduced to the algorithm shown in (B.24): [15]

$$\begin{aligned}\mathbf{w} &\leftarrow E\big[\mathbf{z}g(\mathbf{w^T z})\big] - E\big[g^{'}(\mathbf{w^T z})\big]\mathbf{w} \\ \mathbf{w} &\leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}\end{aligned} \tag{B.24}$$

The g function and the g' function obviously depends on the G function selected, which was described in Section B.5.4. For the experiments in this thesis, the functions seen in (B.25) are used. They were obtained from [15].

$$\begin{aligned}G(y) &= -exp(-y^2/2) \\ g(y) &= y\, exp(-y^2/2) \\ g'(y) &= (1-y^2)\, exp(-y^2/2)\end{aligned} \tag{B.25}$$

B.7 Orthogonalization of vectors

The FastICA algorithm described in Section B.6 essentially estimates *one* component at a time, as only one of the $\mathbf{w_k}$ rotation vectors is estimated. In order to obtain estimates for all the signals, the remaining $n-1$ columns of the rotation matrix W needs to be discovered as well.

It was shown before that to maintain uncorrelation between the components, the rotation matrix W must be orthogonal. Now, it is also the case that the orthogonality of W is only true if and only if the column vectors of W are orthogonal to each other. [22] Thus, ensuring that all the rotation vectors $\mathbf{w_k}$, which are the columns of W, are orthogonal to each other is sufficient. This requirement can be seen even more clearly in (B.26). [15]

$$\begin{aligned}
& E[(\mathbf{w_i^T z})(\mathbf{w_j^T z})^T)] \\
&= E[\mathbf{w_i^T z z^T w_j}] \\
&= \mathbf{w_i^T} E[\mathbf{z z^T}]\mathbf{w_j} \\
&= \mathbf{w_i^T w_j} \\
\text{Also, } E[(\mathbf{w_i^T z})(\mathbf{w_j^T z})^T)] &= E[(\mathbf{s_i s_j^T}] = 0,\, i \neq j \\
\text{Therefore, } \mathbf{w_i^T w_j} &= 0,\, i \neq j \qquad \text{(B.26)}
\end{aligned}$$

Even if all the rotational vectors $\mathbf{w_k}$ estimate one of the source components exactly, there might still be redundancy. It is quite possible that some of the rotational vectors estimate the *same* source component. If an estimation were to be duplicated, that essentially means that some of the other components were not even found. This is obviously highly undesirable. This situation would never happen if the $\mathbf{w_k}$ vectors were orthogonal to each other.

In this thesis, on each full iteration, all the separate $\mathbf{w_k}$ vectors are found separately by FastICA, and are then orthogonalized using a suitable orthogonalization scheme. Subsequently, the new vectors can simply be normalized. The most popular scheme for orthogonalization is the *Gram-Schmidt Orthogonalization* scheme, seen in (B.27). However, this method gives more emphasis to estimates with a lower index number and might cause the compounding of errors. Alternatively, The *Symmetric Orthogonalization* method can be used. This method does not give preference to any particular vector and as such does not have the problem associated with the compounding of errors. [15] Both of these methods are used in the experiments in this thesis.

$$\begin{aligned}\mathbf{w_1} &= \mathbf{a_1} \\ \mathbf{w_j} &= \mathbf{a_j} - \sum_{i=1}^{j-1} \frac{\mathbf{w_i^T a_j}}{\mathbf{w_i^T w_i}} \mathbf{w_i} \end{aligned} \tag{B.27}$$

The Symmetric Orthogonalization algorithm makes the unorthogonalized vectors orthogonal to each other while preserving their *likeness* to the original set of unorthogonalized vectors, as measured by an appropriate matrix norm. If the columns of matrix G are the $\mathbf{w_k}$ vectors, the Symmetric Orthogonalization algorithm is simply (B.28): [15]

$$G \leftarrow (GG^T)^{-1/2} G \tag{B.28}$$

The columns of the resulting matrix will give the new orthogonalized vectors. [15]

B.8 Complete ICA algorithm

Now that all the steps required have been discussed, the complete algorithm is summarized in Table B.1

Table B.1: ICA algorithm

#	**Step**
1	Center the data to obtain a mean of zero
2	Use PCA to diagonalize the covariance matrix
3	Reduce the number of components by eliminating any components with an insignificant variance
4	Randomly select n initial vectors for $\mathbf{w_i}$, i ← 1 to n
5	Update *all* $\mathbf{w_i}$ in parallel using FastICA with a suitable objective function
6	Perform Orthogonalization and normalize
7	Check for convergence of $\mathbf{w_i}$, if not converged goto step 5

B.9 Other possible measures

ICA is certainly not limited to the techniques described in this thesis. There are a multitude of other ICA solutions; such as *Maximum Likelihood, Tensorial Methods, and Nonlinear Decorrelation* based methods. The pre-processing steps are not limited to PCA either. Reference [15] gives further information.

Appendix C

Statistical Tests

C.1 Spearman's Rank Correlation

The Spearman's Rank correlation is a measure of correlation between two variables. One of the desirable properties of the Spearman's Rank correlation is that it does not make any assumptions regarding the distribution of data. The measure is obtained by considering the *rank* of the data item and considering the *difference* in rank of the corresponding values of two random variables.

For example, if the age and the height of an individual are correlated, a person should expect that her age rank in a group of individuals to be more or less equal to her height rank in the same group. Obviously, this is not necessarily the case in real life as there are many other factors, such as genetics.

The Spearman's rank correlation between two random variables, A and B, has a simple form and is shown in Equation C.1.

$$\rho_{A,B} = 1 - \frac{6\sum_{i=1}^{n}(r_{A,i} - r_{B,i})^2}{n(n^2-1)} \tag{C.1}$$

Here, $r_{A,j}$ and $r_{B,j}$ are the ranks of the j^{th} value produced by the random variable A and B, respectively. The values are always ranked with respect to all the values generated by the same random variable. Therefore, the equation compares the difference in rank between an ordered pair of values produced by two random variables, A and B, and judges the manner in which their ranks vary.

C.2 Mann-Whitney test

The essence of the Mann-Whitney test is quite simple. Given two unpaired groups, the goal is to determine whether the two groups are sufficiently statistically different. Given two labeled classes, it is important to know whether *all* the points came from a single distribution or whether they came from two different distributions. The Mann-Whitney test is a measure of this degree of separation. Also, it is a rank based measure.

The Mann-Whitney test measures how far the *actual* rankings differ from the *expected* rankings when both groups come from the same distribution.

Firstly, items are ranked from 1 to N, where N is the number of items in both groups. Then, all the data items of the class with the lower number of items is added and then compared to the *expected* sum of the rankings for this class. If there are n_s items in the class with less elements, the expected sum is simply $n_s(1+N)/2$. The actual form of the equation was obtained from [16] and is shown in Equation C.2.

$$\begin{aligned} Z &= \frac{\left| R_s - E\big(R_s\big) \right| - 0.5}{\sqrt{Var\big(R_s\big)}} \\ \text{where } E\big(R_s\big) &= \frac{n_s(1+N)}{2} \\ \text{and } Var\big(R_s\big) &= \frac{n_s n_m}{12}(1+N) \end{aligned} \tag{C.2}$$

Here, Z is the level of significant difference, R_s is the sum of the ranks of the elements in the class with less elements, n_m is the number of elements in the class with more elements, and $N = n_s + n_m$ is the total number of elements.

Appendix D

Figures of sleep stages

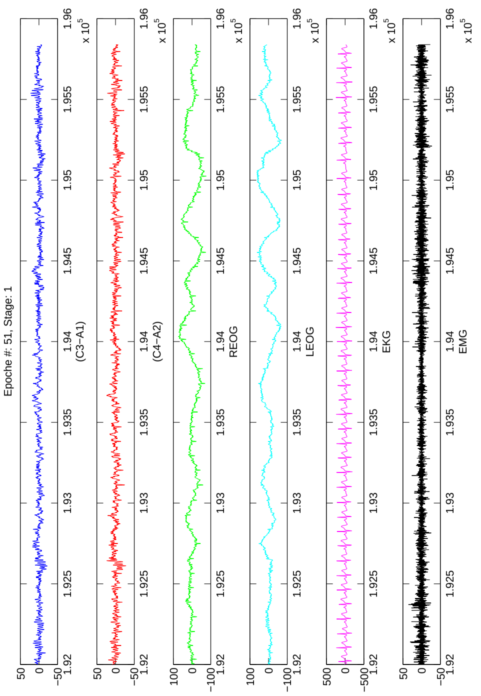

Figure D.1: Observed channels from sleep stage - NREM I

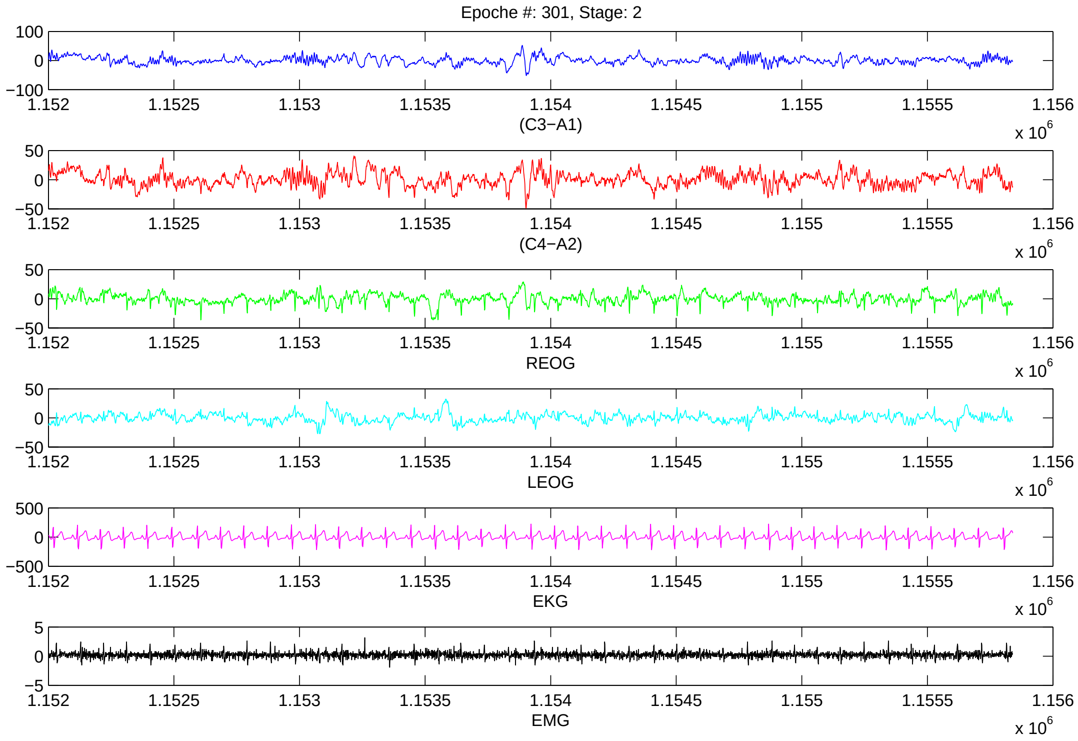

Figure D.2: Observed channels from sleep stage - NREM II

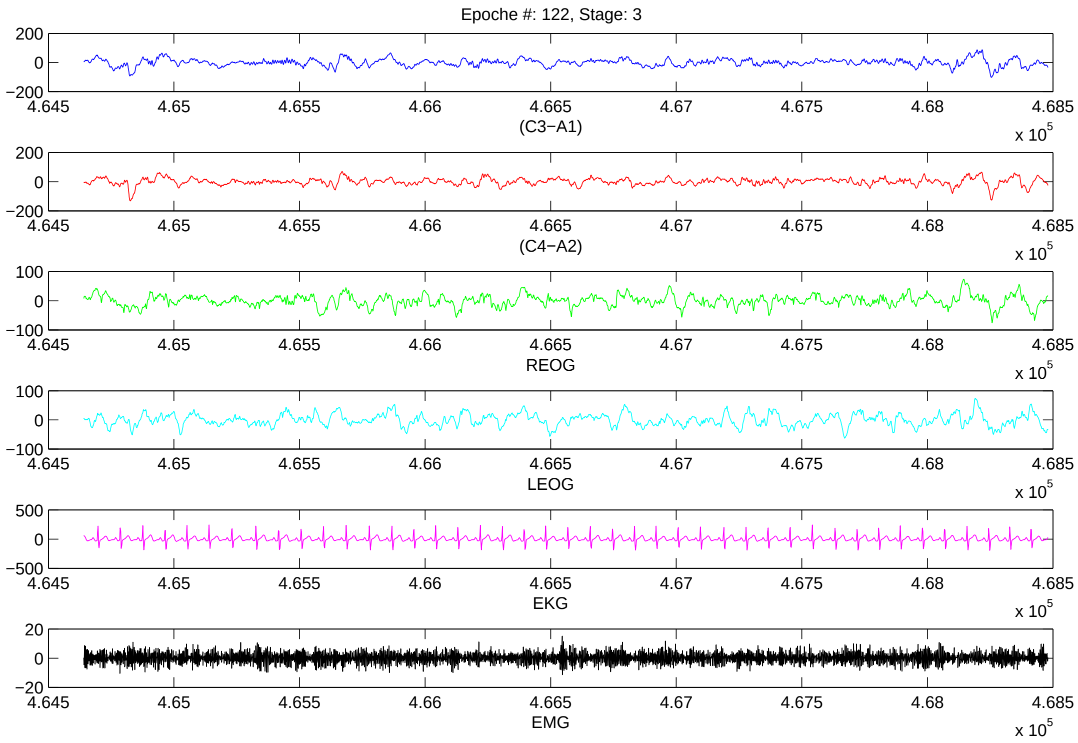

Figure D.3: Observed channels from sleep stage - NREM III

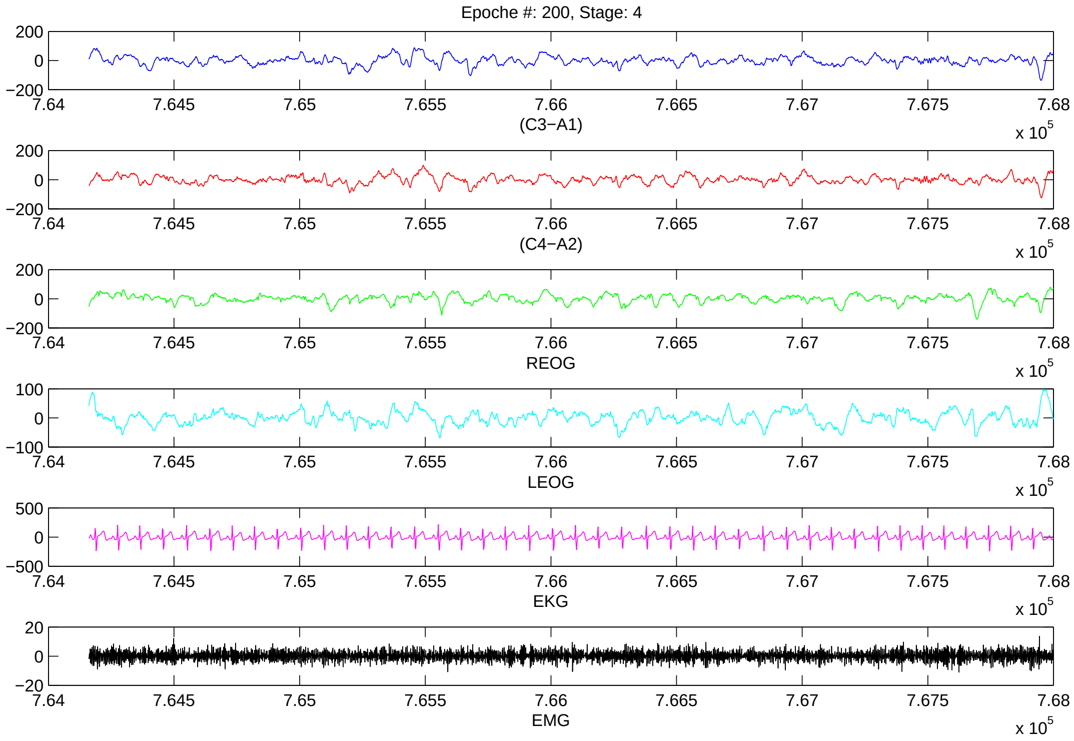

Figure D.4: Observed channels from sleep stage - NREM IV

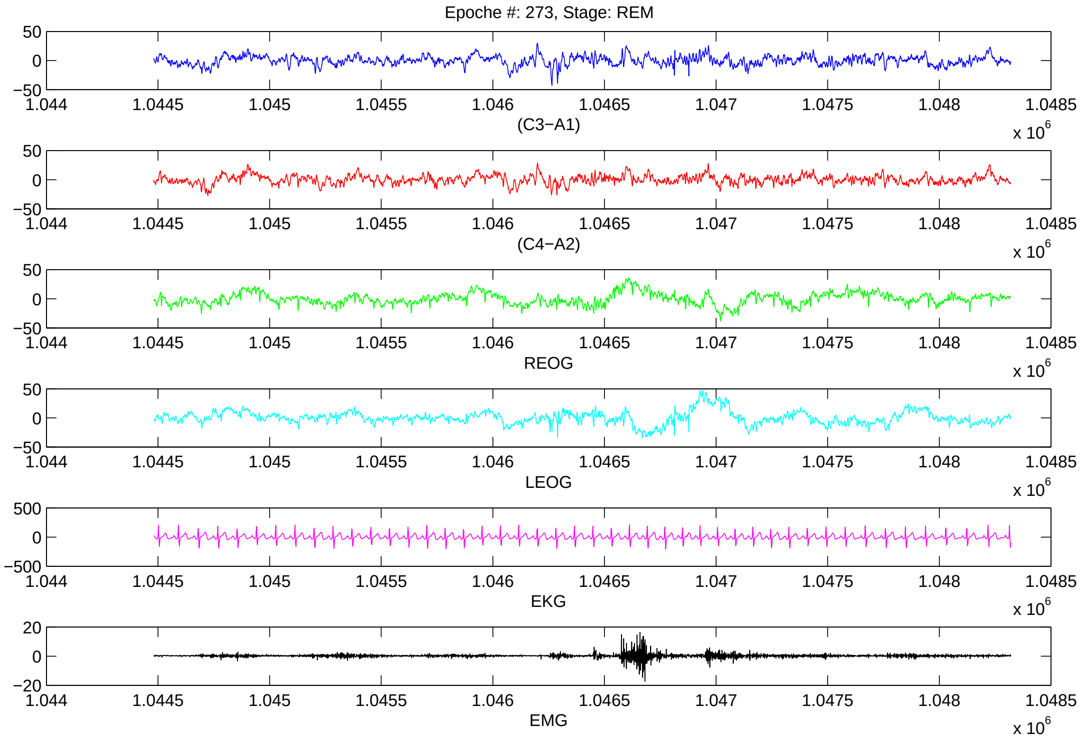

Figure D.5: Observed channels from sleep stage - REM

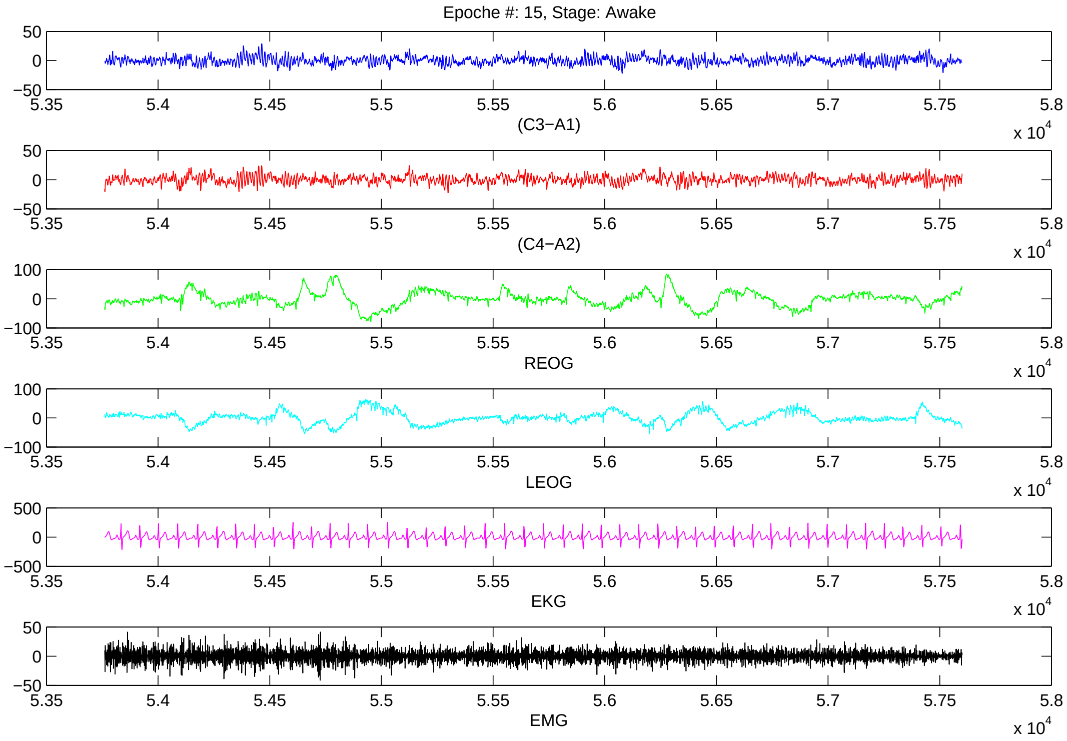

Figure D.6: Observed channels from sleep stage - Awake

Bibliography

[1] B. Y. Kim and K. Park, "Automatic sleep stage scoring system using genetic algorithms and neural network," *Proceedings of the 22nd Annual EMBS International Conference*, July 2000.

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

[3] D.-U. J. Haejeong Park, KwangSuk Park, "Hybrid neural-network and rule-based expert system for automatic sleep stage scoring," *Proceedings of the 22nd Annual EMBS International Conference*, July 2000.

[4] H. Y. Masaaki Hanaoka, Masaki Kobayashi, "Automated sleep stage scoring by decision tree learning," *Proceedings of the 23rd Annual EMBS International Conference*, October 2001.

[5] M. C. Patrick Hamilton, "Adaptive removal of motion artifact," *Proceedings - 19th International Conference - IEEE/EMBS*, October 1997.

[6] T. Richey, *EEG instrumentation and technology.* Springfield, Ill.: Thomas, 1976.

[7] S. C. S. G. M. J. B. S. Romero, M. A. Mananas, "Reduction of eeg artifacts by ica in different sleep stages," *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, Semptember 2003.

[8] e. a. Tatjana Zikov, "A wavelet based de-noising technique for ocular artifact correction of the electroencephalogram," *Proceedings of the Second Joint EMBS/BMES Conference*, October 2002.

[9] T. G. C. Jose C. Principe, Sunit K. Gala, "Sleep staging automaton based on the theory of evidence," *IEEE Transactions on Biomedical Engineering*, vol. 36, May 1989.

[10] Y. S. Takamasa Shimada, Tsuyoshi Shiina, "Sleep stage diagnosis system with neural network analysis," *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1998.

[11] S. J. N. N. Nicolaou, "Temporal independent component analysis for automated artefact removal from eeg," *Cybernetics Intelligence Research Group, University of Reading, UK.*

[12] e. a. Tzyy-Ping Jung, "Removing electroencephalographic artifacts: comparison between ica and pca," *Proceedings of the 1998 IEEE Signal Processing Society Workshop*, August 1998.

[13] e. a. J. Wu, "Intelligent artefact identification in electroencephalography signal processing," *IEEE Proc.-Sci Meas. Technology*, vol. 144, September 1997.

[14] L. Satish and B. Nazneen, "Wavelet-based denoising of partial discharge signals buried in excessive noise and interference," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 10, April 2003.

[15] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, Inc., 2001.

[16] Q. Xu, "A significance test-based feature selection method for the detection of prostate cancer from proteomic patterns." University of Waterloo, 2004.

[17] "traincgb." Available on: `http://www.mathworks.com/access/helpdesk/help/toolbox/nnet/traincgb.html`.

[18] C. H. Kim and R. Aggarwal, "Wavelet transforms in power systems," *Power Engineering Journal*, April 2000.

[19] C.-J. Ku and T. Fine, "Testing for stochastic independence: Application to blind source separation," *IEEE Transactions on Signal Processing*, vol. 53, May 2005.

[20] "Gram-schmidt orthonormalization." Available on: `http://mathworld.wolfram.com/Gram-SchmidtOrthonormalization.html`.

[21] "Fixed-point iteration." Available on: `http://math.fullerton.edu/mathews/n2003/FixedPointMod.html`.

[22] S. Grossman, *Elementary Linear Algebra.* Saunders College Publishing, 1994.