Shirin Najdi

Mestre em Engenharia Electrotécnica e de Computadores

Feature Extraction and Selection in Automatic Sleep Stage Classification

Dissertação para Obtenção do Grau de Doutor em Engenharia Electrotécnica e de Computadores, Especialização em Processamento de Sinais

Orientador: Professor Doutor José Manuel Fonseca

Professor Associado com Agregação da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Júri:

Presidente: Doutor Luís Manuel Camarinha de Matos

Arguentes: Doutor Luís Miguel Parreira e Correia

Doutor Pedro Manuel Cardoso Vieira

Vogais: Doutor Luís Manuel Camarinha de Matos

Doutor José Manuel Matos Ribeiro da Fonseca

Doutor Rui Carlos Camacho de Sousa Ferreira da Silva

Doutora Maria Rita Sarmento de Almeida Ribeiro

Doutor André Teixeira Bento Damas Mora

Doutor António Augusto Ribeiro Lopo Nunes Martins

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Dezembro, 2018

Feature Extraction and Selection in Automatic Sleep Stage Classification

To my family

Acknowledgements

During the last years, I had the opportunity to embrace some research projects and maintain contact with people that helped me in this long journey. For that, I must thank them all: professors, colleagues, friends, and the most important one family.

I would like to express my deepest gratitude to my advisor, Professor José Manuel Matos Ribeiro da Fonseca for all the patient guidance and assistance. His support and valuable suggestions made this work possible. Working with him has been a real opportunity and definitely made me grow up not only in scientific but also in personal terms. His vision of life changed me to be more responsible and hard working. I definatley owe him a lot.

I wish to thank the members of the thesis accompanying committee, professor Rita Ribeiro and professor Rui Camacho, for the useful comments and suggestions that were provided during the development of this thesis.

Thank you to the Department of Electrical Engineering and UNINOVA research institute, for making this work possible, providing available facilities and resources.

I also would like to thank my research colleagues at CA3 group of UNINOVA for the support, companionship, and motivation: Professor André Mora and António Falcão. With them it has been like a journey into knowledge and companionship since I started to do research in this centre.

I have to highlight my colleague and friend Shabnam Pasandideh. Her motivational support was not ignorable and effected the progress of this thesis positively. I wish her all the best in the world for her professional and personal life.

To my dearest friends, I have to show my appreciation for the motivation, support, and friendship in finishing this thesis. Friends are the family that we choose…

I would also like to express my most profound gratitude to my wonderful parents, Noushin Aminzadeh and Hassan Najdi. They raised me teaching the meaning of life and the relevance of having a good education. They also motivated me in trying to achieve all the goals that I proposed to, and still do. I could never reach this without them. I really have to thank their patience in all these years away from home, companionship, love and constant smile that they gave me, without which it would never be possible to finish this work and I feel so blessed that I have both of them. To my brothers, Said and Masoud, a very special thank for the continuous motivation and precise word of encouragement or support in the right moment.

Above all, I would like to specially thank my beloved husband and colleague Ali for his love and constant support, for all the late nights and early mornings, and for always being there for me. Thank you for being my muse, editor, proof-reader, and beside all of them a true soal match. But most of all, thank you for being my best friend. Without you it was impossible.

Abstract

Sleep stage classification is vital for diagnosing many sleep related disorders and Polysomnography (PSG) is an important tool in this regard. The visual process of sleep stage classification is time consuming, subjective and costly. To improve the accuracy and efficiency of the sleep stage classification, researchers have been trying to develop automatic classification algorithms.

The automatic sleep stage classification mainly consists of three steps: pre-processing, feature extraction and classification. In this research work, we focused on feature extraction and selection steps. The main goal of this thesis was identifying a robust and reliable feature set that can lead to efficient classification of sleep stages. For achieving this goal, three types of contributions were introduced in feature selection, feature extraction and feature vector quality enhancement.

Several feature ranking and rank aggregation methods were evaluated and compared for finding the best feature set. Evaluation results indicated that the decision on the precise feature selection method depends on the system design requirements such as low computational complexity, high stability or high classification accuracy. In addition to conventional feature ranking methods, in this thesis, novel methods such as Stacked Sparse AutoEncoder (SSAE) was used for dimensionality reduction.

In feature extration area, new and effective features such as distance-based features were utilized for the first time in sleep stage classification. The results showed that these features contribute positively to the classification performance. For signal quality enhancement, a loss-less EEG artefact removal algorithm was proposed. The proposed adaptive algorithm led to a significant enhancement in the overall classification accuracy.

Keywords: Sleep stage classification, Feature extraction, Feature selection, Rank aggregation, Distance-based features, Accuracy, Stability, Similarity, Feature vector quality.

Resumo

A classificação das fases do sono é vital para o diagnóstico de muitos problemas relacionados com a qualidade do sono sendo a polissonografia (PSG) uma ferramenta muito importante nesse sentido. No entanto, o processo visual de classificação das fases do sono é demorado, subjetivo e caro. Para melhorar a precisão e aumentar a eficiência da classificação das fases do sono, diversos trabalhos têm sido desenvolvidos no sentido de permitir a sua classificação automática através de algoritmos informáticos.

A classificação automática das fases do sono é composto por três etapas principais: pré-processamento, extração de características e classificação. O trabalho apresentado nesta Tese foca-se essencialmente nas etapas de extração e seleção de características. O principal objetivo desta Tese foi identificar um conjunto de características tão reduzido quanto possível mas suficientemente robusto e fiável que possa permitir a classificação eficiente das fases do sono com o mínimo de recursos. Para atingir esse objetivo, são dadas três tipos de contribuições na seleção das sinais adquiridos, na extração de características e no melhoramento da qualidade do vetor de características.

Vários métodos de classificação de características e de agregação características foram avaliados e comparados para encontrar o conjunto de sinais mais adequado à classificação.

Os resultados da avaliação efectuada indicaram que a decisão sobre o método de seleção de características depende dos requisitos da aplicação sendo esta influênciada por diversos parâmetros como a complexidade computacional, a estabilidade da classificação e a sua precisão. Além dos métodos convencionais de classificação de características, nesta tese, novos

métodos como o *Stacked Sparse AutoEncoder* (SSAE), foram utilizados para conseguir reduzir a dimensionalidade do problema.

Na área da extração de características, foram utilizadas pela primeira vez para a classificação das fases do sono características tais baseadas na diferença entre sinais (*distance-based features*) que, de acordo com os resultados obtidos, se revelaram de grande eficácia contribuíndo significativamente para o bom desempenho da classificação. Para melhorar a qualidade do sinal, foi também proposto um algoritmo adaptativo de remoção de artefatos sem perdas para os sinais EEG. Como se demonstra, o algoritmo proposto permitiu um aprimoramento significativo na precisão geral da classificação.

Palavras-chave: Classificação das fases do sono, Extração de características, Seleção das sinais adquiridos, Agregação características, Características baseadas em distância, Precisão, Estabilidade, Similaridade, Qualidade do vetor de características.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AASM	American Academy of Sleep Medicine
ANOVA	Analysis of Variance
AR	AutoRegressive
CMIM	Conditional Mutual Information Maximization
CWT	Continuous Wavelet Transform
DTCWT	Dual Tree Complex Wavelet Transform
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyogram
EOG	Electro-oculogram
FCBF	Fast Correlation Based Filter
FFT	Fast Fourier Transform
IG	Information Gain
KDR	Kernel Dimensionality Reduction
kNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LE	Lyapunov Exponent
LZC	Lempel-Ziv Complexity
MCDM	Multi-Criteria Decision Making
MLF	Multi-layer feed-forward
MODWT	Maximum Overlap Discrete Wavelet Transform
mRMR	minimum Redundancy Maximum Relevance
NREM	Non-Rapid Eye Movement
NSD	Normalized Slope Detectors
OAA	One-Against-All

OAO	One-Against-One
P2P	Peak to Peak
PCA	Principal Component Analysis
PSD	Power Spectral Density
PSG	Polysomnography
REM	Rapid Eye Movement
RKHS	Reproducing Kernel Hilbert Spaces
RQA	Recurrence Quantification Analysis
RRA	Robust Rank Aggregation
SBS	Sequential Backward Selection
SFS	Sequential Forward Selection
SSAE	Stacked Sparse AutoEncoder
STD	Standardization
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
SWS	Slow Wave Sleep
TQWT	Tuneable Q-factor Wavelet Transform
ZCR	Zero Crossing Rate

Chapter 1

1. Introduction

1.1 Problem Statement and Motivation

Sleep is fundamental for physical and mental health. As a physiological condition, it can be defined in many ways. For example, in [1] sleep is defined as a "reversible state of inactivity associated with decreased responsiveness". The decline in the responsiveness to environmental stimuli is like the coma but, unlike coma, this state is rapidly changeable to wakefulness with usually full cognitive capabilities.

Normal human sleep consists of two distinct stages with independent functions known as Non-Rapid Eye Movement (NREM) and Rapid Eye Movement (REM) stages. In their ideal situation, NREM and REM states alternate regularly, each cycle lasting 90 minutes on average. NREM sleep accounts for 75 to 80% of sleep duration and REM sleep accounts for 20-25% [2]. According to the American Academy of Sleep Medicine (AASM) [2], NREM can be subdivided into three stages: stage 1 or light sleep (N1), stage 2 (N2) and stage 3 (N3) [1], [3].

Sleep stage classification is vital for diagnosing many sleep related disorders. For this aim, a multiple-parametric test, called polysomnography (PSG) [1] is usually used. PSG recordings contain several bio-signals including Electroencephalogram (EEG), Electro-oculogram (EOG), chin

electromyogram (EMG), leg electromyogram (EMG), airflow signals, respiratory effort signals, oxygen saturation, body position, and electrocardiogram (ECG) recorded in overnight sleep. During staging, each epoch (i.e. a 30-second segment of PSG) is assigned to one of the five stages (wake, N1, N2, N3 and REM) according to the activity observed in that time interval.

The sleep stage classification process is, mainly done by an expert in a clinic or hospital environment. A collection of rules has been identified in AASM to guide the practitioners. However, the visual process of sleep stage classification is time consuming, subjective and costly. To improve the accuracy and efficiency of this process, researchers have been trying to develop automatic classification algorithms.

The automatic sleep stage classification mainly consists of three steps: pre-processing, feature extraction and classification. The pre-processing step includes artefact rejection and/or correction. In the feature extraction step, researchers try to compactly represent PSG recordings by means of a feature vector. In most cases, to enhance the efficiency of the feature vector dimensionality reduction and feature selection methods are used. Finally, in the classification step, the extracted feature vectors are assigned to one of the five categories using a proper classifier. Although significant amount of work has been done on this area, still there exist challenges and open issues which need to be resolved. Some of these open issues are summarized in the following list:

1. *Large and imbalanced data*: raw data of one subject for 8 hours with sampling frequency of 200 Hz will result in a single file with about 250 MB. Managing and processing this data needs reliable and sufficient computational resources. Moreover, the distribution of

stages is not always fair. For example, over 55% of the records are N2 and about 5% are N1 and N3 [4].

2. *Noisy data*: the presence of noise and artefacts in the data may lead to unusual numerical values in the extracted features and reduce the accuracy of the classification results.
3. *Inconsistency in the human PSG scoring*: the results of sleep scoring from two different practitioners are often not consistent. It has been reported that there is a considerable inter-scorer variability (about 20% disagreement) among scorers. Such differences are typically the result of rapid transitions between stages which create ambiguous stages [5].
4. *Difference between AASM-based scoring and commonly used signal processing methods*: experts learn the shapes and visual characteristics of the waves while signal processing methods cannot always reproduce the AASM rules and in some cases may completely ignore them. This leads to an inconsistency in the results of automatic and visual sleep stage classification [4].

The moment that the existing challenges are solved to a satisfactory level, the automatic sleep stage classification algorithms will be reliable enough to be routinely used in the clinical environments and at-home monitoring systems. In this thesis, we will address the forth open issue, trying to reduce the gap between manual and automatic classification results. The main motivation for this work is to develop a feature set to characterize each sleep stage in a way that extracted features are sufficiently powerful to distinguish sleep stages from each other and, on the other hand, are compact enough to reduce the dimensionality and improve the classifier's performance. Moreover, since having access to labelled PSG recording is not always easy,

this work is aimed to design a system that can work even with small amounts of labelled data.

1.2 Research Question and Hypothesis

The performance of an automatic sleep stage classification algorithm is deeply affected by the features provided to the classifier. Therefore, proper feature extraction and selection play an important role in the automatic sleep scoring process. Besides the significant amount of work done in this area, there are still challenges that need to be addressed. The most important challenge is the characterization of sleep stages in such a way that ambiguity in classification is minimized. For example, most of the classifiers cannot discriminate N1 from REM because the currently used feature sets are inadequate to discriminate them properly.

Non-robust and redundant features are two other challenges that current automatic sleep stage classification systems face. A feature is robust if it has low inter-subject variation as well as low sensitivity to signal acquisition parameters. On the other hand, a feature set is redundant if its features are highly correlated. Addressing these challenges will contribute to the implementation of more efficient automatic sleep stage classification systems.

Having these challenges in mind, the proposed research question is as follows:

> How can a robust and non-redundant feature set, be extracted in a way that it is efficient and reliable for adequately differentiating sleep stages?

To better analyse and interpret the main research question, six research sub questions are proposed:

1. How should a subset of PSG recordings be selected?
2. How can we effectively enhance the signal quality to extract better features?
3. What should be the strategy for feature extraction, or in other words, how should we decide about the type of features to be extracted?
4. What are the measures to assess the discriminative ability of the features?
5. Are there other methods to extract the desirable features rather than conventional methods?
6. How can feature selection methods contribute to find non-redundant and robust features?

Keeping in mind the research question, previously mentioned, the following hypothesis is proposed:

A desired feature set can be designed if,

— The quality of data is enhanced through the use of a loss-less artefact rejection method.

— A suitable dimensionality reduction or feature selection method is adapted/developed to select non-redundant and robust features.

— In addition to conventional feature extraction methods, new features and feature selection methods are utilized for differentiating sleep stages.

Considering the research sub-questions, first, two main issues should be defined, namely the PSG recording subset to be used and the strategy for enhancing the selected signals without losing data. The next step is to determine the type of features to be extracted. The third important issue is to find the most suitable criteria for evaluating the discriminative power and stability of extracted features, other than the existing criterion: accuracy. Finally, for feature selection, it is important that the pros and cons of different feature selection methods for sleep stage classification be investigated and the most suitable method be adapted to the problem at hand.

1.3 Research Method

The proposed work is aimed at performing research in automatic sleep stage classification to improve the process of feature extraction and selection through the usage of innovative signal processing methods. To achieve such result, this thesis work followed the classical research method that consists of seven main phases, as illustrated in Figure 1.

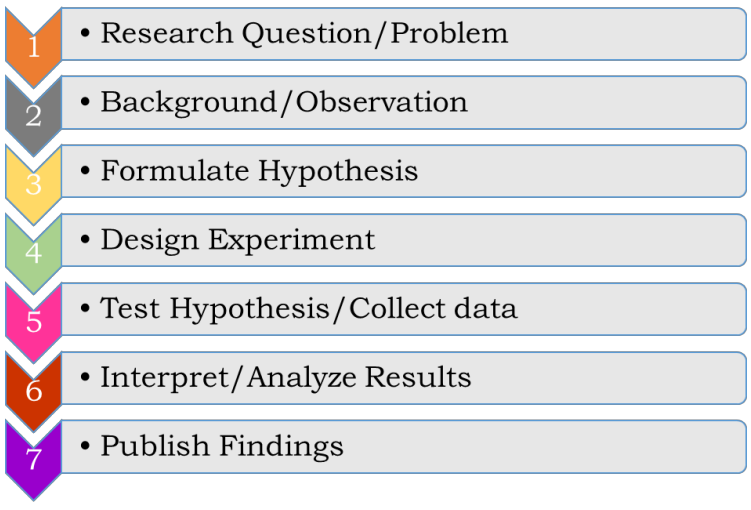

Figure 1. Classical research method adopted from the handouts of the Scientific Research Methodologies and Technologies course of the PhD program in Electrical and Computer Engineering by Professor Luis Camarinha-Matos.

Following this method, the research work was planned and scheduled according to the seven main phases:

1. **Research Question / Problem:** identification of the working context and motivation to formulate the research question.
2. **Background / Observation:** analysis of the state of the art in research and practice. In this observation and analysis, some main topics are addressed, namely: related background in PSG subset selection, feature extraction and selection methods.
3. **Formulate Hypothesis:** formulation of the hypothesis according to some preliminary analysis of the main problem and the current state of the art.
4. **Design Experiment:** split into three phases: first increasing feature vector quality, second the development and implementation of methods for selecting the best manually extracted features, followed by implementation of methods to extract new and innovative features
5. **Test Hypothesis / Collect Data:** application of the widely used open access sleep data for the validation scenarios. Results were collected for analysis and evaluation.
6. **Interpret / Analyse Results**: analysis and evaluation of the model, methodology and proposed tools in selected validation scenarios.
7. **Publish findings**: in parallel to all previous phases, there was a continuous publishing of the work findings, in recognized conferences and journals, being the work finalized with this thesis document, combining all the findings that were published and the final remarks.

Although the described phases might give the impression of a sequence, there are some iterations among them. As an example, after implementing, testing and interpreting some results, there was the need to make some reformulation in the hypothesis and corresponding model design to achieve results that were more accurate.

1.4 Thesis Structure

This thesis document is divided into six chapters:

Chapter 1. Introduction: Introduces the problem and motivation for the proposed research work, related to the improving of the feature vector quality in automatic sleep stage classification using innovative signal processing methods. This leads to the main research question and corresponding hypothesis. This chapter also includes a description of the research method and finishes with outlining the thesis structure.

Chapter 2. Background: Provides a baseline for the proposed research work. The history and technical background of manual and automatic sleep stage classification are described in this chapter.

Chapter 3. Literature Review: Introduces a literature review in techniques for developing a suitable feature vector to be fed to the classifier. This includes various feature extraction and selection methods. Also, some other related areas are considered including PSG subset selection, feature post-processing and normalization methods.

Chapter 4. Data and Methods: Describes the research material, especially the database and the main methods used for the development of the proposed algorithms. This chapter also helps to present the main logic behind selecting the techniques and tools used in this research work.

Chapter 5. Methodology and Results: Presents in detail the main contributions of this thesis work together with the details of the developed experiments designed to validate and support the proposed feature extraction and selection methods. This chapter also includes the corresponding results for the validation experiments.

Chapter 6. Discussion and Conclusion: Provides the discussion for the main findings of this thesis work focusing on the pros and cons of the proposed methods compared to the state-of-the-art methods. This chapter also concludes the thesis document and includes some possible directions for further research.

Chapter 2

2 Background

Sleep is one of the few physiological conditions that has received much attention by the scientists and scholars through the ages. In Aphorism LXII, Hippocrates wrote: SOMNUS, VIGILIA, UTRAQUE MODUM EXCEDENTIA, MORBUS – Disease exists if either sleep or watchfulness be excessive [6], [7]. Sleep is essential for human physical health and cognitive function. It is deeply connected to some of important physiological and cognitive mechanisms such as hormone release and immune function. Alterations in circadian rhythms and chronic sleep deprivation may lead to obesity, hypertension, heart disease and immune system dysfunction [8]. On the other hand, it is possible that disturbances in one's amount or quality of sleep are the symptom of another medical or mental problem. Therefore, sleep qualification and diagnosis of sleep related problems is of crucial importance.

2.1 Polysomnography (PSG)

The ground-breaking advances in understanding the cause of sleep disorders and the anatomy of sleep/wakefulness were only made after the middle of the twentieth century [6]. Currently, it is known that normal human sleep generally consists of two distinct stages with independent functions known as Non-Rapid Eye Movement (NREM) and Rapid Eye

Movement (REM) sleep. To differentiate these two stages, researchers rely on neurophysiological measures. These measures include Electroencephalogram (EEG), Electromyogram (EMG), Electrooculogram (EOG), and Electrocardiogram (ECG). While EEG has been a key element in analysing the sleep quality as well as diagnosing sleep disorders, EMG and EOG turned out to be useful in recognizing REM sleep [6]. Currently in clinical practice, PSG is regarded as the gold standard for recording and objective assessment of sleep related patterns. During PSG test, several bio-signals including EEG, EOG, chin EMG, leg EMG, airflow signals, respiratory effort signals, oxygen saturation, body position and ECG are recorded in a clinical environment.

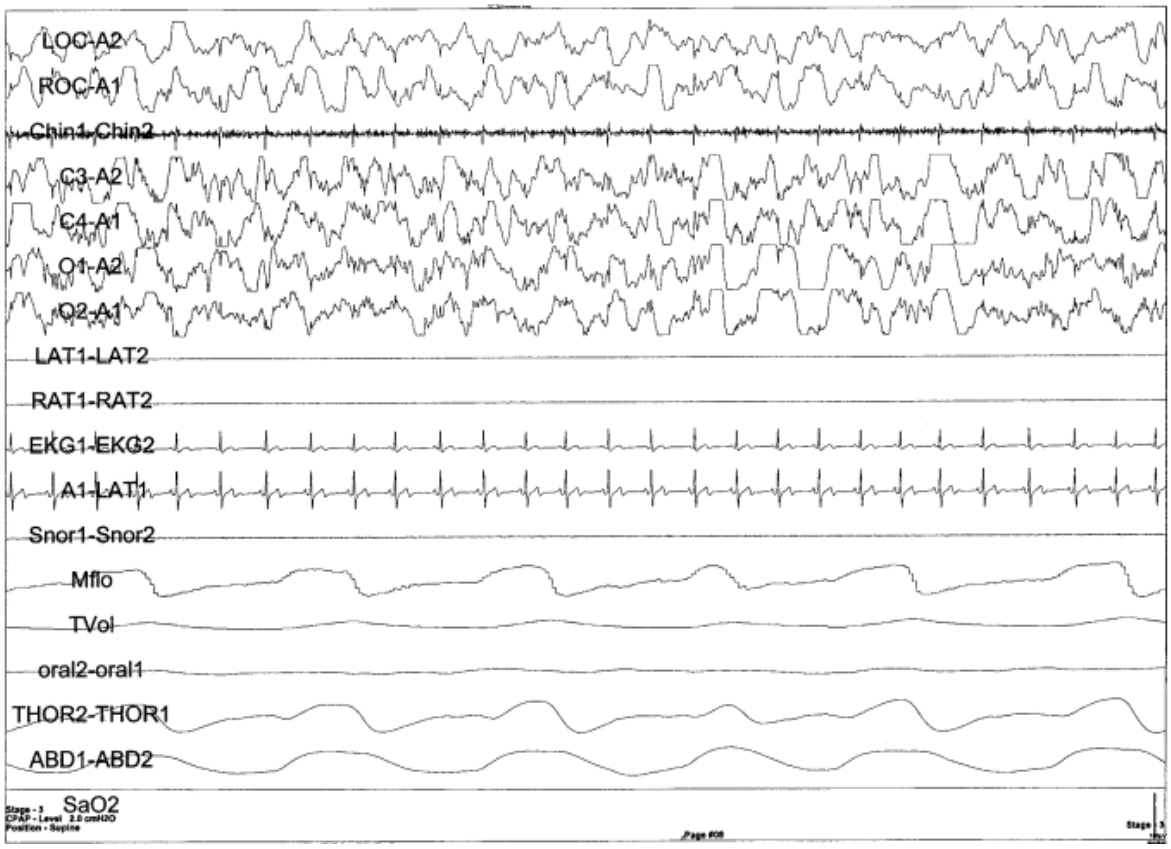

Figure 2. 30 seconds PSG of a 35-year-old woman in N3 stage [9].

2.2 Manual Sleep Stage Classification

Since 1975, AASM has developed guidelines and standards for practicing sleep medicine using PSG. According to the latest version of AASM guidelines [2], NREM stage is subdivided into three stages: N1 or light sleep, N2 and N3 or Slow Wave Sleep (SWS). Therefore, considering wakefulness, five distinct stages are considered in sleep analysis: Wake, N1, N2, N3, and REM. Figure 2 shows an example of PSG recording of a 35-year-old woman in N3 stage.

Usually, sleep stages are scored by a sleep expert through visual inspection in a clinic or hospital environment. According to AASM, each epoch (30-second segment of PSG) is assigned to one of the five sleep stages consistent with the activity observed in that time interval. The resulting series of discrete sleep stages are referred to as hypnogram. Figure 3 shows an example of a hypnogram for an 8-hour long sleep. In this figure, S1 refers to stage 1 (N1), S2 refers to stage 2 (N2) and SWS refers to stage 3 (N3).

Each epoch of the sleep is characterized by the presence of special characteristics of physiological signals. Especially EEG waves have been proven to be useful in distinguishing sleep stages [2]. For instance, the wake stage with eyes open is characterized by the presence of low amplitude mixed EEG frequencies (Alpha and Beta) and probable body movements. Beta waves are defined as low amplitude and high frequency waves being dominant during wake stage. While the eyes are open, alpha wave amplitude

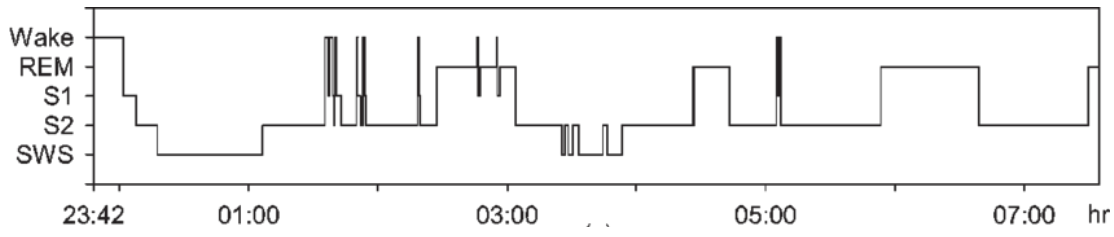

Figure 3. A sample hypnogram for an eight-hour long sleep [10].

is much lower compared to eyes closed state. During the wakefulness with eyes closed more than 50% of the epoch contains alpha activity. Slow eyes movement is also detectable in the EOG [11] .

N1 stage is the transition between the wake and sleep stages. It is identified by the reduction of alpha rhythm and the appearance of low amplitude theta. During this stage, both the respiration rate and the cardiac rate decrease. The N2 stage is characterized by the presence of *k*-complex waves (a negative high voltage sharp wave) followed by sleep spindle bursts (with frequency range between 12 and 14 Hz).

Delta waves usually occur in N3 stage. They are characterized by high amplitude (between 20 and 200 μV) and low frequencies (below 5 Hz). The REM stage is known as paradoxical sleep since it is characterized by low amplitude, irregular and mixed brain waves. The brain activity at REM is like the wake state and the incidence ratios of delta and spindle wave decrease. Rapid eyes movements appear, EOG waves are similar to stage wake and the chin becomes relaxed [12]. In Table 1, the specifications of each stage are summarized. In this table TST stands for Total Sleep Time.

After the acquisition of the PSG, the data is scored by a technician according to a collection of rules set forth by AASM. The presence of skilful technicians and physicians is necessary for assuring the quality of recording and analysis. According to AASM criteria, the scoring should be done on 30-second, sequential epochs starting from the first sample of the data. For each stage, a number of recommended definitions are presented. These definitions mainly include EEG frequency and waveform, eye blinks and movements and EMG amplitude.

Table 1. EEG, EOG and EMG characteristics of sleep stages [13] (TST: Total Sleep Time).

Sleep Stage	TST (%)	EEG	EOG	EMG
Wake	-	Alpha activity (8-12 Hz) or low-amplitude beta (13-35 Hz), mixed-frequency waves	REM (in sync or out of sync deflections), eye blinks	Relatively high tonic EMG activity
N1	2-5	Low-voltage, mixed-frequency waves (2-7 Hz range), mainly irregular theta activity, triangular vertex waves	Slow eye movements, waxing and waning of alpha rhythm	Tonic EMG levels typically below range of relaxed wakefulness
N2	45-55	Relatively low-voltage, mixed-frequency waves, some low-amplitude theta and delta activity	No eye movement	Low chin muscle activity
N3	5-20	≥20%-50% of epoch consists of delta (0.5-2 Hz) activity	No eye movement	Chin muscle activity is lower than N1 and N2
REM	20-25	EEG is relatively low voltage with mixed frequency resembling N1 sleep	Episodic rapid, jerky, and usually lateral eye movements in clusters	EMG tracing almost always reaches its lowest levels owing to muscle atonia

2.3 Automatic Sleep Stage Classification

Manual scoring of sleep stages has some disadvantages. First, it is time consuming. Usually it takes hours to score the PSG of a whole night sleep. Second, the results of sleep scoring from two different practitioners are often not consistent. It has been reported that there is a considerable inter-scorer variability (about 20% disagreement) among scorers. Such differences are typically the result of rapid transitions between stages which create ambiguous stages [5]. Moreover, with the immergence of at-home sleep monitoring systems, there is an urgent need for unsupervised methods that can efficiently score the sleep data in a way that the results are medically

reliable. Therefore, developing automatic sleep stage classification algorithms has been the focus of many researchers.

Figure 4 shows a general block diagram of automatic sleep stage classification. The common approach in automatic sleep stage classification, like any other pattern recognition process, includes pre-processing, feature extraction and classification steps. The pre-processing step includes artefact rejection and artefact correction. Artefacts are unwanted signals not produced by the desired physiological events. Power line noise (50 Hz EU/60 Hz US), electrical equipment noise, sweat and pulse spikes are some examples of non-biological and biological artefacts. Some of these artefacts can be easily removed by using a simple notch filter but some others need more advanced signal processing techniques to be rejected or corrected.

Features are extracted from a subset of raw PSG recordings containing only raw EEG data or EEG data together with other raw PSG signals, acquired. For each sleep stage, most of the features used try to describe the presence of these special waves, their duration and properties. This feature vector should be informative and non-redundant enough to facilitate the subsequent classification step. Various types of features have been extracted from PSG recordings and used in the literature. Besides, different types of dimensionality reduction and feature selection methods have been applied to find the most valuable subset of features. These features and techniques together with other related processing such as PSG subset selection and feature post-processing are described in the next chapter.

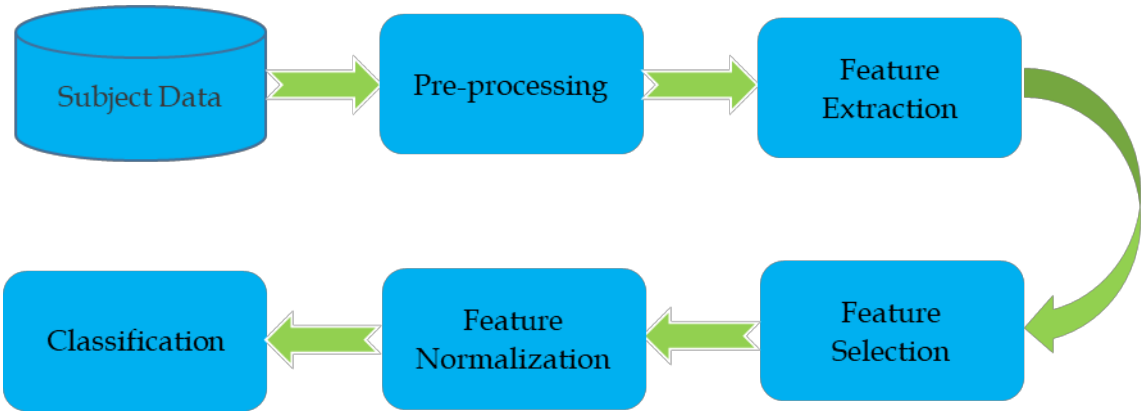

Figure 4. Block diagram of automatic sleep stage classification

2.4 Summary

In this chapter, the basics of manual and automatic sleep stage classification were discussed. Specifically, some of the challenges of manual sleep staging were mentioned as the grounds for the emergence of automatic methods. Finally, the main steps of automatic staging were described. In the next chapter, the state of the art methods for feature extraction and selection will be described in detail.

Chapter 3

3. Literature Review

3.1 PSG Subset Selection

In manual sleep scoring, technicians and doctors use PSG recordings and AASM rules for characterizing sleep. There are a number of recommended parameters that must be reported for a PSG study. At minimum, three EEG channels (frontal, central and occipital derivations) plus two EOG channels (from left and right eyes) and two chin EMG channels are necessary to perform manual sleep scoring. For describing the location of scalp electrodes, AASM uses the international 10-20 system [14] according to Figure 5. Particularly, the recommended EEG channels by AASM include F4-M1, C4-M1 and O2-M1. If it is not possible to use these channels, alternative EEG channels set include Fz-Cz, Cz-Oz, and C4-M1.

Inspired by this procedure, researchers try to mimic the visual sleep scoring process by using a proper subset of PSG recordings in automatic sleep stage classification. This subset usually includes EEG, submental EMG and EOG. There are no clear hints or clues in the literature about the strategy or reason of selecting a special subset of PSG recordings, except in papers that design a system for a specific signal such as single channel EEG. In Table 2, a summary of PSG subsets used in the literature is presented. Papers summarized in this table include studies that classify sleep recordings into 2 stages (REM/Non-REM or Sleep/Wake), 3 stages, 4 stages,

5 stages or 6 stages. Studies that detect patterns such as spindles, k-complex or sleep disorder detection papers are not included in this table.

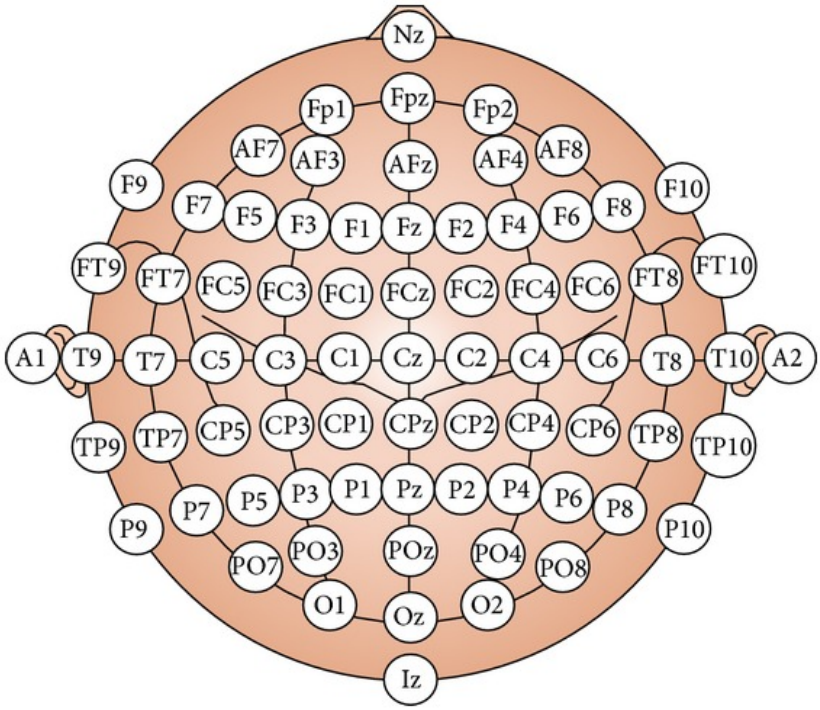

Figure 5. The 10–20 system of electrode placement [15].

Table 2. Summary of PSG subsets used in sleep stage classification.

Subset Type	Signal	Channels	References
Single Channel	EEG	C3-A2	[10], [16]–[32]
		C4-A1	[33][27][30][34]
		C3-A1	[35]
		Fpz-Cz/Pz-Oz	[21], [23], [28], [36]–[54]
		F3-A2	[30]
		F4-A1	[30]
		O1-A2	[30]
		Cz-Pz	[55]
		A1-A2	[56]

		Cz-A1	[57]
	EOG	Left EOG	[58]
		E2-E1	[59]
	ECG		[60]–[62]
Multi-Channel	EEG, EOG, and EMG	EEG (C3-A2), Left and Right EOG, and chin EMG	[63]
		Six EEG (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, O2-A1), Left and Right EOG, and chin EMG	[64], [65]
		Four EEG channels (C3-A2, P3-A2, C4-A1, and P4-A1), one horizontal EOG and one chin EMG	[66]
	EEG and EOG	EEG (C3 and Cz), Left and Right EOG	[67]
		Six EEG channels (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, O2-A1) and two EOG channels (Left and Right)	[68][69]
		EEG (Pz-Oz) and Horizontal EOG	[70]
		Two EEG (Fz and Oz) and two EOG (Left and Right) Channels	[71]
		EEG (C4-M1), EOG	[72]
	Heart Rate, Breathing Rate and Movement In-formation	Heart Rate, Breathing Rate and Movement Information	[73]
	EEG, ECG and Respira-tion Features	EEG (C1-A2), ECG and Respiration Features	[74]
		ECG and respiratory inductance ple-thysmography (RIP)	[75]
	EEG and EMG	EEG (C4-M1) and chin EMG	[76]
		EEG (C3-A2) and chin EMG	[77] [78]

ECG, Respir-atory and actigraphy and signals	ECG, Respiratory and actigraphy and signals	[79]
EEG	Fp1-C3, Fp2-C4, Fp1-T3 and Fp2-T4	[80]
	Pz, Cz, Pz, T3, T4	[81]
	Six EEG channels (Fp1–M2, C3–M2, O1–M2, Fp2–M1, C4–M1, and O2–M1)	[82]
	Fpz-Cz and Pz-Oz	[83], [84]
	C3-A2 and C4-A1	[85] [86]
	Six EEG Channels (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1 and O2-A1)	[87]
	C4-A1, O2-A1 and C3-O1	[88]
EOG	Left and Right	[89][90]
EEG, EOG, EMG and ECG	Six EEG channels (Fp1–M2, C3–M2, O1–M2, Fp2–M1, C4–M1, and O2–M1), two EOG channels (Left and Right), one chin EMG channel and ECG.	[82][91]

According to this table, there are, in general, two different approaches: single and multi-channel. In single channel approaches, it is assumed that one signal is sufficient and contains enough information to classify epochs into sleep stages. Therefore, the algorithms can be implemented on a portable device suitable for home environment, clinical care and online applications [30], [33], [35]. Single channel EEG systems are the most common ones in this category. For multi-channel studies, there are several alternatives for channel combinations. The most common combination is a set of EEG, EOG and chin EMG signals. Although multi-channel systems have more computational complexity than single channel ones, several studies have shown that using the information from other channels rather

than single EEG channel can improve the distinguishing ability of the system between stages especially on REM and N1 [72], [82].

3.2 Feature Extraction in Sleep Stage Classification

Feature extraction is the first of the three main steps of automatic sleep stage classification. A wide range of features have been extracted and used in the literature from different subsets of PSG recordings. To evaluate and analyse the effectiveness of feature extraction methods, it is necessary to have an overview of the methods used in the literature.

It should be considered that although feature extraction is a critical step in automatic sleep stage classification, the final performance of the scoring system, in addition to the extracted features, depends on the quality of the signals used (noisy or clean), selected PSG subset, and classification algorithm. In the following, we will review the different features used in the literature for sleep stage classification.

The main categories of the features used in sleep stage classification include frequency domain, time-frequency domain, time domain, and nonlinear features. In this section, the most common features of each category together with their advantages and disadvantages will be described.

Frequency Domain Features

Frequency domain features are the most widely used features in sleep stage classification. The prevalence of their usage is due to their ability in estimating EEG frequency bands that characterizes sleep stages. Also, they are not dependant on the age and gender of the subject. The Fast Fourier Transform (FFT) has been mainly used to describe the frequency content of EEG.

The most common spectral features are as follows:

- *Spectral power:* The absolute spectral power in four significant frequency bands is among the most widely used features in sleep stage classification. In addition to the absolute value, relative spectral power and spectral power ratios have been considered important due to the proportional changes of brain waves in different sleep stages. Relative spectral power is calculated by dividing the absolute power in each frequency band by the total spectral power. Power ratio is the relative spectral power in different frequency bands such as (alpha/beta) [24], [30], [33], [78], [88], [92].

- *Spectral entropy*: This feature is calculated based on Shannon's entropy and is a measure of the flatness in Power Spectral Density (PSD). Spectral entropy is considered suitable for discriminating between N1 and N3 [30], [33], [91]–[93].

- *Statistical parameters*: Spectral moments describe the shape of the PSD of the PSG recordings. Spectral mean, variance, skewness and kurtosis fall into the category of statistical parameters extracted from PSG signals [24], [30], [33], [94].

- *Harmonic parameters*: This type of features, although not very common, are used in some papers [12]. Central frequency, bandwidth and power of the central frequency are some of the harmonic features extracted from PSG recordings.

- *Other spectral features*: There are other spectral features used in sleep stage classification that cannot be completely categorized in one of the above groups. Spectral edge frequency is one of those features, commonly interpreted as the frequency which 95% of the total spectral power is located below it. In [91] this feature has been found useful

for discriminating the wake-N2 and wake-N3 stages. Peak power frequency [33] that was originally used for estimating the depth of anaesthesia, is also common in sleep analysis applications. Percentile is another feature that provides some useful information about the amplitude of the signal. For example, percentile75 EEG provides an indication on the amplitude level of electrical brain activity and can be useful to distinguish relatively high amplitude activity during wakefulness and N3 stages [66].

Most of the spectral features mentioned above are usually extracted from the EEG signal. However, it is also possible to find papers in the literature that extract some of these features from EMG or EOG [58], [78], [95]

The most important shortcoming of frequency-based features is their disability in analysing non-stationary signals. Since PSG recordings are non-stationary by nature, joint time-frequency methods like Wavelet transform can be considered suitable alternatives.

Time-Frequency Domain Features

The range of time-frequency domain features is very diverse in sleep stage classification. The coefficients calculated by time-frequency methods are sometimes treated like Fourier coefficients to calculate spectral energy features [25]. Sometimes, they are regarded as a different representation of PSG recordings and used to extract temporal or nonlinear features that are usually extracted from the signal in the time domain [71].

For analysing non-stationary PSG recordings, Continuous Wavelet Transform (CWT) [35], Discrete Wavelet Transform (DWT)[16], [45], [52], [94] Maximum Overlap Discrete Wavelet Transform (MODWT), Choi-Williams distribution [35], Empirical Mode Decomposition (EMD) [43], [53], Hilbert-Hung transform [36] and Wigner Ville distribution [25], [96] are the most

commonly used time-frequency methods. In addition, recently the performance of two new signal decomposition methods, Dual Tree Complex Wavelet Transform (DTCWT) [17] and Tunable Q-factor Wavelet Transform (TQWT) [20], [97], were evaluated in sleep stage classification.

Time Domain Features

Time domain features can represent the morphological characteristics of a signal. They are simply interpretable and suitable for real-time applications. This category of features is used in sleep stage classification because they usually have less computational complexity and simulate the manual scoring process. There are several time domain features including:

- *Statistical parameters*: If the PSG recording is considered as a random process, stochastic modelling can be used for its analysis. Several papers in the literature [12], [22], [23], [25], [33], [94], [98], [99] have used stochastic modelling to extract statistical parameters such as first to forth moments, average amplitude, maximum or minimum amplitude and percentile from PSG recordings and especially from EEG. These parameters are computed for each epoch to measure the dispersion, the central tendency and the distribution and describe the wave shapes in the time domain. In [91] the EEG variance has been found useful in discriminating between N2-REM and N3-REM. In the same paper, skewness also showed acceptable performance in distinguishing N2 from REM.
- *Autoregressive model parameters*: The AutoRegressive (AR) model is a parametric model that represents the current value of a PSG recording as a linear combination of its previous samples plus a stochastic term that is imperfectly predictable. The computed regression coefficients

are commonly used as features in EEG analysis. Several methods exist for estimating AR coefficients such as least squares, Yule-Walker and Burg's method. By looking at the literature, it can be found out that the AR model parameters are not anymore among primarily used features in sleep stage classification. Although exact reasons for this issue should be sought, the stationarity requirement can be a cause for this method's unpopularity.

In [10] the goal is single-channel sleep stage classification. In this paper, the order of autoregressive model for EEG is chosen as eight and the computed eight auto-regression coefficients from theta band together with multiscale entropy features are fed to a Linear Discriminant Analysis (LDA) classifier.

In [100], sleep spindle detection has been done by using AR modelling for feature extraction. The authors tried to prove that the time domain characteristics of a signal can be used to discriminate EEG rhythms. For defining the model order, they didn't use the optimal model order selection methods like Akaike's information-theoretic criteria or Parzen's criterion of autoregressive transfer function [101], [102]. Instead, they tried different model orders to find out which order gives the best separable class of patterns. Their simulation results show that, although AR model coefficients provide a good representation of the EEG data, Short Time Fourier Transform (STFT) works better in characterizing spindle and non-spindle regions.

- *Hjorth Parameters*: In 1960, Bo Hjorth [103] proposed normalized slope detectors (NSD) as indicators of statistical properties of a signal in time domain. NSDs include three features: activity, mobility and complexity. These features are used in the analysis and characterization of EEG and sleep stage classification [25], [42], [104], [105].

- *Period Analysis-based Features*: Features like Zero Crossing Rate (ZCR) and its derivatives and peak to peak (P2P) amplitude are commonly used since they describe the time domain characteristics of the signal and are similar to manual scoring of sleep stages [25], [89], [104]. About ZCR, although it seems that for high accuracy scoring of sleep stages it should be used in combination with other features, it has some advantages like low computational complexity and ability to detect transient waves like sleep spindles and *k*-complexes.

Nonlinear Features

In the brain's neural network, nonlinearity is apparent even on the cellular level since the dynamic behaviour of individual neurons is governed by threshold and saturation phenomena. Moreover, the brain's ability to perform sophisticated cognitive tasks rejects the hypothesis of an entirely stochastic brain. In addition to the EEG, other signals acquired from the body neither have completely stochastic nature nor are stationary. Therefore, nonlinear signal processing techniques have been widely used for characterizing sleep signals. In the following, the most important nonlinear features used in sleep stage classification will be discussed.

- *Energy features*: Energy based features are the most common type of nonlinear features extracted from different sub-bands of PSG recordings in time domain [25], [48], [106]. In addition to the usual energy, Teager energy operator also has been proved to be useful in analysing sleep recordings [25].
- *Entropy estimators*: Entropy is a measure for evaluating the unpredictability of information content. So far, numerous entropy estimators have been proposed and used for discriminating sleep stages including:

- *Shannon Entropy*: This measure is usually considered as the most classic and foundational entropy measure. It has been used for EEG signal analysis in many applications including epilepsy detection, abnormality detection and emotional states discrimination [107], [108]. In [82], [106], Shannon entropy, in combination with other entropy features, is used for sleep stage classification.
- *Renyi Entropy*: In 1960, Alfréd Rényi introduced Rényi's general notion of entropy [110]. Since Rényi Entropy generalizes several distinct entropy measures, it turned out to be theoretically interesting and found many applications in various research areas such as pattern recognition [111] and biomedicine including sleep stage classification [25], [35], [71], [104].
- *Permutation Entropy*: Permutation Entropy [112] is a simple complexity measure, which can be applied to any type of time series including regular, chaotic, noisy and time series from reality. Low computational complexity of permutation entropy facilitates its use in the characterization of PSG recordings [25].
- *Approximate Entropy*: In time series analysis, approximate entropy is regarded as a measure to quantify the amount of randomness or equivalently regularity of time series [42], [113], [114]. A high value of this measure indicates randomness and unpredictability. In [26] changes in approximate entropy of EEG has been assessed during eyes-closed wake and other sleep stages in healthy subjects. Significant changes in approximate entropy have been found during different stages of

sleep with lowest values during stage 3 and highest values during REM.

- *Sample Entropy*: Sample entropy is a modified form of approximate entropy in which the bias existing in approximate entropy due to self-match patterns has been removed [115]. This measure has been widely used in sleep stage classification [71], [83].
- *Multiscale Entropy*: As previously mentioned, entropy measures the complexity of physiological signals. A wide range of diseases are associated with degraded physiological information and loss of complexity. However, certain pathologies exist that are associated with highly unpredictable fluctuations. For such cases, conventional methods would estimate an increase in the entropy compared to the healthy subjects. Multiscale entropy [116] estimates the long-range temporal correlation of time series to solve this problem. This measure has been applied to the analysis of ECG, heart rate and sleep EEG [40], [83], [90].

- *Fractal Dimension*: A structure exhibits fractal properties if similar details are observed on different scales [117]. Also, a time series can display fractal properties if statistical similarity emerges at different time scales of its dynamics. A signal is fractal if the scaling properties fit a scale-free behaviour, meaning that the same features of small-time scales emerge in large ones. This relationship is quantified by the fractal dimension. In other words, fractal dimension is a measure of signal complexity. The fractal dimension of a time series including PSG recordings can be computed by several different techniques such as Petrosian fractal dimension, Higuchi fractal dimension, Katz fractal

dimension and correlation dimension [42], [61], [83], [92]. Mean curve length was also proposed in the context of reducing the complexity of Katz fractal dimension algorithm and it provides results almost equivalent to Katz fractal dimension [118].

This measure has been used for analysing sleep signals in several applications. In [119] the behaviour of the fractal dimension during each of the neonatal EEG sleep stages and during the wake stage has been studied and the results are compared to the classical spectral parameters and zero crossing values. In [120] fractal dimension is used to analyse sleep EEG in healthy and insomniac subjects. The results show that each sleep stage can be characterized by a certain range of EEG fractal dimension, though no statistical significance was observed between healthy and insomniac subjects in any sleep stage. Finally, in [91], fractal dimension demonstrated satisfying performance in describing stage 1 as well as distinguishing wake stage from N3.

- *Hurst Exponent*: Hurst exponent is a non-linear chaotic parameter that has been used for assessing self-similarity and correlation properties of time series. Its values vary between 0 and 1 and when it exceeds 0.5, the signal is called persistent with consecutive trends. In sleep stage classification, there is no significant study that specifically evaluates Hurst exponent's ability in discriminating each sleep stage. Siiram et al. [121] evaluates its ability in distinguishing wake from sleep. Also in [25], [42], [54], [76], [122] Hurst exponent is used in combination with other linear and non-linear features in sleep stage classification.

Due to the non-stationary nature of physiological signals, often Detrended Fluctuation Analysis (DFA) is preferred to Hurst exponent. DFA permits the detection of intrinsic self-similarity embedded in a seemingly nonstationary time series, and avoids the false detection of apparent self-similarity, which may be an artefact of external trends. This method has been successfully used in a wide range of sleep studies including [61], [85].

- *Lyapunov Exponent*: Lyapunov Exponent (LE) gives the rate of exponential divergence from perturbed initial conditions. A system with a large LE is said less predictable. In [123], changes in the largest positive LE were investigated by using the sleep data of 15 healthy men. LE decreased from stage 1 to N3 and for REM, it was slightly lower than for stage 1. In general, the results show that LE decreases as the sleep goes deeper. Inspired by this work, the nonlinear analysis of sleep has become a major research topic. Generally, it can be concluded from these works that deeper sleep stages are associated with lower complexity as demonstrated by the LE values and this adds to the value of nonlinear features in sleep stage classification [42], [55], [124].
- *Lempel-Ziv complexity*: The Lempel-Ziv Complexity (LZC) for sequences of finite length was proposed by Lempel and Ziv [125] and represents a simple way to measure signal complexity. Although LZC still remains a rather unexplored measure, the studies show that it has a high potential to investigate neurophysiological events during sleep and wakefulness. One of the open issues about LZC is the number of necessary samples to robustly estimate LZC for different sleep stages [126]. In [127], the authors use LZC with the aim of going beyond results obtained with conventional techniques of signal analysis. Their

results reveal that the activated brain states (wake and REM sleep) are characterized by higher LZC compared with NREM sleep. In many other works LZC has been used in combination with other features for sleep stage classification [33].

LZC is not the only feature that remains rather unexplored. There are other features and analysis methods that can still be considered infantile regarding their application in sleep stage classification. Recurrence Quantification Analysis (RQA) is an example of such methods. RQA was originally a visual tool used for detecting the patterns of recurrence in the data. To go beyond visual evaluation, several recurrence quantification estimators are devised. These measures were used in [71]. Figure 3 shows the recurrence plots of two EEG segments at drowsy and alert states. As described in the Introduction chapter, alpha rhythm is dominant in drowsy state while beta rhythm is significant in alert state. This difference is apparent in the recurrence plots of these two states according to Figure 3.

- *Itakura Distance (ID)*: ID is a feature based on the AR assumption of the analysed process. It is widely used in speech processing and measures the similarity between two AR processes. In [128], ID has been used for extracting features from EEG for sleep stage classification. In 2005 Estrada et al. [72] tried to capture the temporal similarity of EEG and EOG by using ID. Their results were very promising. In addition to AR coefficients, the distance between spectral representations of the signals can also be used to measure similarity [129]. In this case, the extracted distance feature is called Itakura spectral distance.

As mentioned before EEG signals arise from a highly nonlinear nervous system and nonlinear features play an important role in this regard and yet

it is important to know if the information provided by nonlinear features can be obtained by conventional spectral features or not. Fell et al. [55], in a pilot study compared the spectral and nonlinear measures of EEG signals during sleep. They concluded that nonlinear features provide additional information that is not redundant to the information gained from spectral features. In other words, the information obtained from these two groups complement each other. For example, nonlinear measures like correlation dimension and Lyapunov exponent perform better in discriminating stage 1 and 2 compared to the spectral measures. On the other hand, spectral measures outperformed the nonlinear ones in separating stage 2 from N3. An overall review of the literature also confirms that researchers boost their proposed system's performance by taking advantage of different types of features.

In [46], a review of the existing EEG signal-based methods in three phases of pre-processing, feature extraction and classification is presented. Different features used in the design of sleep stage classification systems were analysed from the popularity point of view and results are shown in Figure 5. According to this analysis, 35% of the studies use non-parametric-based frequency-domain features (such as power, spectral flatness, spectral centroid, etc.), 24% use the Wavelet-transform-based time-frequency domain features, 25% use statistical standards based on the time domain and 6% use approximate entropy based on nonlinear, domain feature extraction measures. The standard statistics of the time domain, non-parametric statistics of the spectral domain and Wavelet transform of the time-frequency domain are the top three feature extraction methods that have received more attention in sleep stage classification schemes.

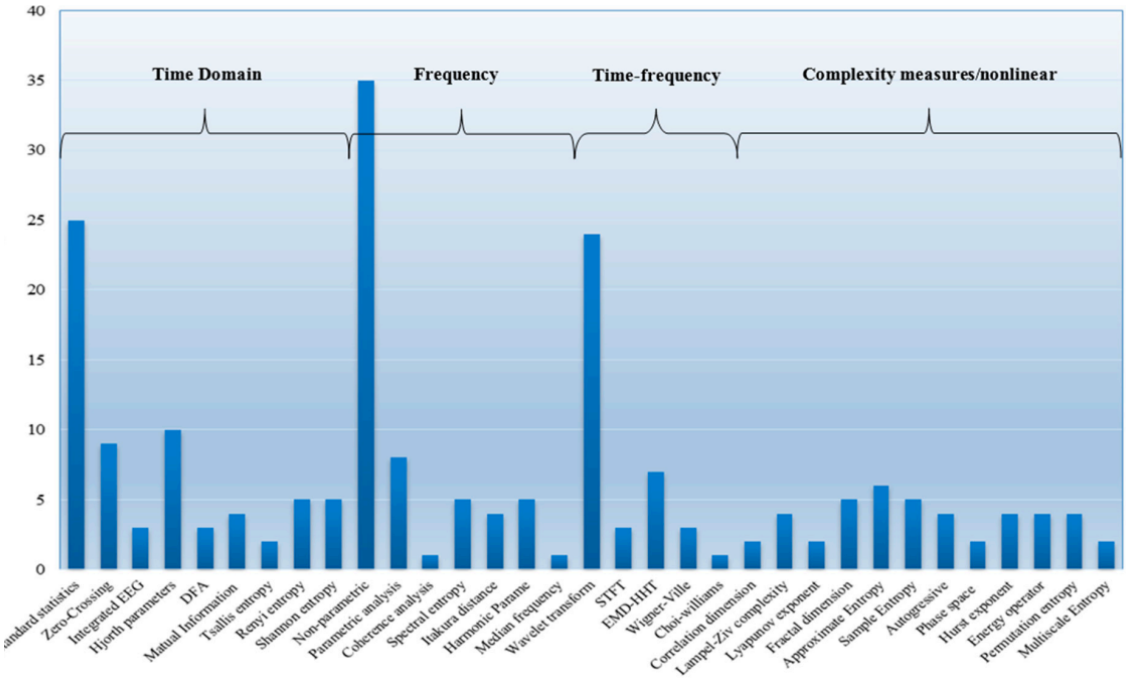

Figure 6. Summary of Feature in Automatic Sleep Stage Classification [46]

3.3 Dimensionality Reduction and Feature Selection in Sleep Stage Classification

As discussed before, in the feature extraction stage, several types of features can be extracted from PSG signals in different time and frequency domains. Nevertheless, some of these features may be redundant and/or irrelevant and increase the complexity of the model. Therefore, dimensionality reduction and feature selection have been important research topics for the researchers in data mining and machine learning areas.

Basically, the aim of feature selection in a classification task is to select the subset of features that best explain the difference between the different classes of the input data. Feature selection offers many advantages making it an apparent prerequisite on many classification systems. By selecting an

adequate subset of features, *more compact and simpler models* can be reached for the problem at hand reducing the computational time necessary for the classifier. The elimination of redundant and/or irrelevant features may also enhance the generalization ability as well as increase the classification power through reduced overfitting. Less storage memory and simplified visualization are further benefits of feature selection in classification tasks [130], [131].

Given the wide range of features utilized for sleep stage classification, the choice of the most efficient features to be implemented is difficult. There is no complete comparative study that considers the features performance (including the temporal, spectral and nonlinear features) and their accuracy to identify sleep stages. The major focus of the existing literature is on the proper feature extraction and dimensionality reduction. Feature selection methods are relatively overlooked. In the following, the sleep stage classification algorithms that incorporate one or more dimensionality reduction and feature selection methods will be discussed.

3.3.1 Dimensionality Reduction Methods

Principal Component Analysis (PCA) is a feature transformation method that reduces problem's dimension by projecting the original high dimensional data into a lower dimensional space. In other words, PCA transforms the original feature vector to a vector with linearly uncorrelated elements called principal components. These principal components are in such a way that the first one has the greatest variance and each succeeding principal component in its own turn has the greatest variance and is orthogonal to the preceding component [132].

In the context of sleep stage classification, Rempe et al. [133] used PCA for compacting the 7-dimensional energy-based feature vector extracted

from EEG and EMG signals. To answer the question why they applied PCA, they explained that by using the original feature vector, each epoch could be represented by a point in a seven dimensions space. If all the epochs of data were visualized in this space, at the end, a random cloud of data with no distinct pattern would be achieved. But if epochs were demonstrated using their principal components, they would be arranged in one or more directions different from the original coordinate axes. These directions are the most important components accounting for the greatest part of the variance in feature space. In this work, data dimension was reduced to three by keeping only the first three principal components.

Figure 7 shows the data plotted by first two principal components scored by human and naïve Bayes classifier. Distinct clusters are noticeable indicating that PCA could effectively separate the sleep states. Also, from the classification point of view, it is clear that human and machine scored data in a similar way.

In [134], the authors tried to identify the vigilance state of the rats through the analysis of their EEG data. 32 features were extracted from the power spectral density of the EEG recordings and PCA was applied to the feature vector. Using the variance of the principal components, the three most important components were selected and used for classification.

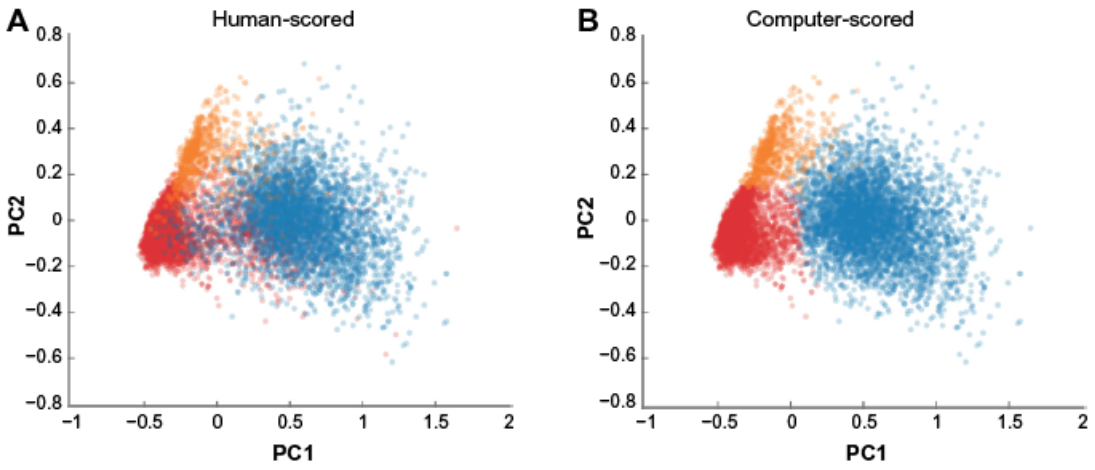

Figure 7. First two principle components of a 43-hour recording scored in 10-second epochs, (A) scored by human, (B) scored using machine learning algorithm [133].

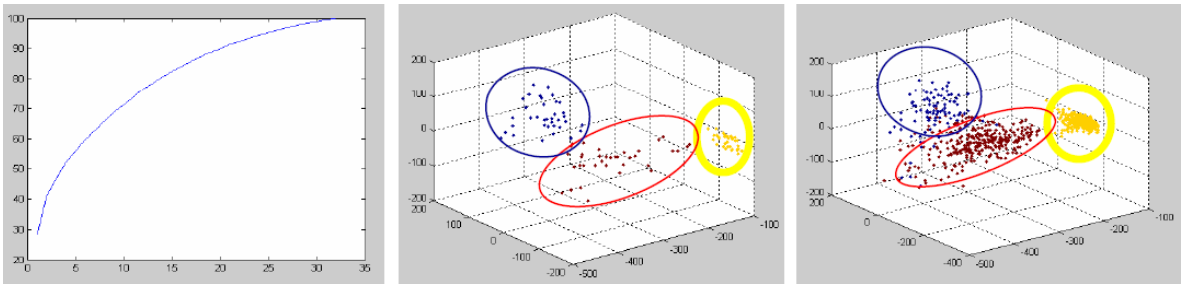

Figure 8. (Left) Percentage of variance explained by the 32 components. (Middle) Training patterns are projected into 3-dimensional subspace by PCA. (Right) Test patterns are projected into 3-dimensional subspace by PCA [134].

Figure 8 shows the variance percentage explained by each component and the training and test patterns projected into the 3-dimensional space created by PCA. According to this figure, the patterns of wake (red), REM (blue) and N3 (yellow) are separated into elliptical clusters and wake and REM stages overlap in the data patterns.

Kernel Dimensionality Reduction (KDR): To the best of our knowledge, excluding PCA, KDR is the only dimensionality reduction algorithm used in the context of sleep stage classification. Given a classification problem in which the goal is to predict *Y* from the feature vector X, KDR treats the

problem by finding a low-dimensional space called “effective space” in which the statistical relationship between X and *Y* are preserved. In this method, no assumptions are made regarding the probability distribution of X or conditional probability distribution of *Y* and X. KDR is based on a particular class of operators on Reproducing Kernel Hilbert Spaces (RKHS) [135].

In [60], four time domain and five frequency domain features were extracted from ECG signal of 16 healthy subjects. The performance of KDR is assessed comparing the classification performance with and without dimensionality reduction. To determine the effective dimension in this study, the original feature vector dimension was reduced from seven to 2, 3 and 4. Simulation results showed that the performance of KDR depends on the classifier used for sleep scoring. The classification accuracy decreased when applying the k-Nearest Neighbour (kNN) and the random forest classifier on the data reduced by KDR. On the other hand, KDR with effective dimension of 2 and Support Vector Machine (SVM) classifier implementation led to an improvement in the classification accuracy.

3.3.2 Feature Selection Methods

Unlike dimensionality reduction methods based on projection or those based on compression, feature selection methods don’t make any changes in the original features. Therefore, it is possible to understand the properties of data by analysing the features [136]. Several different types of feature selection methods exist in the literature. Among them, the most common methods are divided into three main categories: filter methods, wrapper methods and embedded methods.

Filter methods perform feature selection by considering some intrinsic characteristics of the data, usually providing a rank and/or a score for each feature. Low-rank or low scored features are removed experimentally or

according to a user defined threshold. Filter methods offer simple and fast feature ranking independent of the classifier. Wrapper methods, on the other hand, embed a search method in the space of possible feature subsets. Various subsets are produced and evaluated by training and testing with the specific classification algorithms. Since the number of possible subsets grows exponentially with the number of features, heuristic search algorithms are used for finding optimal feature subsets. With higher computational complexity and risk of overfitting, the main benefits of wrapper methods over filter methods are considering feature dependencies as well as interaction between the selected subsets and the specific classification method. Embedded methods integrate the optimal feature subset selection with the classification algorithm with less computational complexity compared to wrapper methods. The results of both wrapper methods and embedded methods are classifier-specific [136].

Filter methods: In sleep stage classification, filter methods are more common than wrapper or embedded methods. Among filter methods, Fast Correlation Based Filter (FCBF), Fisher Score, ReliefF, Chi-square, Information Gain (IG), Conditional Mutual Information Maximization (CMIM) minimum Redundancy Maximum Relevance (mRMR) algorithms [25], [69] and R-square [87] are the most common.

mRMR is a feature selection method which selects a subset of features with maximum relevance with the target class and, at the same time, minimum redundancy between the selected features [137]. In [69], automatic sleep/wake detection and multi-class sleep classification algorithms were designed using six EEG and two EOG channels. Several temporal, nonlinear and spectral features were extracted from these signals and a large feature vector was created. To reduce the number of features, the mRMR method was applied. Figure 9 shows the structure of this system.

The extracted feature types and corresponding number of selected features are shown in Table 3. Most of the relevant features are extracted from Maximum Overlap Discrete Wavelet Transform (MODWT) coefficients (such as energy, mean and standard deviation (47 features)) and harmonic parameters (39 features) and the least effective features were Kurtosis, Renyi Entropy and Tsallis entropies and Peak-to-Peak amplitude.

In addition to using mRMR, identifying the proportion of selected features per each EEG and EOG channel is an interesting aspect of this paper. According to AASM, the recommended EEG channels for sleep scoring are F4, C4 (or alternatively C3) and O2. The same channels are found suitable in this paper according to Figure 10.

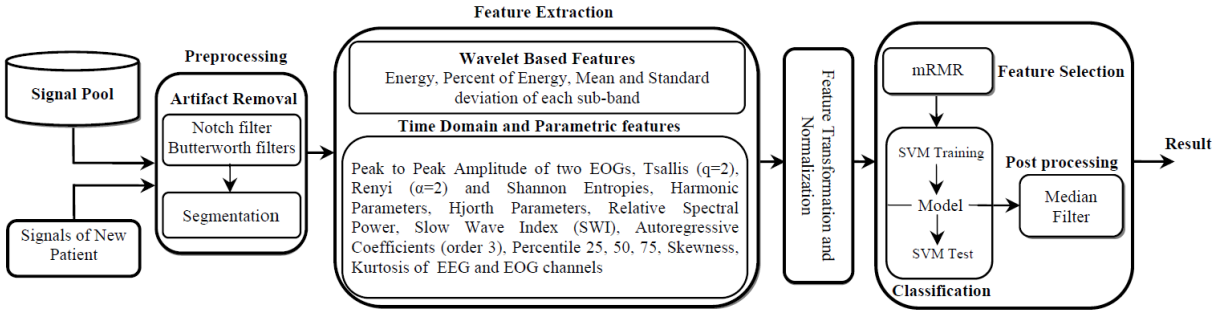

Figure 9. System structure [69]

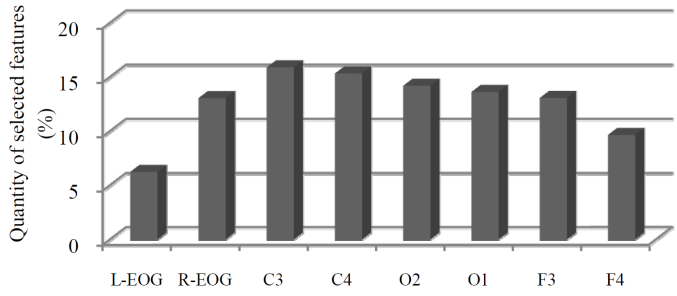

Figure 10. Proportion of selected features of each channel (EEG and EOG) in a total of 176 selected features [69].

Table 3. Extracted feature types and corresponding number of selected features [69].

Features	Selected/Total	Features	Selected/Total
MODWT Features	47/160	Skewness	2/8
Harmonic Parameters	39/120	Percentile 25, 50, 75	1/24
Relative Power	32/40	Kurtosis	0/8
Spectral Analysis	26/104	Renyi Entropy	0/8
Hjorth Parameters	14/24	Tsallis Entropy	0/8
AR coefficients	10/48	Peak to Peak amplitude	0/2
Shannon Entropy	5/8	-	-

[25] and [104] are two other papers that used mRMR for feature selection. In both papers, the performance of different feature selection methods was compared. In [104], the features selected by mRMR showed the best performance form the accuracy point of view, while in [25], mRMR with 37 selected features had the second best performance after Fisher score with 12 selected features.

A new filter method called 'Mahal' is proposed in [79]. According to the authors, the main motivation for proposing this method was the challenge of feature selection in small datasets with many features. In this paper, Mahal method is described as suitable for classifiers that are sensitive to the dimension of feature vector like LDA. Maximum class discrimination and minimum correlation were the design criteria of Mahal method. Inter-class distance and correlation were measured by Mahalanobis distance and Spearman's ranked-order correlation. The performance of Mahal was compared with Sequential Forward Search (SFS) that is a wrapper method. The simulation results show that the Mahal method selected on average 10.33 features, nearly half of the 21 features selected by SFS, with a small

difference in the classification accuracy. Although authors propose Mahal as an adequate method for small datasets with a large number of features, still it should be justified why authors did not use a conventional filter method. In case Mahal is comparable with other filter methods, a comparative study seems necessary.

Wrapper Methods: Sequential feature selection algorithms including SFS and Sequential Backward Selection (SBS) are the most common wrapper methods used in automatic sleep stage classification. Chapotot et al. in [76] tried to improve the applicability of automatic sleep scoring through the design of a formal classification framework to 1) select robust feature set, 2) follow artificial neural network classifiers, and 3) use flexible decision rules to assign sleep/wake stages. Table 4 shows the feature list used for this aim.

For selecting the best feature subset, they took advantage of the SFS algorithm that started to search the feature space with an empty set, then added features one after the other by optimizing a given criterion. Suppose *d* features are available. SFS starts by learning *d* models with one feature and selects the feature that maximizes the performance criterion.

In the second step, it tests the *d*-1 models constructed with the candidate feature selected in the first step and one of the *d*-1 remaining features. At the end, *d* subsets are available with their associated performances ($\{f_{r1}\}$, $\{f_{r1}, f_{r2}\}$ … $\{fr1, f_{r2...} f_{rd}\}$). According to the Occam's razor principle, the feature subset having the best trade-off between model dimension and performance is selected [138].

The results of the feature selection obtained by applying the SFS algorithm to the feature set of training data are illustrated in the performance curve shown in Figure 11. The optimal feature set contains five features that are: Hjorth mobility, Hjorth activity, EMG spectral edge frequency 95%, beta

relative power and sigma relative power. About the selected features, authors discussed that since Hjorth activity was computed from the raw signals acquired from recording devices, its value differed at various sampling rates and quantization scales. Therefore, the inclusion of this feature might affect adversely the robustness of the method. Considering the main objective of this work, for designing an automatic sleep stage classification framework that operates independently of the recording devices and time resolution, Hjorth activity-like features should be concerned about. Re-sampling and re-quantization to constant value can be an alternative for calculating amplitude or sampling frequency dependent features.

In another state of the art work [66], the performance of SFS and SBS methods was compared for accurate sleep stage classification. Another interesting contribution of this work was analysing the role of EOG and EMG features in improving classification performance of different stages, especially stage 1, which is a transition between sleep and wakefulness.

Table 4. Candidate features extracted for their potential independence regarding differences in PSG acquisition settings and signal conditioning [76].

Features	**Source**	**Features**	**Source**
Shannon Entropy	EEG	Theta Relative Power	EEG
Sample Entropy	EEG	Alpha Relative Power	EEG
Hjorth Activity	EEG	Sigma Relative Power	EEG
Hjorth Mobility	EEG	Beta Relative Power	EEG
Hjorth complexity	EEG	Gamma Relative Power	EEG
Hurst Exponent	EEG	Shannon Entropy	EMG
Spectral Edge Frequency 95%	EEG	Spectral Edge Frequency 95%	EMG
Delta Relative Power	EEG	Gamma Relative Power	EMG

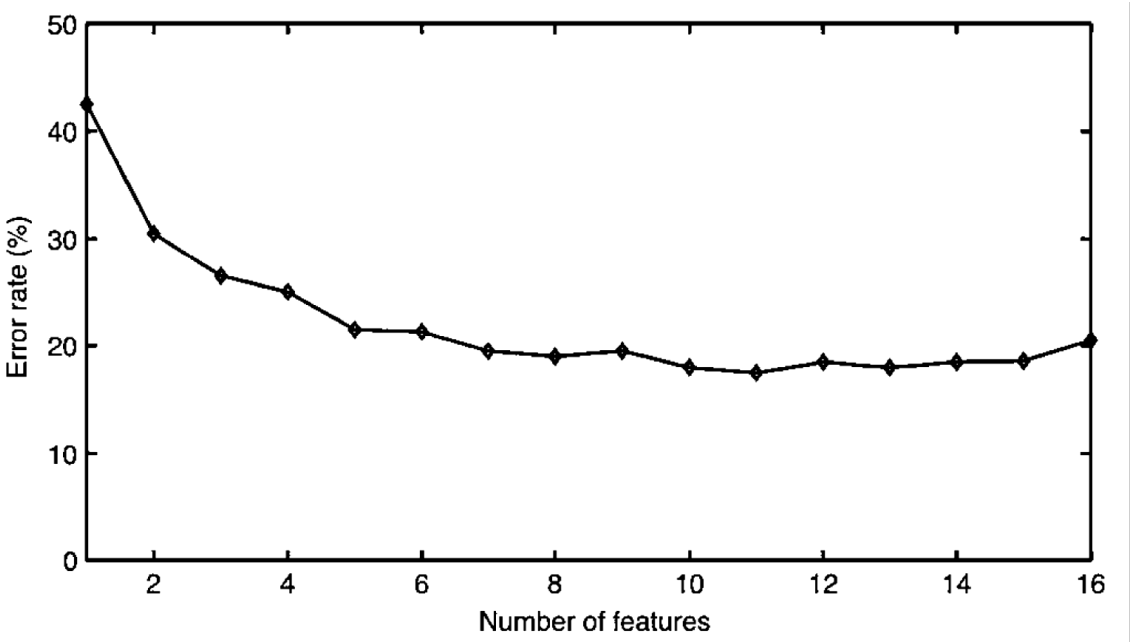

Figure 11. Performance curve resulting from SFS algorithm during candidate feature search. Progression of the classification global error is given as a function of the feature subset dimension [76].

The feature set used included 5 relative power features for describing spectral activity. These features were calculated in two ways: one by using Fourier transform and the other by using DWT coefficients. Their similarities and differences will be discussed later. In addition to spectral features, five other features were used to describe the signal in the time domain, namely, entropy, 75th percentile of the signal distribution, standard deviation, skewness and kurtosis.

The same features were used for describing the EMG and EOG signals. In addition to these features, the EMG signal was processed in the frequency domain by the relative power in high frequency band. The optimization criterion for sequential feature selection was the percentage of epochs correctly classified. Three different classifiers were used to reduce the influence of the classifier in the final accuracy. SFS and SBS methods were applied to the extracted feature set. In this feature set, DWT based features

were removed and the subset of features representing relative power of EEG in the frequency bands obtained with the Fourier transform was considered as a single feature. The SBS algorithm steps are like the SFS algorithm, except that, instead of starting with an empty feature set, the algorithm starts with the complete set of features and removes one feature in each step.

The results obtained using the SFS with the neural network classifier are shown in Figure 12. The dots show the classification accuracy while the bars express the corresponding standard deviation. Stars signal those steps where the addition of a feature generated a significant increase in the accuracy. The optimal feature set is {EEG relative power, EMG entropy, EOG entropy, EOG kurtosis, EEG 75 percentile}. According to Figure 13, the same set of optimal features was obtained using SBS.

To demonstrate the effect of EOG and EMG features on the accuracy, the percentage of correct classification for different sleep stages obtained by each feature is shown in Figure 14. It can be seen that wake, N2, REM and N3 were correctly classified by using EEG spectral information feature (with accuracy higher than 80%). The addition of new information processed from the EMG and EOG, improved the percentage accuracy of N1, where it is hard to discriminate from REM only by EEG spectral features.

About the ability of DWT compared to Fourier transform in processing EEG signals, the authors concluded that their results were quite similar, and the best accuracy was achieved when the relative EEG powers were calculated using Fourier transform and classified using a neural network.

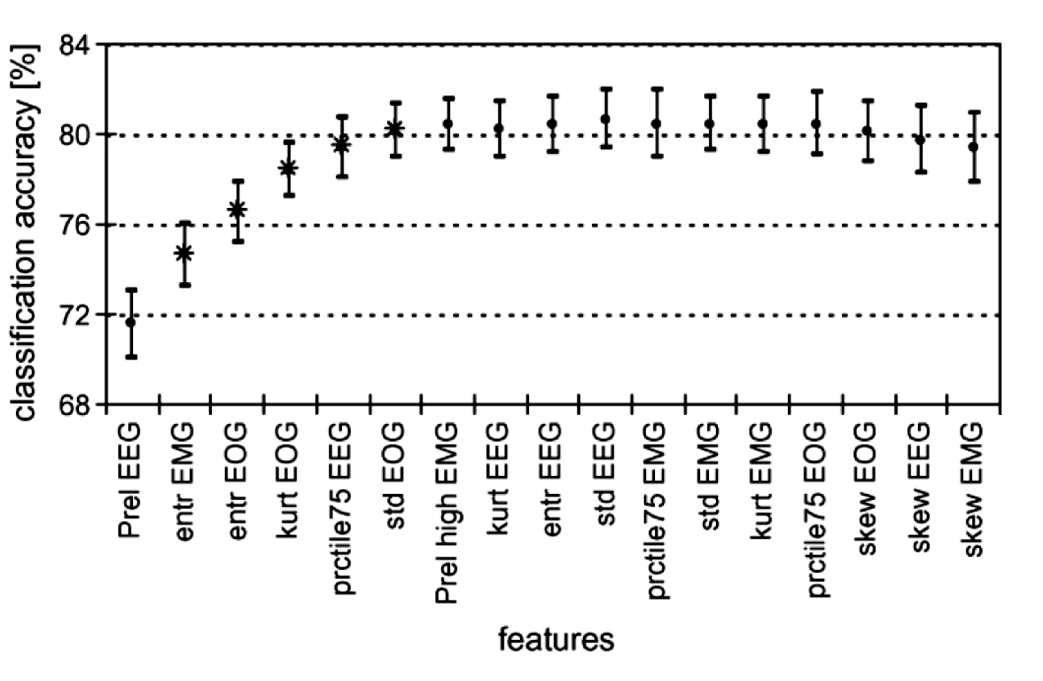

Figure 12. Selection of features by SFS performed by the neural network classifier [66].

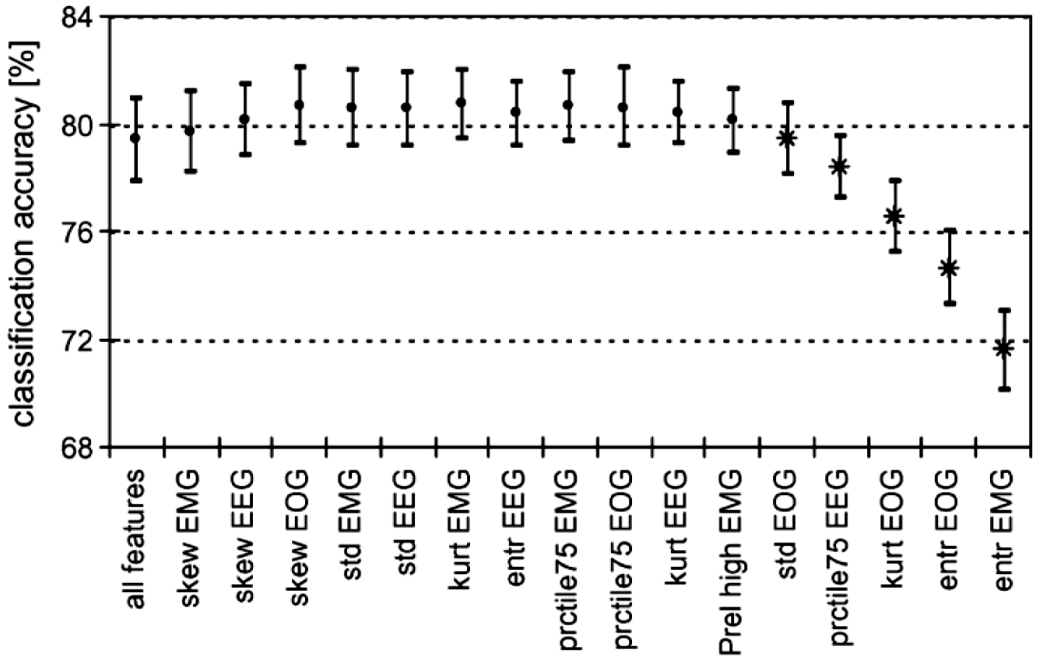

Figure 13. Selection of features by SBS performed by the neural network classifier [66].

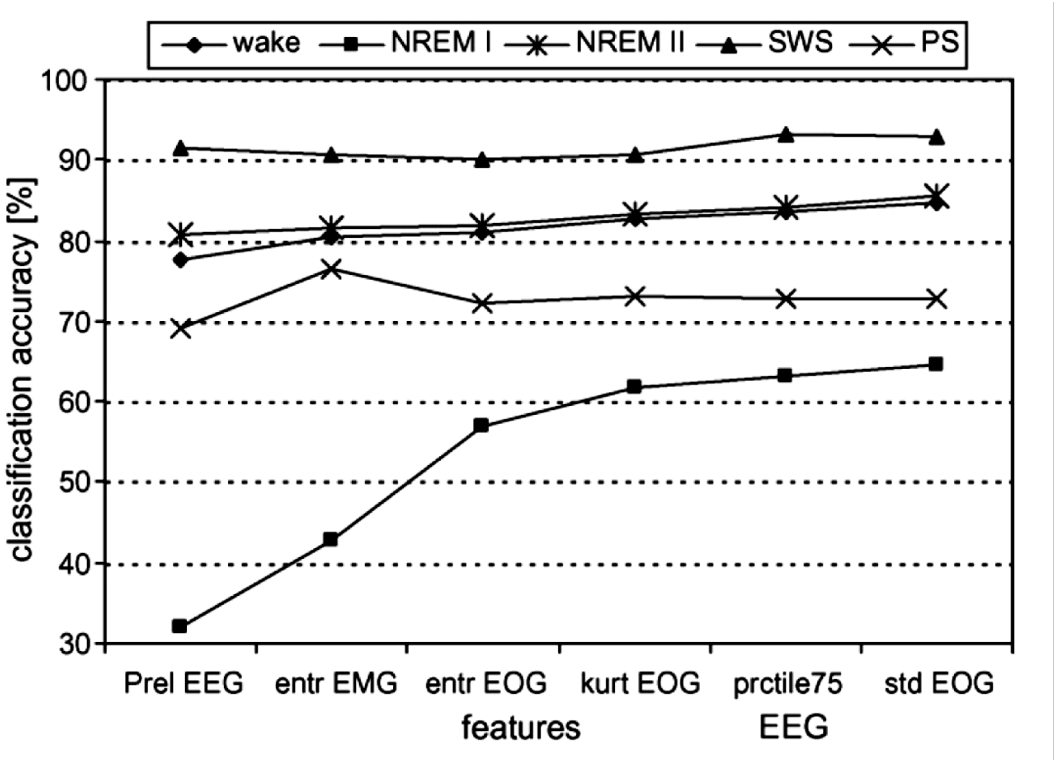

Figure 14. Classification accuracy of each sleep/wake stage obtained at each step of SFS [66].

3.3.3 Statistical Hypothesis Testing Methods

Statistical hypothesis testing methods play an important role in the dimensionality reduction and feature selection steps of classification. In sleep stage classification, these methods are used for three different purposes:

1. Dimensionality reduction,
2. Feature selection,
3. Assessment of the selected feature set's discriminatory capability.

In [67], Lajnef et al. performed a three-step feature selection process for sleep stage classification. Once all features including temporal, nonlinear and spectral features were extracted, first they rejected the outliers (features with values two times higher than the standard deviation of all values of the same feature in the same class). Second, they applied *t*-test for reducing the

dimension of the feature space. Then they ran *t*-test to compare the mean of each feature across all pairs among the five sleep stages. Finally, after removing the least discriminant features, they selected the most relevant ones using SFS. *t*-test is a widely used univariate statistical approach which determines if the means of two groups differ statistically. The probability of null hypothesis (the means of two groups don't differ significantly) is expressed in terms of *p*-value. The lower the *p*-value, the more significant is the difference. Usually a predefined level (α-value) is considered for this comparison.

In another work, Sen et al. [25] used the *t*-test approach for feature selection. If one simply runs the *t*-test on the features and ranks them according to the *p*-values, the most 'powerful' features for a classification task can be found.

In the work by Hassan et al. [53], non-parametric Kruskal–Wallis one-way analysis of variance test was used to ascertain whether the discriminatory capability of the selected features was statistically significant. Kruksal-Wallis test is the non-parametric version of one-way analysis of variance (ANOVA). ANOVA test is used to compare means of three or more groups. Unlike ANOVA, Kruksal-Wallis test doesn't assume normal distribution of data samples.

In a different work, Gunes et al. [34] reduced the feature dimension from 129 down to 4 by using statistical operators. First, they segmented each epoch to 129 overlapping segments. Then, they extracted 129 features using the average Welch spectral analysis method. To reduce the dimension of the feature space, the statistical measures including minimum value, maximum value, mean value and standard deviation were used.

3.4 Feature Post Processing

The physiological differences from subject to subject and equipment related variations have considerable impact on the features extracted from PSG recordings. Moreover, since usually there is a wide variety of feature types extracted for characterizing sleep stages, the amplitude and unit of features may also vary. The features may also get extremely low or extremely high values. Data post-processing is an important step in this respect. The aim of feature post-processing is to enable classification algorithms to uniformly handle the features with different units and ranges as well as reducing the influence of extreme values. Feature post-processing can be a feature scaling (normalization/standardization) or a feature transformation operation.

Feature standardization refers to rescaling the features, so that they have zero mean and unit variance. On the other hand, feature normalization refers to scaling the features to a predefined range such as [0 1] or [-1 1]. Feature transformation differs from standardization and normalization in the sense that the goal of transforming features is to reduce the impact of extreme values that, in some cases, even with standardization, are still a problem. In [139], a useful logarithmic transformation was proposed for obtaining normally distributed spectral features for EEG. Later, Becq et al. [77] proposed a set of transformations including $\log x, \log(1+x), \sqrt{x}, \sqrt[3]{x}, \log\frac{x}{1-x}, \frac{1}{\sqrt{x}}, \arcsin\sqrt{x}$ with the aim of transforming the features towards normal distribution in sleep stage classification. These transformations were reported to be very useful by several researchers [64], [66], [68], [80].

Usually, feature scaling (normalization/standardization) follows the feature transformation step. However, some researchers don't always use both

feature transformation and scaling. For example, feature scaling is considered enough in [61], [69].

There are some important questions regarding feature post-processing that need to be answered before using it such as: is feature post-processing always essential? What is the effect of this step on the consecutive classification step? What are the different feature post-processing algorithms? Which algorithm is proper for a specific problem at hand? In the following we will try to answer these questions.

Basically, feature scaling is necessary when the dimensionality reduction, the feature selection or the classification algorithms to be used are sensitive to the variations in the range of the features. This sensitivity can be related with the nature of the dimensionality reduction algorithm, the classifier's objective function or the metric function that is used.

PCA is a dimensionality reduction algorithm in which feature scaling plays an important role. PCA aims to find the directions of maximum data variance under the orthogonality restriction. Through feature scaling (specifically standardization) equal importance is assigned to different features so that the PCA algorithm is not tricked by the features with higher variance. In addition to PCA, some of the most common classification algorithms such as kNN, SVM and neural network classifiers need feature scaling. The Gradient Descent algorithm is often used as an optimization algorithm in SVM, perceptron and neural networks. Feature standardization will give better error surface shape (round counters instead of highly skewed elliptic ones), preventing from getting stuck in local minima and helping weight decay to be conveniently done. The kNN classifier typically uses the Euclidean distance to measure the distance between two points. If one feature has broader range, the distance will be greatly affected by this feature. In contrast, tree-based methods are scale-invariant and don't need standardized features.

3.5 Summary

This chapter addressed a literature review on topics related to this thesis research work, the topic of sleep stage classification, with special relevance on feature extraction and selection. When relating all the different existing features and selection techniques in the literature, it is noticeable that deeper research work is required in sleep stage classification to apply these methods as a reliable tool in clinical environments. In particular, deeper research is essential regarding the strategy of constructing the PSG feature vector to address the existing challenges. Some of these challenges are related to the reliability and stability of feature vectors. A specific feature vector should be stable enough to provide consistent quality when extracted from different subjects and datasets. This issue seems to be overlooked in the literature. Moreover, considering that the quality of raw signal has significant impact on the feature vector quality as well as final classification performance, effective and loss less methods should be developed to enhance the signal quality.

Chapter 4

4. Data and Methods

In this chapter, we describe data and methods used to achieve the goals of this thesis. First, the data sets used for evaluation of the proposed methods will be presented. Then, the methods applied for pre-processing, feature extraction and selection, classification and feature assessment will be described.

4.1. Database

For evaluation of the sleep stage classification system's performance annotated data is essential. Since in this research work the goal was using mainly supervised classification to evaluate the developed feature extraction and selection methods, PSG signals and the corresponding hypnograms were required. In this work, two different open access databases were considered, namely The Sleep-EDF database [Expanded], Physionet [140] and ISRUC-sleep dataset [141].

4.1.1. The Sleep-EDF database [Expanded], Physionet

The collection of 61 PSG recordings with the corresponding hypnograms in The Sleep-EDF database [Expanded] were acquired from two different sleep studies. PSG recordings of the first study were named SC files (SC=Sleep Cassette) recorded in 1987-1991 and PSG recordings of the

second study were named ST files (ST=Sleep Telemetry) recorded in 1994. All recordings were obtained from whole night sleeps containing EEG (from Fpz-Cz and Pz-Oz channels), horizontal EOG, and submental chin EMG. The signals were sampled at 100 Hz. The data was segmented into 30-second epochs and all epochs were scored according to R&K guidelines [142] for human sleep staging into six sleep stages.

Since EMG data for first study was a zero-amplitude or no data recording, in our evaluations we used ST files which were a collection of PSG signals from 22 Caucasian male and female subjects recorded in the hospital during two nights for about 9 hours. Except for a slight difficulty in falling asleep, subjects were healthy without any sleep related medication.

Through careful analysis of ST recordings, a number of issues were detected that made some of the recordings unsuitable for being used in the evaluations. These issues were as follows:

- Lack of stage 4 (according to R&K guidelines),
- Artefacts such as severe movement or sensor misconnection,
- Unsynchronized EEG data and hypnogram,
- Lack of stage 3 epochs,
- Severely corrupted EEG data.

Therefore, six recordings were selected out of twenty-two and the corresponding hypnograms were converted from R&K to AASM. Table 5 illustrates the number of stages available per subject.

Table 5. Summary of the data provided by six selected subjects in The Sleep-EDF database [Expanded], Physionet.

	Wake	**REM**	**N1**	**N2**	**N3**
Subject #1	146	122	101	527	136
Subject #2	41	159	71	351	284
Subject #3	85	226	120	392	180
Subject #4	40	143	47	266	152
Subject #5	149	80	102	428	218
Subject #6	131	142	135	378	198

4.1.2. ISRUC sleep database

ISRUC-Sleep database is an open-access comprehensive database that includes data from healthy subjects, subjects with sleep disorders and subjects under the effect of sleep medication. PSG recording was performed using a bio-signal acquisition equipment namely, SomnoStar Pro sleep system, in the sleep medicine centre of Coimbra University Hospital (CHUC) between 2009 and 2013. The PSG signals were recorded during a whole-night of sleep (approximately eight hours) according to the recommendations of AASM. Sampling frequency was 200Hz for all EEG, EOG, chin EMG and ECG signals. After segmenting the data into 30-seconds epochs, two different experts performed manual sleep scoring using AASM.

To improve the quality of the recordings, in this database a pre-processing step was already taken by the database providers. The details of this pre-processing are as follows:

- A notch filter was applied to eliminate the 50 Hz electrical noise from EEG, EOG, chin EMG and ECG,

- EEG and EOG recordings were filtered using a bandpass Butterworth filter with a lower cut-off frequency of 0.3 Hz and higher cut-off frequency of 35 Hz, and
- EMG channels were filtered using a bandpass Butterworth filter with a lower cut-off frequency of 10 Hz and higher cut-off frequency of 70 Hz.

4.2 Methods

As mentioned is chapter 2, automatic sleep stage classification algorithms consist of four main steps, namely pre-processing, feature extraction, feature selection and classification. In the following, the methods used in this thesis for each step are described.

4.2.1 Pre-processing

In this thesis, PSG recordings were examined carefully both from quality and agreement with AASM points of view. Thus, when necessary, three types of pre-processing operations were done before feature extraction stage with the aim of enhancing the quality of signals and synchronizing with the corresponding hypnogram. These operations include:

Band pass filtering: AASM manual recommends a filtering interval for each one of PSG recordings to remove the unnecessary waves and oscillations. For example, for EEG and EOG the preferred frequency band is 0.3-35 Hz, and for EMG 10-100 Hz is recommended. In this thesis, for filtering, wavelet multi-level decomposition and reconstruction was used [143]. This filtering technique has high fidelity to the original wide-band signal in contrast to Butterworth filtering that produces a highly distorted “valley” shape.

Windowing: As mentioned before, each 30 seconds of PSG recordings is considered as an epoch and during sleep scoring one of five sleep stages is

associated with this epoch. Therefore, it is essential to window the signals to epochs and associate each of them with the corresponding hypnogram slot.

PSG trimming: PSG recordings get contaminated with several artefacts such as power line noise, electrode movements, sweating, body movements. Even, zero-energy epochs may appear due to the possible failure of the recording device. In this thesis, epochs with zero energy were identified through examination of the signal's time domain energy and removed.

4.2.2 Feature Extraction

Throughout this thesis, two main sets of features were used, namely conventional feature set and distance-based feature set. In the following the description and details of each feature set are presented.

4.2.2.1 Conventional Feature Set

Conventional feature set consists of 48 features extracted from EEG, EOG, and EMG signals. We tried to use the most common features in sleep stage classifcation to explore the information contained in these signals [25], [33]. These features can be mainly categorized into temporal, time-frequency domain, entropy-based and non-linear features. Each epoch's feature vector contains 35 EEG, 6 EOG, and 7 EMG features. Table 6 summarizes these features that were extracted from 30-second epochs along with their handy description.

All the features in this table were already described in chapter 3. For generating F13 to F26, WP analysis was selected since it provides a valuable joint time-frequency domain analysis. According to the scheme proposed in [52], a WP tree with 7 decomposition levels is sufficient to estimate the necessary frequency bands of EEG rhythms, sampled at 100 Hz, with

adequate accuracy. These bands include α, δ, β1, β2, θ and k-complexes + Delta and spindles bands. Table 7 shows the corresponding frequency range to these bands (check frequencies with chapter 2). Features F13 to F26 were extracted using the corresponding WP coefficients.

Table 6. Summary of the conventional features extracted from PSG recordings.

Ref.	Signal	Description	T*	TF*	F*	E*	NL*
F1	EEG	Arithmetic Mean	●				
F2		Maximum	●				
F3		Minimum	●				
F4		Standard Deviation	●				
F5		Variation	●				
F6		Skewness	●				
F7		Kurtosis	●				
F8		Median	●				
F9		Petrosian Fractal Dimension					●
F10		Rényi Entropy				●	
F11		Spectral Entropy				●	
F12		Permutation Entropy				●	
F13		Approximation Entropy				●	
F14		Hjorth Parameter (Activity)	●				
F15		Hjorth Parameter (Mobility)	●				
F16		Hjorth Parameter (Complexity)	●				
F17		Mean Curve Length					●
F18		Zero Crossing Number	●				
F19		Mean Energy					●
F20		Mean Teager Energy					●
F21		Hurst Exponent					●
F22		Mean Quadratic Value of WP Coefficients in Delta Band		●			
F23		Mean Quadratic Value of WP Coefficients in Theta Band		●			
F24		Mean Quadratic Value of WP Coefficients in Alpha Band		●			

F25		Mean Quadratic Value of WP Coefficients in Spindle Band		●			
F26		Mean Quadratic Value of WP Coefficients in Beta1 Band		●			
F27		Mean Quadratic Value of WP Coefficients in Beta2 Band		●			
F28		Mean Quadratic Value of WP Coefficients in All Frequency Bands		●			
F29		F24/(F22+F23)		●			
F30		F22/(F24+F23)		●			
F31		F23/(F22+F24)		●			
F32		F24/F23		●			
F33		F22/F23		●			
F34		Mean of the Absolute Values of WP Coefficients in All Bands		●			
F35		Standard Deviation of WP Coefficients in All Bands		●			
F36	EMG	Spectral Power			●		
F37		Maximum of the Spectral Power Distribution			●		
F38		Mean of the Spectral Power Distribution			●		
F39		Standard Deviation of the Spectral Power Distribution			●		
F40		Temporal Energy					●
F41		Ratio of the Temporal Energy of Current Epoch to The Energy of Previous Epoch					●
F42		Ratio of the Temporal Energy of Current Epoch to the Energy of Next Epoch					●
F43	EOG	Mean	●				
F44		Energy					●
F45		Maximum	●				
F46		Standard Deviation	●				
F47		Skewness	●				
F48		Kurtosis	●				

* (T: Temporal, TF: Time-Frequency, F: Frequency, E: Entropy, NL: Non-Linear)

Table 7. EEG frequency bands used in time-frequency features of conventional feature set.

Frequency Band Name	**Frequency Range (Hz)**
k-complexes + Delta	0.4-1.55
Delta (δ)	1.55-3.2
Theta (θ)	3.2-8.6
Alpha (α)	8.6-11
Spindle	11-15.6
β1	15.6-22
β2	22-37.5

4.2.2.2 Distance-based Feature Set

As mentioned in chapter 2, feature vector quality is an important factor for the development of a reliable classification system. Features used in a specific machine learning problem can perform reasonably well in other problems as well. Therefore, researchers often evaluate and explore the applicability of various features in different machine learning areas. Kong et al. in [144] assumed that EEG signals can be modelled as an AR process and used Itakura distance to measure the similarity of the EEG signals. The Itakura distance is a very popular distance measure in speech signal processing. Nevertheless, it has been found effective in distinguishing hypoxia and asphyxia. Later in 2004, Estrada et al. [128] used the Itakura distance for measuring similarity of a baseline EEG epoch to the rest of the EEG in the context of sleep stage classification. In addition to the similarity of EEG signal with itself, in [70], [72] it is demonstrated that the Itakura distance between EEG and EOG is also a useful similarity measure for sleep stage classification.

Suppose $x(t)$ is the baseline epoch and $y(t)$ is an epoch from the rest of the signal. If we model $x(t)$ and $y(t)$ as AR processes with order *p*, then the

vectors $\mathbf{a}_x$ and $\mathbf{a}_y$ would contain the AR coefficients, respectively. Itakura distance of a baseline epoch with others is calculated as:

$$D_I = \ln\left(\frac{\mathbf{a}_y^T \mathbf{R}_x(p)\mathbf{a}_y}{\mathbf{a}_x^T \mathbf{R}_x(p)\mathbf{a}_x}\right) \tag{1}$$

where $\mathbf{R}_x(p)$ and $\mathbf{R}_y(p)$ are the autocorrelation matrixes of $x(t)$and $y(t)$ with size p + 1, respectively. Itakura distance, defined in this way, is asymmetric, i.e. D_I of $x(t)$ and $y(t)$ is not equal to $D_I$of $y(t)$ and $x(t)$[129]. In order to add symmetry to this measure, the mean of these two distances is usually calculated, as follows [128]:

$$D_I = \frac{1}{2}\left(\ln\left(\frac{\mathbf{a}_y^T \mathbf{R}_x(p)\mathbf{a}_y}{\mathbf{a}_x^T \mathbf{R}_x(p)\mathbf{a}_x}\right) + \ln\left(\frac{\mathbf{a}_x^T \mathbf{R}_y(p)\mathbf{a}_x}{\mathbf{a}_y^T \mathbf{R}_y(p)\mathbf{a}_y}\right)\right) \tag{2}$$

In addition to AR coefficients, the distance between spectral representations of the signals can be used to measure similarity [129]. Suppose $S_x(\omega)$ and $S_y(\omega)$ are the power spectra of $x(t)$ and $y(t)$. The Itakura distance between these two spectra, in its asymmetric form, is calculated as:

$$D_I(X,Y) = \ln\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S_x(\omega)}{S_y(\omega)}d\omega\right] \; D_I(X,Y) = \ln\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S_x(\omega)}{S_y(\omega)}d\omega\right] \tag{3}$$

The same averaging (Equation (2)) can be applied for adding symmetry property to this distance. Along with Itakura distance, there are two other distance measures that are common in speech processing, namely Itakura-Saito and COSH distances [145]. Following the definitions of variables made for Itakura distance, Itakura-Saito distance is calculated as:

$$D_{IS}(X,Y)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S_x(\omega)}{S_y(\omega)}-\ln\frac{S_x(\omega)}{S_y(\omega)}-1\right]d\omega \tag{4}$$

COSH distance is the symmetrical version of Itakura-Saito distance and is calculated as:

$$\begin{aligned} D_{Cosh} &= \frac{1}{2}\left(D_{IS}(x,y)+D_{IS}(y,x)\right) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{S_x(\omega)}{S_y(\omega)}+\frac{S_y(\omega)}{S_x(\omega)}-2\right)d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}2\cosh\left(\ln\frac{S_x(\omega)}{S_y(\omega)}-1\right)d\omega \end{aligned} \tag{5}$$

where $\cosh(x)=\frac{e^{x}+e^{-x}}{2}$ is the hyperbolic cosine function. Like Itakura distance, Itakura-Saito and COSH distances can be calculated using AR coefficients as well.

Considering the previous work in this area, in this thesis a set of 32 distance-based features, was used for sleep stage classification as summarized in Table 8. Two types of distance-based features were considered: features measuring the similarity of a baseline epoch of a signal with other epochs of the same signal and features measuring the similarity of a baseline epoch of a signal with the epochs of another signal. Except for three features (F49, F51, F65), the remaining features have not been used in sleep stage classification before [70], [72], [144]. For calculating F49 to F52 and F73 to F74, the wake EEG epoch was considered as the baseline. The same applies for features F53 to F64 and F75 to F80 corresponding to EMG, EOG, and ECG signals. For calculating F65 to F72, wake EEG epoch was considered as the baseline, and the distance was found between EEG-EOG, EEG-EMG, and EEG-ECG. We used VOICEBOX, a MATLAB speech processing toolbox [146], consisting of MATLAB routines that are mostly written and maintained by Mike Brookes from department of electrical &

electronic engineering, Imperial College, UK. We used the routines for calculating Itakura, Itakura-Saito and COSH distances from this toolbox.

Table 8. Summary of distance-based features extracted from PSG recordings.

Ref.	Signal	Description
F49	EEG	Itakura Distance of AR Coefficients
F50		Itakura Distance of Spectral Coefficients
F51		Itakura-Saito Distance of AR Coefficients
F52		Itakura-Saito Distance of Spectral Coefficients
F53	EMG	Itakura Distance of AR Coefficients
F54		Itakura Distance of Spectral Coefficients
F55		Itakura-Saito Distance of AR Coefficients
F56		Itakura-Saito Distance of Spectral Coefficients
F57	EOG	Itakura Distance of AR Coefficients
F58		Itakura Distance of Spectral Coefficients
F59		Itakura-Saito Distance of AR Coefficients
F60		Itakura-Saito Distance of Spectral Coefficients
F61	ECG	Itakura Distance of AR Coefficients
F62		Itakura Distance of Spectral Coefficients
F63		Itakura-Saito Distance of AR Coefficients
F64		Itakura-Saito Distance of Spectral Coefficients
F65	EEG & EOG	Itakura Distance of AR Coefficients,
F66		Itakura Distance of Spectral Coefficients
F67		Itakura-Saito Distance of AR Coefficients
F68		Itakura-Saito Distance of Spectral Coefficients
F69	EEG & EMG	Itakura Distance of AR Coefficients
F70		Itakura Distance of Spectral Coefficients
F71		Itakura-Saito Distance of AR Coefficients
F72		Itakura-Saito Distance of Spectral Coefficients
F73	EEG	COSH Distance of AR Coefficients
F74		COSH Distance of Spectral Coefficients
F75	EMG	COSH Distance of AR Coefficients

F76		COSH Distance of Spectral Coefficients
F77	EOG	COSH Distance of AR Coefficients
F78		COSH Distance of Spectral Coefficients
F79	ECG	COSH Distance of AR Coefficients
F80		COSH Distance of Spectral Coefficients

4.2.3 Feature Post-processing

The features extracted from PSG signals are in different ranges and this variety can bias the results of the subsequent steps. Feature scaling methods are utilized for avoiding this bias. In this thesis, two different types of scaling methods were used: standardization (or Z-score normalization) and Min-Max scaling.

4.2.3.1 Standardization

This rescaling is necessary for many machine learning algorithms. Each feature (x_{ij}) is independently scaled to have zero mean and unit variance (x'_{ij}) using the following equation:

$$x'_{ij} = \frac{x_{ij} - \overline{\mathbf{x}}_i}{\sigma_{\mathbf{x_i}}} \tag{6}$$

where $\overline{\mathbf{x}}_i$ and $\sigma_{\mathbf{x}_i}$ are the mean and the standard deviation of each independent feature vector.

4.2.3.2 Min-Max Normalization

In Min-Max normalization, features are scaled to the fixed range of [0 1]. Suppose $x_{\min}$ and $x_{\max}$ are the minimum and maximum of feature vector X. The values of this feature vector are normalized according to the following equation:

$$x'_{ij} = \frac{x_{ij} - x_{i\min}}{x_{i\max} - x_{i\min}} \quad (7)$$

4.2.4 Feature Similarity Reduction

In order to remove features with high levels of similarity, a feature selection method was proposed in this thesis. This method works as follows:

First, the L1-norm between each pair of feature vectors is calculated, then considering the range of the extracted L1-norm, a similarity threshold is defined. The feature pair whose L1-norm is less than the threshold level is considered strongly similar. In this way, the features are clustered into groups of similar ones and one feature per cluster is selected as representative. The representative feature has the lowest computational complexity. Alternatively, it is possible to use Principal Component Analysis (PCA) for finding the most dissimilar features. However, there are two main reasons that we didn't use PCA. First, using PCA for finding a non-redundant feature set would lead to keeping and calculating all the features in the classification and practical application steps while by using the similarity threshold the most redundant features can be detected and omitted from feature set in the application step. Second, PCA would generate combinations of the features. Since in this thesis the aim is to evaluate individual features without combining them, it is necessary to preserve the information on the features and PCA is not suitable in this regard.

4.2.5 Feature Selection

In this thesis, to select a subset of features containing most of the original feature set information, seven different feature ranking methods were used namely, ReliefF, minimum Redundancy-Maximum Relevance (mRMR-MID

and mRMR-MIQ), Fisher Score, Chi-Square, Information Gain (IG) and Conditional Mutual Information Maximization (CMIM).

4.2.5.1 Feature Ranking Methods

- *ReliefF*: In 1992, Kira and Rendell [147] proposed Relief, an instance based method, for estimating the quality of features. In this method for a randomly selected sample two nearest neighbours were considered: one from the same class (nearest hit) and another from a different class (nearest miss). The quality estimation value for each feature is updated according to the randomly selected sample's distance from the nearest hit and miss. The Relief method is restricted to two-class problems and is highly sensitive to noisy and incomplete data. An extension of Relief, called ReliefF [148], was proposed improving the original method by estimating the probabilities more reliably and extending the algorithm to multi-class problems. The ReliefF algorithm uses k-nearest hits and k-nearest misses for updating the quality estimation for each feature.
- *minimum Redundancy-Maximum Relevance (mRMR)*: MRMR [149] is a feature selection method which selects a subset of features with maximum relevance for the target class and at the same time minimum redundancy between the selected features. In MRMR method the redundancy (R) and relevance (D) are expressed in terms of mutual information. To select the final feature set, an objective function $\varphi(D,R)$ is maximized. The $\varphi(D,R)$ can be defined either as the mutual information difference (MID), D-R, or the mutual information quotient (MIQ), D/R.

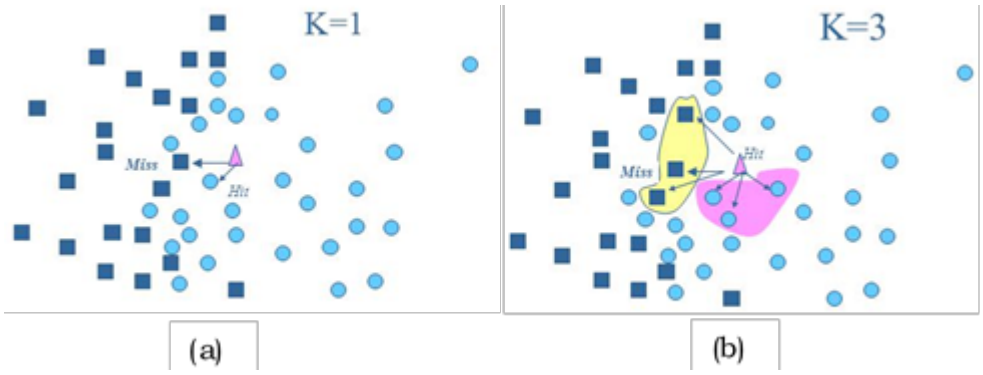

Figure 15. (a) Relief feature selection method, (b) ReliefF feature selection with K=3 [150].

- *Fisher Score*: This method is one of the most efficient and most widely used feature ranking methods. The key idea is to find a subset of features with maximum distance between the data points from different classes and minimum distance between data points of the same class in the feature space [151].
- *Chi-square*: Chi-square is another very common class sensitive feature selection method which ranks the features according to their Chi statistics without considering the interactions between features. Originally proposed for categorical data, this method was later extended to the continuous data [152]. For calculating Chi-square statistics of each feature, the range of the numerical feature should be discretized into intervals.
- *Information Gain (IG)*: Ross Quinlan proposed an algorithm for generating decision trees from a set of training data [153]. In this algorithm, IG is the measure for selecting the effective feature at each node. Generally, IG can be described as the change in the marginal entropy of a feature set considering the conditional entropy of that feature set with the given class set.

- *Conditional Mutual Information Maximization (CMIM)*: This method [154] is based on mutual information in such a way that all the selected features are informative and have two-by-two weak dependency. A feature is added to the selected feature set if it carries information about the specific class and this information is not caught by any other previously selected feature.

4.2.5.2 Rank Aggregation Methods

In many machine learning problems, performing a single round of feature selection can give unstable results which are sensitive to small changes in the input data. New techniques are required to reliably select features in a consistent manner. One of the more promising methods for resolving this problem is ensemble feature selection. In general, an ensemble feature selection technique takes the results of multiple feature and aggregates the resulting ranked feature lists into a single ranked list. Therefore, more robust and global feature subsets are generated which are as good as (if not better than) the feature subsets created by individual feature ranking methods [155].

There are several ways to aggregate feature ranking methods [156]. In this thesis, we have implemented two different rank aggregation methods namely, Borda and Robust Rank Aggregation (RRA), to evaluate the ability of these methods to produce better feature rankings compared to the conventional feature ranking methods. A brief description of the used rank aggregation methods is provided below.

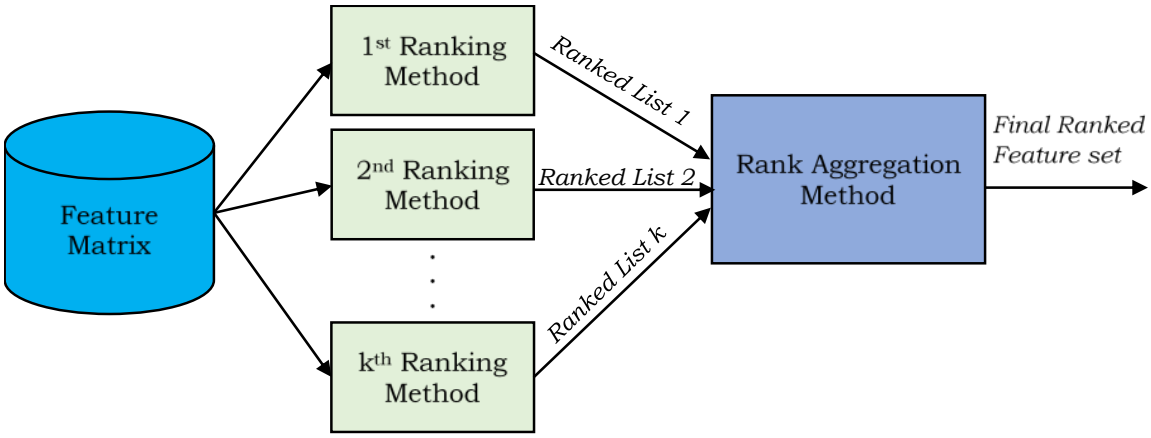

Figure 16. Block diagram of feature rank aggregation method. k is the number of ranking techniques.

- *Borda*: Borda methods ranks each feature based on its mean position in the different ranking methods, i.e.

$$Borda(f_i) = \sum_{j=1}^{N} \pi_j(f_i) \tag{8}$$

where $\pi_j(f_i)$ is the rank of the feature f_i in the ranking method π_j.The feature with the highest Borda rank is considered the best [156].

- *Robust Rank Aggregation (RRA)*: This method, proposed by Kolde et al. [157], compares the results from several feature ranking methods with a randomly ranked feature list. The RRA first looks how a specific feature is ranked by the various methods and lists the corresponding values in a so-called rank order, from best to worst. Then, the probability of a random list producing better ranking than the values seen in the actual rank order for that specific feature is determined. The features with the smaller probability are selected as the better ones [155].

4.2.5.3 Stacked Sparse AutoEncoder (SSAE)

An autoencoder is a special type of neural network whose output values are equal to the inputs. Typically, it consists of an encoder and a decoder and it is trained in an unsupervised manner using backpropagation. During training, a cost function that measures the error between the input and output of the autoencoder is optimized. In other words, the autoencoder tries to learn the identity function (Figure 19). By applying special constraints on the network such as the number of hidden units, an autoencoder can learn new representation or coding of the data [158].

Suppose the input vector to the autoencoder is a set of un-labelled data $\mathbf{x} \in \sim^{D_x}$. This vector is encoded to another vector $\mathbf{z} \in \sim^{D_1}$ in the hidden layer as follows:

$$\mathbf{z} = h^1\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right) \tag{9}$$

where h^1 is the transfer function of the encoder, $\mathbf{W}^1$ is the weight matrix and $\mathbf{b}^1$ is the bias vector of the encoder. Then, the autoencoder tries to decode this new representation back to the original input vector as follows:

$$\hat{\mathbf{x}} = h^2\mathbf{z} = h^2\left(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1\right) \tag{10}$$

where h^2 is the transfer function of decoder, $\mathbf{W}^2$ is weight matrix and $\mathbf{b}^2$ is the bias vector of the decoder. Sparse autoencoder is a specific type of autoencoder in which to encourage the sparsity of the output of the hidden layer, a constraint is imposed on the number of active hidden neurons. The cost function of the sparse autoencoder is slightly different from the original autoencoder as follows:

$$E = \underbrace{\frac{1}{N}\sum\sum\left(\mathbf{x} - \hat{\mathbf{x}}\right)^2}_{\text{mean squared error}} + \underbrace{\lambda\,\Omega_{weights}}_{\text{weight regularization}} + \underbrace{\beta\,\Omega_{sparsity}}_{\text{sparsity regularization}} \tag{11}$$

where N is length of the input vector, λ is the weight regularization parameter β is the sparsity regularization parameter [159].

A Stacked Sparse Autoencoder (SSAE) is a neural network with several sparse autoencoders. In this architecture, the output of each autoencoder is fully connected to the inputs of the next autoencoder. Greedy layer-wise training strategy is usually used for training SSAE. After the training of each layer is complete, a fine tuning is usually performed for enhancing the learned weights using the backpropagation algorithm. Fine tuning can greatly improve the performance of the stacked autoencoder [158]. Figure 20 [160] shows the training steps of a two layers stacked autoencoder. The training of this stacked autoencoder has three steps:

- Step1: initial pretraining of layer 1,
- Step 2: optimize the weights of the second layer using the weights of the first layer,
- Step 3: model fine-tuning by connecting all the layers together.

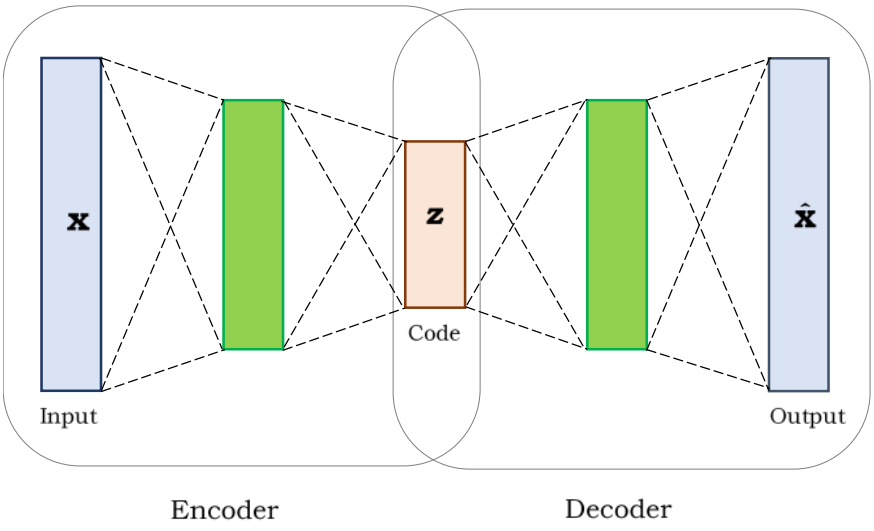

Figure 17. Schematic structure of an autoencoder with 3 fully-connected layers.

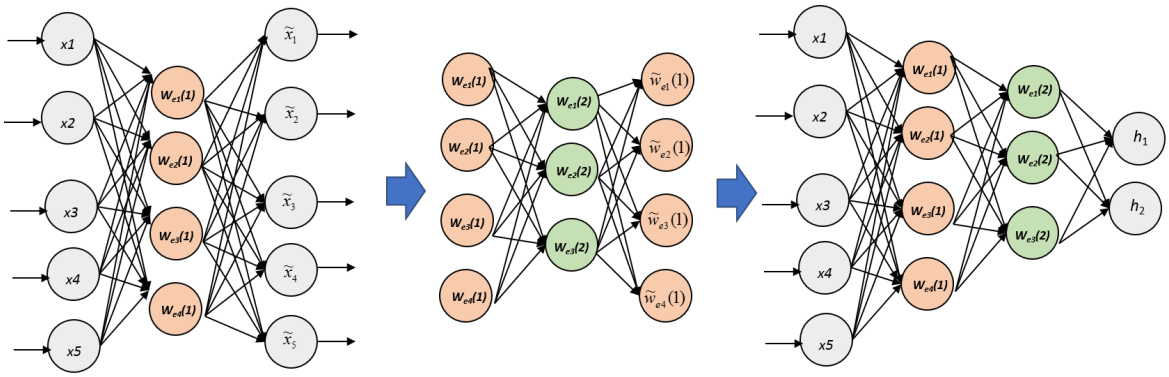

Figure 18. Training of a two-layer stacked autoencoder [160].

4.2.6 Classification

In this thesis, four types of classifiers were used for the classification of extracted feature vectors. In the following a brief description of each classifier is presented.

4.2.6.1 k-Nearest Neighbours (kNN)

kNN method is one of the most common classification techniques. It classifies an unknown sample based on the known classification of its neighbours. Suppose that a training set with a known classification is available. Intuitively, if the classification of a sample is unknown, then it could be predicted by considering the classification of its nearest neighbours. In kNN, for an unknown sample and a training set, all the distances between the unknown sample and all the samples in the training set can be computed. The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample. Therefore, the unknown sample can be categorized into the class of its nearest neighbour [161].

4.2.6.2 Multi-layer Feed-Forward Neural Network

Multi-layer feed-forward (MLF) neural network trained with backpropagation algorithm [162] is one of the most popular neural networks and were used in this thesis.

On a MLF neural network the first layer is called the input layer, the last layer is called the output layer and the layers in between are called hidden layers. Each neuron in a specific layer is fully connected to the neurons of the next layer. The strength of this connection is defined with the weight coefficient. The weighted sum of input and bias are fed to the transfer function, which usually generates a nonlinear mapping of its input. In supervised training process, the weights are varied to minimise the sum of squared errors between the computed and the desired outputs. In back propagation algorithm, the steepest descent minimisation method is used [163].

4.2.6.3 Softmax Classifier

The softmax classifier [158] is a generalization of the binary Logistic Regression classifier to multiple classes. Logistic regression is a statistical method used for predicting a binary outcome such as pass/fail, win/lose, 1/0. Softmax classifier is a model that converts the unnormalized values at the end of a linear regression to normalized probabilities for classification.

Suppose $\mathbf{x}$ is the classifier's input, $\mathbf{W}$ is the matrix of weights and $\mathbf{b}$ is the bias, the output of liner regression model $\mathbf{y}$ is calculated as follows:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{12}$$

To go from arbitrary values y_i to normalized probability estimates for each class (p_i) in a classification problem with K classes, exponentiation and normalization are used in Softmax classifiers as follows:

$$p_i = \frac{\exp(y_i)}{\sum_{k=1}^{K} \exp(y_k)} \tag{13}$$

4.2.6.4 Dendogram-based Support Vector Machine (DSVM)

Support Vector Machines (SVM) are discriminative classifiers defined by a separating hyperplane [164]. There are two types of approaches for multi-class classification using SVM classifiers, namely One-Against-All (OAA) and One-Against-One (OAO) approaches. OAA framework consists of a binary SVM to distinguish each class from all other classes and the decisions obtained from applying a winner-takes-all strategy. In contrast, in the OAO approach, a dedicated classifier is trained for each of all possible pairs of classes.

Lately, a new variation of SVM classifier was proposed which is based on decomposing of the multiclass problem to several binary classification problems [165]. First, these methods build a dendrogram of classes, according to Figure 19, and then, a binary SVM is learned for each internal node of that hierarchy in order to separate the examples of each class.

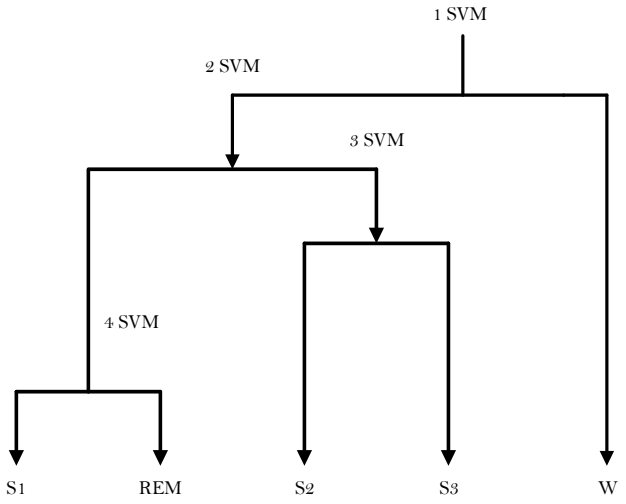

Figure 19. Dendorgam-based SVM structure.

4.2.7 Multi-Criteria Decision Making (MCDM)

In this thesis, to find the trade-off between the number of features used, and the classification accuracy, a Multi-Criteria Decision Making (MCDM) technique, called Vikor [166], [167] was used. The Vikor method was originally developed for MCDM problems with contrasting and conflicting criteria. In our case, the accuracy and number of features are two conflicting criteria. This method ranks and selects a set of alternative solutions for the problem at hand, helping decision makers to reach a final decision. The various J alternative solutions are denoted as $a_1, a_2, \ldots, a_J$. Suppose that there are n criteria, f_{ij} is the value of the i^{th} criterion for j^{th} solution, a_j. The compromise ranking is performed by comparing the closeness to the ideal solutions of the criteria (utopian solution F^*). The distance measure of the Vikor method is developed from the L_p-metric as:

$$L_{p,j} = \left\{ \sum_{i=1}^{n} \left[w_i \left(f_i^* - f_{ij} \right) / \left(f_i^* - f_i^- \right) \right]^p \right\}^{\frac{1}{p}}, \qquad (14)$$
$$1 \le p \le \infty; \quad j = 1, 2, ..., J,$$

where f_i^* and f_i^- are the best and worst solutions of the i^{th} criterion. After determining the best and worst solutions for all criteria, the Vikor algorithm has the following steps:

1. Compute the values S_j and R_j, $j = 1, 2, \ldots, J$ as:

$$S_j = \sum_{i=1}^{n} w_i \left(f_i^* - f_{ij} \right) / \left(f_i^* - f_i^- \right), \qquad (15)$$

$$R_j = \max_i \left[w_i \left(f_i^* - f_{ij} \right) / \left(f_i^* - f_i^- \right) \right], \qquad (16)$$

where υ is the maximum group utility, here $\upsilon = 0.5$.

2. Sort the values of *S*, *R* and *Q* in decreasing order, obtaining three ranked lists.

3. The alternative that minimizes *Q* is selected as the compromise solution if two conditions of "acceptable advantage" and "acceptable stability in decision making" are satisfied. For more information about these conditions, refer to [167].

4.2.8 Evaluation Criteria

In this thesis, four criteria (stability, similarity, discrimination ability and accuracy) are considered for evaluating and comparing the different features and feature selection techniques. In the following, each of these criteria are briefly described.

4.2.7.1 Stability

Stability of a feature selection method is defined as its sensitivity to variations in the training set. In this study, in order to measure the stability of feature rankings produced by different methods, a similarity based approach proposed by Kalousis et al. [168] is used. In this method, similarity between two selected feature sets s and s', is calculated using the Tanimoto distance which measures the overlap between two sets of arbitrary cardinalities:

$$S_s(s,s') = 1 - \frac{|s| + |s'| - 2|s \cap s'|}{|s| + |s'| - |s \cap s'|} \tag{17}$$

S_s takes values in the range of [0 1], with 0 meaning there is no overlap or similarity between two rankings and 1 meaning that the two rankings are identical. Then *N* subsets of the original training set are drawn using a random resampling technique such as cross validation or bootstrapping. Each specific ranking algorithm produces a feature preference list for each

N subsets. The similarity between all pairs is calculated. The stability of that specific feature ranking algorithm is simply the average of the similarities over all possible pairs, i.e. $\frac{N(N-1)}{2}$ pairs.

4.2.7.2 Similarity

The stability measure used for assessing the internal stability of a feature selection technique can also be used in a different context to measure the similarity of different feature selection techniques. The similarity measure provides information about the consistency and diversity of different feature selection algorithms. The similarity between two feature subsets s and s' can be calculated using Equation (9) with a slight difference in the definition of s and s'. Instead of two lists of features produced by a specific feature selection technique from different subsets of the training set, they are now two lists produced by two different feature selection techniques derived from the complete training set [168].

4.2.7.3 Accuracy

To measure the classification accuracy, the overall accuracy value was calculated as follows [169]:

$$\text{Accuracy} = \frac{\text{No. of true detections}}{\text{Total no. of epochs}} \tag{18}$$

4.2.7.4 Discrimination Ability Analysis

The neurophysiological signals recorded for analysing the sleep quality show similarities with each other [65] especially in REM and N1 stages. This similarity affects the performance of staging algorithm negatively. Therefore, in automatic sleep stage classification, one of the most important quality measures for a feature is the ability of that feature to distinguish pairs of

sleep stages. These pairs include Wake-REM, Wake-N1, Wake-N2, Wake-N3, REM-N1, REM-N2, REM-N3, N1-N2, N1-N3, and N2-N3.

In this thesis, the ability of each feature in total feature set to discriminate between each specific pair of sleep stages was evaluated using two-tailed student's t-test [170]. Student's t-test is a hypothesis testing method for comparing the means of two populations.

4.3 Summary

This chapter presented the datasets and methods utilized in this thesis for developing the proposed techniques for feature extraction and selection. Details of PSG data in each database together with the applied pre-processing steps were described. Two feature sets (conventional and distance-based) were used in this thesis work. Conventional feature set is a collection of the most common features used in automatic sleep stage classification. On the other hand, distance-based feature set consists of three main types of features measuring the distance, (using Itakura distance, Itakura-Saito distance or COSH distance). For the first time in sleep stage classification, a total 31 distance-based features were generated to be used and extensively assessed.

Next, feature ranking and rank aggregation methods were described. These methods will be used in evaluation of the individual features described in the next chapter. The classification techniques used throughout the thesis were also described in this chapter. Finally, the evaluation criteria for assessing the potency, similarity, stability and discrimination ability of the proposed features and feature extraction methods were presented. In the next chapter, the methodology of the contributions together with the details of validation experiments and their corresponding results will be described.

Chapter 5

5. Methodology and Results

This chapter focuses on the contributions and main findings of this thesis work. It is divided into two main subsections, feature selection and feature extraction.

In the feature selection subsection, first the performance of several feature ranking methods applied on the conventional feature set is evaluated. Then two rank aggregation techniques are utilized for the first time in sleep stage classification and their performance is compared to feature ranking methods. The stability and similarity of the generated feature lists is evaluated with three different criteria namely, accuracy, stability and similarity. This contribution is supported by the following publications:

- S. Najdi, A. A. Gharbali, and J. M. Fonseca, "A Comparison of Feature Ranking and Rank Aggregation Techniques in Automatic Sleep Stage Classification Based on Polysomnographic Signals," in 4th International Conference, IWBBIO, 2016, pp. 230–241.
- S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature ranking and rank aggregation for automatic sleep stage classification: a comparative study," *Biomedical Engineering OnLine*, vol. 16, no. S1, p. 78, Aug. 2017.

Next, to compactly represent the feature vector in sleep stage classification, a feature transformation and dimension reduction method based on SSAE is proposed. The performance of the proposed method is evaluated by classification accuracy. This contribution is supported by the following publication:

- S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature Transformation Based on Stacked Sparse Autoencoders for Sleep Stage Classification," in *Technological Innovation for Smart Systems*, 2017, pp. 191–200.

In the feature extraction subsection, first the contribution of a distance-based features in sleep stage classification is assessed and compared to the performance of the conventional features. The evaluation criteria in this work is the classification accuracy and the discrimination ability. This contribution is supported by the following publication:

- A. Gharbali, S. Najdi, and J. M. Fonseca, "Investigating the contribution of distance-based features to automatic sleep stage classification," *Computer in Biology and Medicine*, vol. 96, pp. 8–23, May 2018.

Finally, to enhance the PSG signal quality before feature extraction, a loss-less artefact removal algorithm based on adaptive filtering is proposed. The effect of proposed method is evaluated by the classification accuracy. This contribution is supported by the following publication:

- A. Gharbali, J. M. Fonseca, S. Najdi, and T. Y. Rezaii, "Automatic EOG and EMG Artefact Removal Method for Sleep Stage Classification," in *7th IFIP Advanced Doctoral Conference on Technological Innovation for Cyber-Physical Systems*, 2016, pp. 142–150.

All simulations for the validation of proposed methods were performed using a PC with 3.40 GHz Intel® Core™ i7-3770 CPU, 8 GB of RAM, Windows 10 (64 bits), and MATLAB R2015b.

5.1 Feature Selection

In the following, our contribution in feature selection step of sleep stage classification will be described.

5.1.1 Feature Ranking and Rank Aggregation

To the best of our knowledge, the performance of various feature selection methods from the same category in sleep stage classification has not been compared so far. Moreover, the potential of ensemble feature selection methods has not been explored in this area. In this section, we utilize six feature ranking techniques together with two different heuristic rank aggregation methods to blend the ranking results of several methods. Their performance is evaluated by three criteria: accuracy, stability and similarity. For classification two different classifiers are used, nearest neighbour, and MLF neural networks.

5.1.1.1 Methodology

Figure 20 shows the block diagram of sleep stage classification methodology implemented for investigation and evaluation of several feature ranking and rank aggregation techniques. The data used in this study was obtained from The Physionet Sleep-EDF database [Expanded], [140]. Pz-Oz EEG channel together with submental chin EMG and horizontal EOG, sampled at 100 Hz, were used in the evaluations. In this study for reducing the artefacts, and guarantee the reliability of the classification results, all three pre-processing steps, including band pass filtering, windowing and trimming (described in chapter 4) were applied to the selected PSG subset.

For the WP-based filtering, a Daubechies order 20 (db20) was used as mother wavelet.

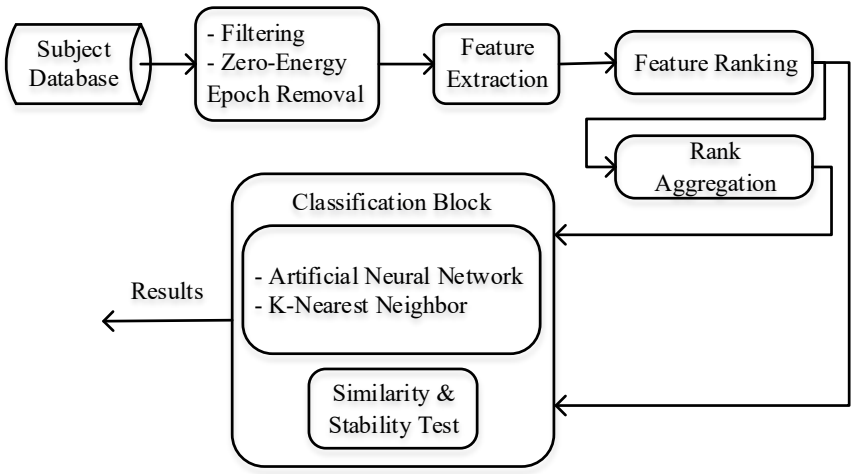

Figure 20. Block diagram of the proposed method for comparing the conventional and the ensemble feature ranking methods.

In order to explore the information contained in PSG recordings, a feature set including 49 features was extracted from each epoch (48 features from Table 6 and F49 from Table 8 in chapter 4). The extracted features can be categorized into time, frequency, joint time-frequency domain, entropy-based and nonlinear types. To avoid that features with greater numeric values, dominate those with smaller numeric values, affecting the classification performance, the extracted features were normalized using standardization method to achieve zero mean and unit variance.

After feature extraction and normalization, the feature set was fed into seven feature ranking methods, namely ReliefF, Minimum Redundancy-Maximum Relevance (MRMR-MID and MRMR-MIQ), Fisher Score, Chi-Square, Information Gain (IG) and Conditional Mutual Information Maximization (CMIM). In order to combine the resulting ranked feature lists, Borda and RRA techniques were also implemented, producing two

additional ranked list of features. In the classification stage, the Euclidean distance was chosen as the distance metrics for the nearest neighbour classifier. In addition to the nearest neighbour classifier, an MLF neural network with 12 neurons and sigmoid transfer function was also used in our simulations. The Levenberg-Marquardt training algorithm was preferred for minimizing the cost function because of its fast and stable convergence. For performance assessment, three main criteria including stability, accuracy and similarity were considered. In the following section the evaluation results are presented.

5.1.1.2 Results

In this study, in order to assess the stability of feature rankings, a similarity based approach proposed by Kalousis et al. [171] (described in chapter 4) was used. For each feature selection method $N = 50$ subsets were generated by bootstrapping. The stability of each method was evaluated as a function of the number of selected features (d) in which $d = 1,\ 3,\ 5 \ldots\ 29$. The corresponding results are shown in Figure 21. Table 9 provides significant information about the variations of stability with regards to the number of features,. In this table the mean value of stability is calculated for fifth, thirteenth and twenty-ninth features.

Classification accuracy was calculated as the ratio of truly classified epochs to the total number of epochs [172]. To estimate the generalization ability of the classifier, *repeated random sub-sampling validation* with 200 runs was used. Figure 22 shows the accuracy of the classifiers with respect to the number of selected features. As this figure shows, starting from one feature, each additional feature typically leads to an increment in the classification accuracy. However, at some point, the increment of the classification accuracy for each additional feature is not significant, leading to an elbow in the graph. Inspired by the “elbow” point in the cost-benefit

curves, in this work we used the Kneedle algorithm proposed in [173] for determining the optimal feature number that provides a satisfactory trade-off between selected number of features and classification accuracy.

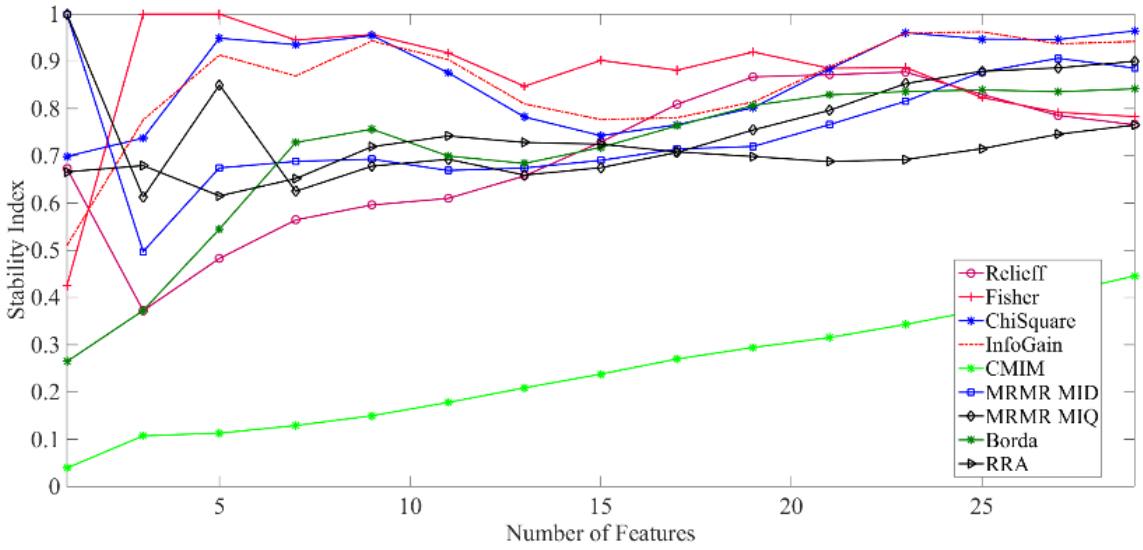

Figure 21. Stability measure of each feature selection method.

The optimum number of features for each classifier, selected by the Kneedle algorithm, together with the corresponding classification accuracies are shown in Table 10. This table also illustrates the top 10 features selected by each feature ranking technique.

Table 9. Mean stability for 5th, 13th, and 29th features by different ranking techniques.

	ReliefF	**Fisher**	**Chi-square**	**IG**	**CMIM**	**MRMR-MID**	**MRMR-MIQ**	**Borda**	**RRA**
Mean stability up to 5th feature	0.50	0.80	0.79	0.73	0.20	0.72	**0.82**	0.39	0.65
Mean stability up to 13th feature	0.66	**0.99**	0.95	0.92	0.21	0.79	0.82	0.68	0.78
Mean stability up to 29th feature	0.69	0.86	0.86	**0.94**	0.24	0.75	0.77	0.70	0.70

The stability measure used for assessing the internal stability of a feature ranking technique can also be used in a different context to assess the similarity of these techniques. Table 11 shows the similarity results for all the ranking techniques used in this study. The similarity index has been calculated for the first 29 features selected by each method.

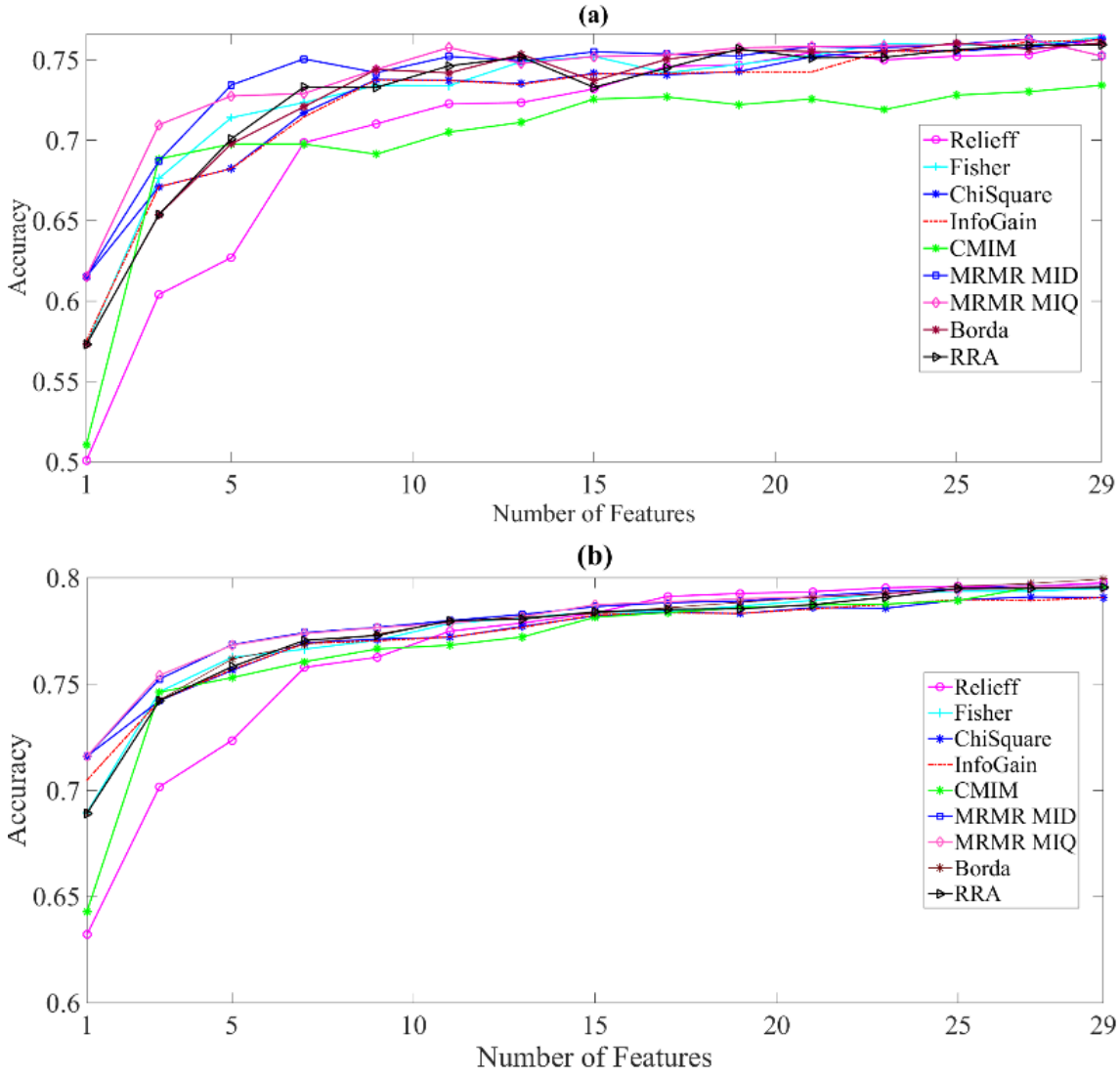

Figure 22. Classification accuracy for different feature ranking and rank aggregation methods, (a) nearest neighbour classifier, (b) MLF neural network.

Table 10. Top 10 features selected by each method and the corresponding optimum number selected by Kneedle algorithm.

	ReliefF	Fisher	CHI	IG	CMIM	MRMR-MID	MRMR-MIQ	Borda	RRA
Top 10 Features	F28	**F36**	F35	F9	F15	F35	F35	**F36**	**F36**
	F36	F35	F9	F35	**F36**	F39	F42	F35	F35
	F7	F31	F11	F11	F9	**F36**	F15	F9	F9
	F49	F9	F31	F31	F8	F22	**F36**	F31	F31
	F41	F29	**F36**	**F36**	F1	F15	F22	F22	F27
	F27	F11	F27	F4	F34	F31	F23	F27	F22
	F20	F25	F26	F27	F35	F29	F31	F29	F17
	F23	F27	F4	F26	F28	F23	F38	F11	F29
	F6	F12	F25	F25	F6	F9	F29	F15	F11
	F22	F22	F14	F29	F48	F38	F9	F20	F20
MLF	7 (0.75)	5 (0.76)	7 (0.76)	7 (0.76)	3 (0.74)	5 (0.76)	5 (0.76)	5 (0.76)	7 (0.77)
Nearest Neighbours	7 (0.69)	5 (0.71)	9 (0.73)	9 (0.73)	3 (0.68)	7 (0.75)	11 (0.75)	9 (0.74)	7 (0.73)

Table 11. Similarity of the feature ranking and rank aggregation techniques.

	ReliefF	Fisher	CHI	IG	CMIM	MRMR-MID	MRMR-MIQ	Borda	RRA
ReliefF	1	0.26	0.18	0.18	0.35	0.40	0.40	0.31	0.31
Fisher		1	0.58	0.52	**0.11**	0.58	0.65	0.72	0.65
CHI			1	**0.90**	0.15	0.35	0.35	0.52	0.52
IG				1	0.18	0.35	0.35	0.46	0.46
CMIM					1	0.22	0.22	0.22	0.22
MRMR-MID						1	**0.90**	0.72	0.65
MRMR-MIQ							1	0.72	0.65
Borda								1	0.72
RRA									1

5.1.2 Feature Transformation Based on Stacked Sparse Autoencoders

One of the main challenges of automatic sleep stage classification is to compactly represent the subject's data in the form of a feature vector. As mentioned in chapter 2, some conventional feature transformation methods such as PCA [133] and KDR [135] were used for reducing the dimensionality and enhancing the descriptive power of feature vector.

Considering the fact that deep learning methods have found their way into many artificial intelligence applications with successful results reported from academia and industry, the main motivation for the current work was to explore the potential of deep learning for feature transformation and classification in the automatic sleep stage classification area. Therefore, we proposed a deep learning-based dimension reduction, feature transformation and classification method for automatic sleep stage classification.

5.1.2.1 Methodology

Figure 23 shows an overview of sleep stage classification framework with the proposed deep learning-based feature transformation scheme.

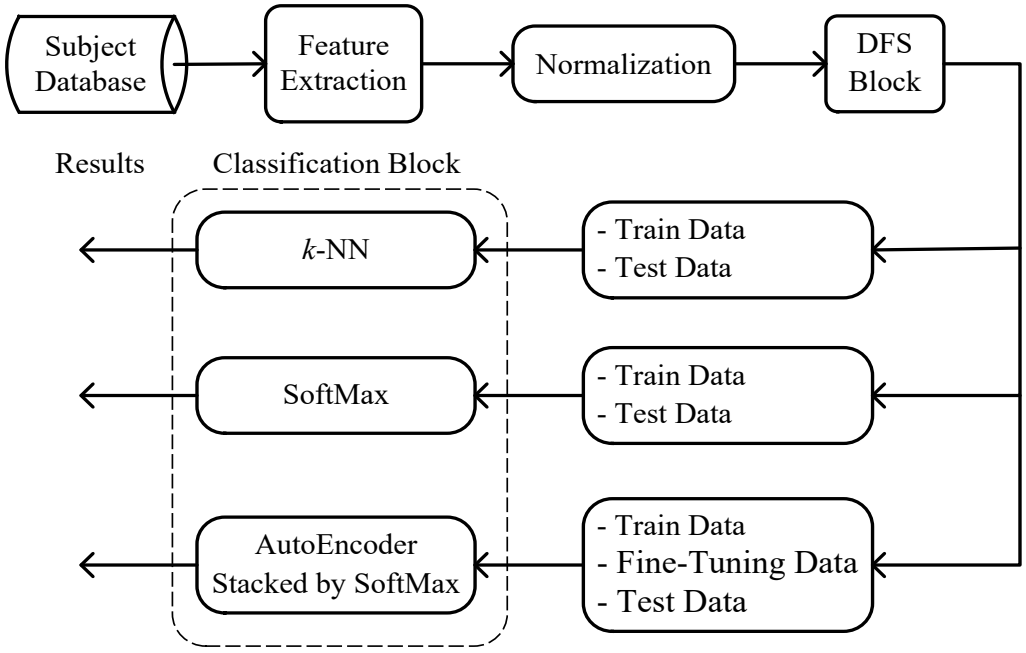

Figure 23. Block diagram of the sleep stage classification framework with deep learning-based feature transformation.

We used a publically available dataset, called ISRUC-Sleep [141]. The data was acquired from 10 healthy adults, including 9 male and 1 female subjects aged between 30 and 58. For the evaluation of the proposed method, we used C3-A2 EEG channel, right EOG and chin EMG channels. The number of epochs, available in this dataset, for these 10 subjects is 954, 941, 824, 794, 944, 853, 814, 1000, 969, and 796. To avoid overfitting we used all of 8889 epochs from healthy subjects available in this database.

All signals used in this study, were divided into 30-second epochs. A set of features were extracted from each epoch of EEG, EOG and EMG recordings of each subject. This feature set included 49 features that can be considered as time, frequency, joint time-frequency domain, entropy-based and nonlinear types. For a comprehensive description regarding the features (F1 to F48 and F49) see Chapter 4, Tables 6 and 8. Next, Min-Max

normalization method was applied to standardize the range of the extracted features.

In this work, a Discriminative Feature Selection (DFS) algorithm was proposed to remove the "near-zero variance" features. Suppose, a feature that has a single value for all of the samples. According to [174], this feature is called "zero-variance predictor". Even if it has little effect on the next steps, this feature should be discarded from the feature set, because it has no information and increases the computational complexity of the overall system. Similarly, some features may have few unique values that occur with low frequency. These features are called "near-zero variance predictors". Kuhn et al. [174] defines two criteria for detecting near-zero variance features as follows:

1. The ratio of unique values to the number of samples is low, for example 10%.
2. The ratio of the frequency of the most dominant value to the frequency of the second dominant value is high, for example 20.

Using these two criteria, we applied DFS to remove the features that didn't have enough discriminative power. As a result, 12 features were recognized as near-zero variance features and removed from our sleep data model. The features are as follows: maximum value (F2), minimum value (F3), variation (F5), median (F8), Petrosian fractal dimension (F9), permutation entropy (F12), Hjorth parameter (Activity) (F14), ZCR (F18), EMG spectral power (F37), mean of the EMG spectral power distribution (F39), EMG temporal energy (F41), maximum value of time domain EOG signal (F46).

After the feature vector was set, data was divided into two parts, training and testing, using 10-fold cross validation. For the fine tuning step of SSAE, part of the training data was utilized. Our deep learning consists of three

layers: a two-layer SSAE and a Softmax layer. The number of hidden units for the first and second layer of SSAE was 20 and 12, respectively. For finding the best hyper-parameters for the autoencoders, we tried several models by adjusting sparsity regularization parameter, weight regularization parameter and the number of iterations. We used autoencoders with logistic sigmoid activation function for both layers.

The performance of the proposed algorithm was compared with two other classifiers, Softmax and kNN classifiers. The number of neighbours was set to 18 and Euclidean distance was used as a measure of distance for kNN.

5.1.2.2 Results

To evaluate the performance of deep learning-based feature selection algorithm, we used classification accuracy as the evaluation criterion. Table 12 shows the individual sleep stage and overall classification accuracy extracted from confusion matrix for three different classifiers. The boldface numbers indicate the best performance. To confirm the advantage of DFS block, the performance of SSAE-based sleep stage classification with and without this step was also investigated. Without using DFS block, 49 original features were fed to SSAE. The classification accuracy achieved in this way was 74.1% which is almost 8% less than the accuracy with DFS block.

Table 12. Results of the statistical analysis for comparison of each stage and overall accuracy.

Classifiers	Wake (%)	REM (%)	N1 (%)	N2 (%)	N3 (%)	Overall Accuracy (%)
Softmax	80	61.66	65	**90**	78.33	74.9
kNN	85	66.66	61.66	70	83.33	73.33
SSAE	**91**	**77**	**69**	87	**87**	**82.2**

5.2 Feature Extraction

In the following, our contribution in feature extraction step of sleep stage classification will be described.

5.2.1 Investigating the Contribution of Distance-based Features to Automatic Sleep Stage Classification

One of the main motivations for this thesis was to evaluate new features to characterize each sleep stage in such a way that extracted features were more powerful than conventional features to distinguish sleep stages from each other, and to improve classifiers accuracy. Considering the outstanding performance of Itakura and Itakura-Saito distances in sleep and speech signal processing [52], [70], [144] and COSH distance in speech signal processing [145], [146], we aimed to extensively evaluate the performance of distance-based features together with conventional features in automatic sleep stage classification. The distance-based features were extracted by calculating Itakura, Itakura-Saito and COSH distances of autoregressive and spectral coefficients of EEG, EMG, EOG and ECG signals according to Table 8 in chapter 4.

5.2.1.1 Methodology

In this work, we used the open-access comprehensive ISRUC-Sleep dataset [141]. For our evaluations, we used PSG recordings from healthy subjects. Nine male and one female subjects aged between 30 and 58 participated in the recordings. Each recording contains signals from 19 channels. The data include six EEG channels: F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, and O2-A1 from which we selected the C3-A2 EEG channel. The C3-A2 channel is the commonly used EEG channel in sleep stage classification [10], [16], [25], [27] and is among the recommended channels by AASM. In

addition to one EEG signal, we used the signals from right EOG and chin EMG, and ECG channels of all ten subjects.

Figure 24 shows the framework used in this study. In the following, each part will be described in detail. In this study two groups of features namely,

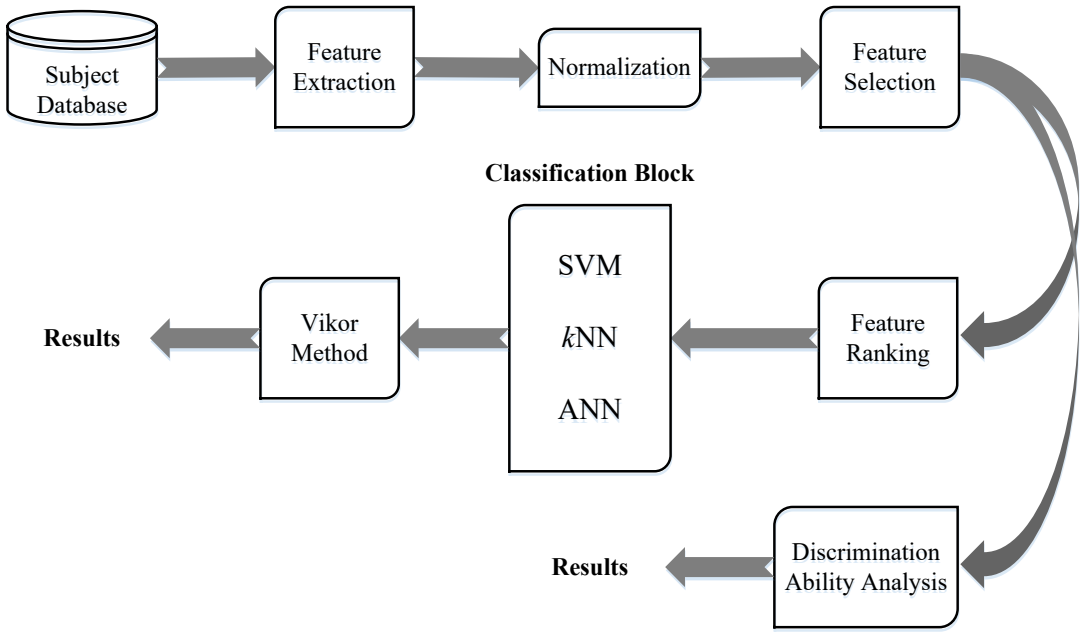

Figure 24. Sleep Study Framework for analysing the contribution of distance-based features.

conventional feature set and distance-based feature set were extracted from 30-second long the epochs of selected PSG subset. The conventional feature vector consists of 48 features extracted from EEG, EOG, and EMG signals. These features were described in Table 6 of chapter 4 as F1 to F48. In this study, the contribution of a set of 32 distance-based features, extracted from EEG, EOG, EMG and ECG, was evaluated for sleep stage classification as described in Table 8 chapter 4, F49 to F80. A third feature set was also created, named total feature set composed of pruned distance-based and pruned conventional feature sets.

The features extracted from PSG signals were in different ranges, and this variety could bias the results of the following steps. Therefore, two different types of normalization methods were used namely, standardization and Min-Max. The effect of each method in the overall system performance was evaluated. Next, to remove the features with high levels of similarity, a feature selection method was proposed and used. Existence of similar features negatively affect the stability [168] of the feature ranking results; therefore, excluding similar features from the feature set can improve the overall performance of the proposed algorithm [175]. The proposed algorithm worked as follows:

After the L1-norm between each pair of feature vectors was calculated, a similarity threshold was defined. The feature pair, whose L1-norm was less than the threshold level, was considered strongly similar. In this way, the features were clustered into groups of similar features, and one feature per cluster was selected as representative. The representative feature had the lowest computational complexity.

Alternatively, it was possible to use PCA for finding the most dissimilar features. However, there are two main reasons why we did not use PCA. First, using PCA for finding a non-redundant feature set would lead to keeping and calculating all the features in the classification and practical application steps, whereas by using the similarity threshold, the most redundant features can be detected and omitted from the feature set in the application step. Second, PCA would generate combinations of the features. Since our aim was to evaluate the performance of the distance-based and compare it with the performance of the conventional features, it was necessary to preserve the information of the features and PCA was not a proper option in this regard.

To analyse the potential of individual features in sleep stage classification, six feature ranking techniques were adopted. In particular, we used ReliefF, mRMR-MID, mRMR-MIQ, Fisher score, Chi-square and IG techniques. The description of these methods was provided in chapter 4. Each of these methods was applied on the conventional, distance-based and total feature, and all in all, 3*6=18 ranked lists of features were achieved.

For classification, three different classifiers were used: kNN, MLF neural network and DSVM. The reason for choosing these three different classifiers is that we did not want to restrict the significance of the comparison to one specific family of classifiers, and on the other hand, we aimed to choose a variety of classifiers including the simplest, most used and the one that usually shows the best performance. Euclidean distance was used as the distance measure for the kNN classifier. In each experiment, the classification accuracy for the 1, 2, …20 neighbourhood was calculated, and the one leading to maximum accuracy was selected as the optimum neighbourhood number.

For the MLF neural network classifier, a three-layered feed forward neural network with 20 hidden neurons for the conventional and total feature sets and 12 hidden neurons for the distance-based feature set were used. DSVM was used instead of conventional multi-SVMs. The reason for choosing DSVM was that it outperforms conventional multi-SVMs (OAO and OAA) while utilizing lower number of SVM in the structure [165], [176]–[178]. Radial Basis Function (RBF) was selected as the kernel function, and sigma was set to 3.0 for the conventional and total feature sets and 1.1 for the distance-based feature set.

For each ranked list of features, created by one of the ranking methods, and each specific classifier, the classification accuracy was calculated for the top 1, 2, … 25 features. Since it is always desirable to achieve the

maximum accuracy with the minimum complexity, to find the optimum number of features, Vikor method was used for multi-criteria (i.e. classification accuracy and number of features) decision making [166]. Finally, the ability of the top 25 features in the total feature set, selected by different feature ranking methods, to discriminate between each specific pair of sleep stages was evaluated using two-tailed student's t-test. These pairs include Wake-REM, Wake-N1, Wake-N2, Wake-N3, REM-N1, REM-N2, REM-N3, N1-N2, N1-N3, and N2-N3.

5.2.1.2 Results

In this section, the evaluation results of the framework depicted in Figure 24 considering different normalization methods, feature ranking techniques and classifiers are presented.

After feature extraction and normalization, the highly similar features in both conventional and distance-based feature sets were detected. The threshold value of L1-norm between each pair of feature vectors was empirically set to $1e^{-15}$. This value was chosen empirically. For conventional and distance-based feature sets, the similar groups were detected and are listed in Table 13.

Table 13. Similar feature groups from the conventional and distance-based feature sets.

Conventional Feature Set	**Group 1**			**Group 2**		
	F36, F38 and F40			F6 and F14		
Distance-based Feature Set	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**
	F52 and F74	F55 and F75	F56 and F76	F60 and F78	F63 and F79	F64 and F80

According to this table, several similar cases were found using this measure. For example, the Hjorth activity parameter is the same as the

variation. Also, the COSH distance is the symmetric version of the Itakura-Saito distance. From each group of similar features, one feature with the lowest computational complexity was selected as representative of the group. Therefore, F14, F38 and F40 were removed from the conventional feature set. F74, F75, F76, F78, F79 and F80 were also removed from the distance-based feature set. After removing the redundant features, 45 features remained in the conventional feature set, and 26 features remained in the distance-based feature set.

To assess the usefulness of pruning feature sets, the sleep stage classification accuracy before and after feature selection was evaluated using the conventional, distance-based, and total feature sets. The results obtained by the kNN classifier with Euclidean distance are shown in Table 14. The optimum number of neighbours for each case was found (shown in brackets in Table 14) by evaluating the performance of the classifier for different numbers of neighbours. According to the results, removing similar features led to an average improvement of 0.61% for all the cases. The maximum improvement (2.07%) was observed in the pruning of the conventional feature set using the standardization method. Additionally, it is notable that the accuracy of the classification with the Min-Max method is, in all cases, higher than the one with the standardization method. This emphasizes the importance of selecting a proper feature normalization method before classification.

Table 14. Classification accuracy for the original and pruned feature sets using the kNN classifier. The numbers in brackets refer to the nearest neighbours used in each case.

Features / Normalization	Distance-Based	Pruned Distance-Based	Conventional	Pruned Conventional	Total
STD	60.88 (15)	61.03 (5)	70.90 (15)	72.97 (26)	73.26 (12)
Min-Max	62.30 (10)	62.37 (5)	73.94 (8)	74.10 (8)	74.42 (6)

For determining the features that should be given a high priority when dealing with the description of PSG signals, six feature ranking techniques were applied on three feature sets: conventional, distance-based and total feature sets. Furthermore, each feature set was considered with two different normalization methods. From each group, the top 25 features were selected for comparison as shown in Tables 15-17. Table 15 shows the feature ranking results for the conventional feature set. The results of this table are summarized in Figure 25. According to this figure, temporal and time-frequency domain features are preferred by the ranking methods, whereas frequency domain features are the least preferred ones. Nonlinear and entropy features are always among the top 25 and occupy five to six places on the list. Detailed assessment of these features leads to the following observations about conventional features:

- EEG ZCR (F18) has been chosen as the best feature by most of the ranking methods with either the standardization or Min-Max method. Even the methods that did not select F18 as the first feature such as ReliefF, have it ranked in the top five best features.
- Petrosian fractal dimension (F9), Hjorth parameter (Mobility) (F15), and Hurst exponent (F21) are among the top ranked-features by all the methods.
- ReliefF, mRMR-MID and mRMR-MIQ methods include EEG-, EMG-, and EOG-related features in their top 25 list, whereas Fisher, Chi-square, and IG only contain EEG-related features.
- Between EMG and EOG features, those related to EOG are more preferred by the ranking methods, such as EOG kurtosis, maximum, and standard deviation.

- Features from time-frequency domain that were extracted using WP are ranked in the top 25 features by all methods.

Table 15. Feature ranking results for the conventional feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
1th	F45	F13	F18	F18	F18	F18	F21	F18	F18	F18	F18	F15
2nd	F16	F9	F34	F11	F34	F11	F18	F15	F21	F15	F21	F18
3rd	F15	F21	F10	F9	F10	F45	F15	F21	F15	F21	F15	F21
4th	F18	F15	F21	F35	F37	F35	F9	F9	F9	F9	F34	F9
5th	F13	F18	F35	F21	F29	F9	F13	F13	F34	F16	F9	F16
6th	F29	F16	F15	F45	F13	F32	F34	F34	F35	F11	F35	F11
7th	F21	F32	F13	F15	F23	F31	F35	F35	F4	F26	F4	F2
8th	F9	F29	F29	F32	F21	F10	F11	F16	F28	F13	F28	F13
9th	F32	F45	F23	F31	F45	F21	F4	F4	F22	F2	F23	F34
10th	F7	F7	F46	F10	F35	F30	F16	F25	F16	F27	F22	F22
11th	F31	F31	F9	F13	F15	F15	F29	F29	F23	F20	F5	F35
12th	F48	F6	F26	F30	F25	F29	F22	F30	F36	F22	F19	F3
13th	F41	F25	F11	F29	F11	F34	F30	F22	F5	F34	F11	F26
14th	F6	F10	F4	F34	F48	F13	F28	F33	F19	F25	F16	F4
15th	F25	F41	F25	F4	F9	F23	F25	F28	F11	F29	F36	F20
16th	F11	F48	F2	F25	F2	F25	F33	F27	F27	F3	F13	F27
17th	F10	F46	F31	F23	F26	F4	F31	F26	F13	F30	F27	F29
18th	F36	F11	F16	F33	F32	F33	F23	F31	F29	F35	F29	F30
19th	F39	F42	F32	F16	F31	F16	F2	F2	F20	F4	F2	F36
20th	F46	F34	F37	F2	F4	F2	F5	F5	F26	F36	F30	F28
21th	F27	F3	F45	F22	F46	F22	F27	F19	F30	F33	F20	F25
22nd	F26	F43	F3	F36	F16	F3	F19	F20	F25	F37	F26	F37
23th	F37	F47	F30	F46	F8	F36	F3	F3	F39	F28	F3	F33
24th	F24	F27	F48	F7	F39	F28	F26	F10	F2	F39	F25	F5
25th	F47	F2	F24	F28	F3	F46	F45	F45	F33	F45	39	F19

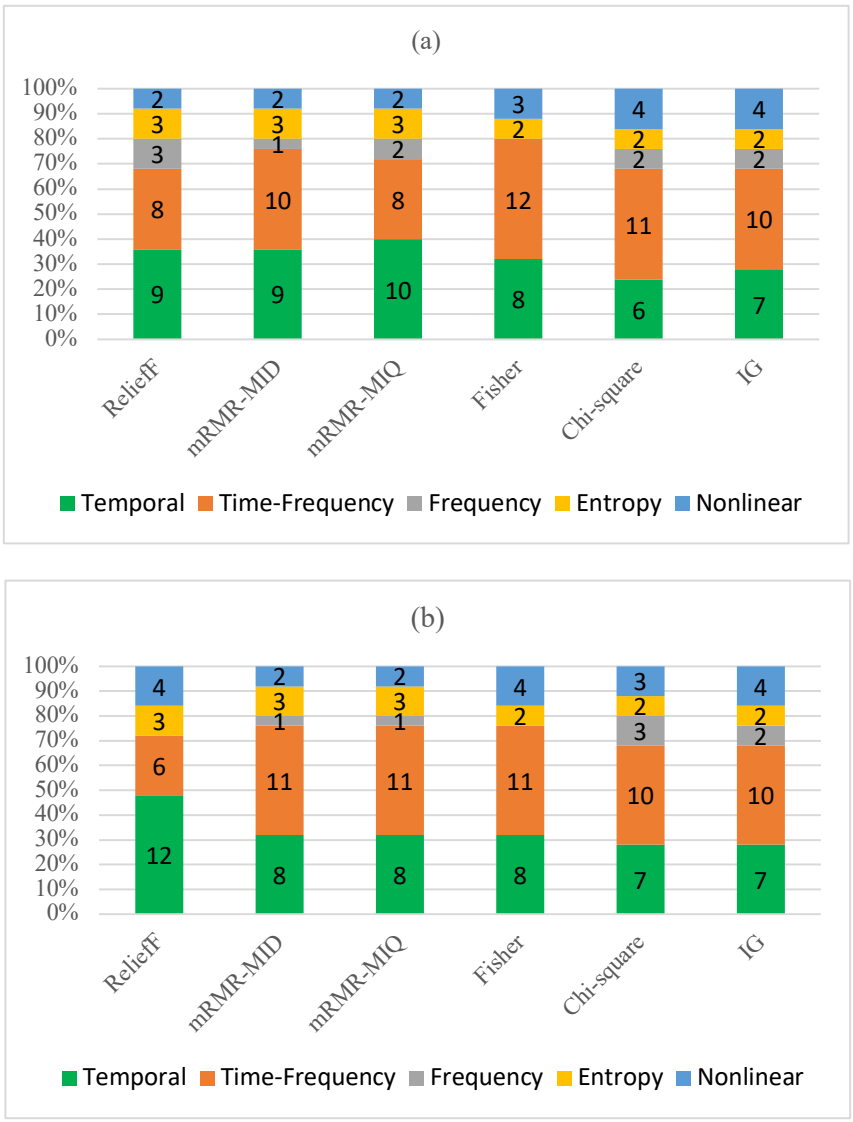

Figure 25. Graphical representation of the feature ranking results for the conventional feature set, (a) normalized with STD and (b) normalized with Min-Max.

Table 16 shows the feature ranking results for the distance-based feature set. Like the conventional feature set, the ranking results are summarized as a graphical representation in Figure 26. According to these charts, Itakura and Itakura-Saito distances were much more effective than COSH distance in discriminating the sleep stages and, at the same time, were preferred equally by the ranking methods. These results imply that the

Itakura and Itakura-Saito features can be used interchangeably in sleep stage classification. Detailed assessment of top 25 distance-based features leads to the following observations:

- Among several types of distance-based features, two are ranked as the best by all methods. These features are similarity between a baseline EEG epoch and the rest of the EEG measured by Itakura distance (F49 and F50) and similarity of EEG and EOG signals measured by either Itakura or Itakura-Saito distance (F65-F68).
- Itakura-Saito distance of AR or spectral coefficients of EEG (F51 and F52) are also seen in the top five.
- All methods rank one of the features related to the similarity of a baseline EOG epoch to the rest of the EOG (F57-F60), measured by Itakura or Itakura-Saito distance, in the top 25.
- The features related to the similarity of a baseline ECG epoch to the rest of the ECG (F61-F64), measured by Itakura or Itakura-Saito distance, are considered important mostly by three methods: ReliefF, mRMR-MID and mRMR-MIQ. The same applies to the similarity between EEG and EMG (F69- F72).
- Among the COSH distance-based features (F73- F80), only COSH distance of EEG AR coefficients (F73) and COSH distance of EOG spectral coefficients (F77) could find their way to the top 25 features list.
- There are no noticeable differences in the number of occurrences of AR or spectral-based features.

Table 16. Feature ranking results for the distance-based feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
1th	F65	F65	F49	F66	F49	F66	F67	F65	F50	F68	F50	F67
2nd	F66	F66	F53	F53	F55	F53	F68	F66	F49	F67	F49	F68
3rd	F49	F50	F68	F67	F68	F67	F66	F68	F51	F49	F51	F49
4th	F50	F49	F61	F57	F58	F57	F65	F67	F73	F50	F73	F50
5th	F70	F70	F58	F61	F65	F49	F54	F53	F52	F51	F52	F51
6th	F69	F69	F65	F49	F61	F55	F53	F54	F57	F73	F58	F73
7th	F61	F58	F69	F69	F53	F65	F56	F55	F58	F52	F57	F52
8th	F62	F57	F55	F68	F50	F68	F55	F56	F67	F57	F67	F57
9th	F72	F72	F50	F65	F67	F69	F49	F49	F68	F58	F68	F58
10th	F71	F71	F67	F55	F57	F54	F50	F50	F65	F65	F65	F65
11th	F52	F62	F71	F51	F66	F61	F57	F57	F66	F66	F66	F66
12th	F73	F61	F57	F63	F54	F51	F58	F58	F60	F60	F59	F60
13th	F51	F60	F59	F54	F69	F50	F70	F70	F59	F59	F77	F59
14th	F63	F77	F66	F59	F51	F70	F69	F69	F77	F77	F60	F77
15th	F64	F59	F54	F52	F56	F56	F51	F73	F53	F53	F53	F53
16th	F57	F63	F70	F71	F63	F52	F73	F51	F54	F54	F54	F54
17th	F58	F52	F51	F64	F59	F58	F52	F52	F55	F55	F55	F55
18th	F60	F51	F63	F56	F73	F63	F60	F60	F56	F56	F56	F56
19th	F77	F73	F72	F50	F70	F73	F77	F77	F61	F70	F61	F70
20th	F59	F64	F56	F70	F52	F59	F59	F59	F62	F69	F62	F69
21th	F55	F53	F60	F73	F60	F64	F72	F72	F63	F72	F63	F72
22nd	F56	F54	F73	F72	F64	F62	F71	F71	F64	F71	F70	F71
23th	F53	F56	F77	F62	F77	F60	F63	F62	F69	F63	F69	F63
24th	F54	F55	F52	F60	F71	F77	F64	F61	F70	F64	F64	F64
25th	F68	F68	F64	F77	F62	F71	F61	F64	F71	F61	F71	F61

Figure 26. Graphical representation of feature-ranking results for the distance-based feature set (a) normalized with STD and (b) normalized with Min-Max.

Table 17 shows the feature ranking results for the total feature set. Furthermore, Figure 27 shows the percentage that each feature group occupies in top 25 feature list. Like the conventional feature set, temporal and time-frequency domain features are the most preferred types by the ranking methods. Distance-based features are always in the top 25. Itakura and Itakura-Saito features were more popular than the COSH features.

Among the ranking methods, only IG and Chi-square have COSH features in their top 25 feature list. Detailed assessment of ranking results leads to the following observations:

- On average, 28% of the top-ranked features was selected from the distance-based feature set. The selected distance-based features in Table 16 belong to one of these categories: similarity of EEG and EOG (F65-F67), similarity of a baseline EEG epoch with the rest of EEG (F49-F52 and F73), similarity of a baseline epoch of EMG with the rest of EMG (F53-F55), and similarity of a baseline EOG epoch with the rest of EOG (F57 and F58).
- Among the feature ranking methods, the Chi-square and IG methods had the maximum percentage of distance-based features (44%) in their top 25. These features include the similarity between a baseline EEG epoch with the rest of EEG, measured by Itakura, Itakura-Saito and COSH distances, (F49-F52 and F73) and the similarity of EEG and EOG, measured by the Itakura-Saito distance (F67 and F68).
- The ReliefF method has the minimum percentage of distance-based features (13%) in its top 25-list. The similarity between EEG and EOG, measured by Itakura distance (F65 and F66), is the selected distance-based feature by this method.
- F73 is the only COSH distance-based feature that appears in top 25 list of the total feature set, and it is related to the similarity of a baseline EEG epoch with the rest of EEG.
- Zero-crossing number (F18) is selected as the best feature by all methods.
- Besides the zero-crossing number, Hjorth parameter (mobility) (F15), approximation entropy (F13), Petrosian fractal dimension (F9), Hurst

exponent (F21) and at least one of the WP-based features (F22-F35) are in the top-ranked features by all methods.

Table 17. Feature ranking results for the total feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
1th	F16	F13	F18	F18	F18	F18	F18	F18	F18	F18	F18	F18
2nd	F15	F9	F34	F11	F34	F11	F21	F15	F21	F15	F21	F15
3rd	F13	F21	F53	F9	F53	F45	F15	F21	F15	F21	F15	F21
4th	F29	F29	F21	F66	F48	F53	F9	F9	F9	F9	F34	F9
5th	F32	F32	F35	F35	F68	F9	F13	F13	F49	F16	F9	F16
6th	F45	F16	F68	F21	F32	F35	F34	F65	F50	F68	F35	F49
7th	F18	F15	F15	F45	F46	F32	F67	F66	F34	F67	F4	F50
8th	F7	F7	F46	F15	F21	F66	F68	F34	F35	F49	F49	F68
9th	F9	F18	F13	F31	F35	F31	F66	F35	F4	F50	F50	F67
10th	F21	F31	F29	F32	F13	F21	F65	F68	F51	F11	F22	F11
11th	F65	F11	F23	F53	F10	F10	F35	F67	F73	F51	F23	F51
12th	F66	F6	F2	F13	F11	F30	F11	F16	F52	F73	F28	F73
13th	F10	F45	F57	F10	F25	F15	F54	F4	F22	F52	F52	F52
14th	F6	F10	F11	F4	F23	F23	F53	F54	F16	F13	F51	F2
15th	F48	F34	F9	F29	F15	F29	F4	F53	F28	F26	F73	F13
16th	F41	F25	F26	F30	F58	F13	F16	F25	F23	F2	F5	F34
17th	F36	F47	F4	F65	F55	F34	F29	F30	F11	F27	F19	F22
18th	F31	F66	F55	F34	F2	F67	F30	F29	F68	F20	F11	F3
19th	F39	F65	F65	F23	F29	F25	F25	F33	F67	F22	F68	F35
20th	F37	F24	F49	F25	F9	F4	F33	F22	F36	F34	F67	F26
21th	F61	F48	F25	F54	F26	F65	F31	F28	F58	F65	F16	F4
22nd	F62	F41	F31	F33	F65	F33	F56	F27	F5	F66	F58	F66
23th	F2	F37	F10	F67	F4	F54	F55	F31	F57	F3	F57	F65
24th	F34	F46	F67	F68	F37	F68	F22	F5	F19	F25	F13	F57
25th	F46	F43	F32	F49	F31	F69	F27	F26	F13	F29	F36	F58

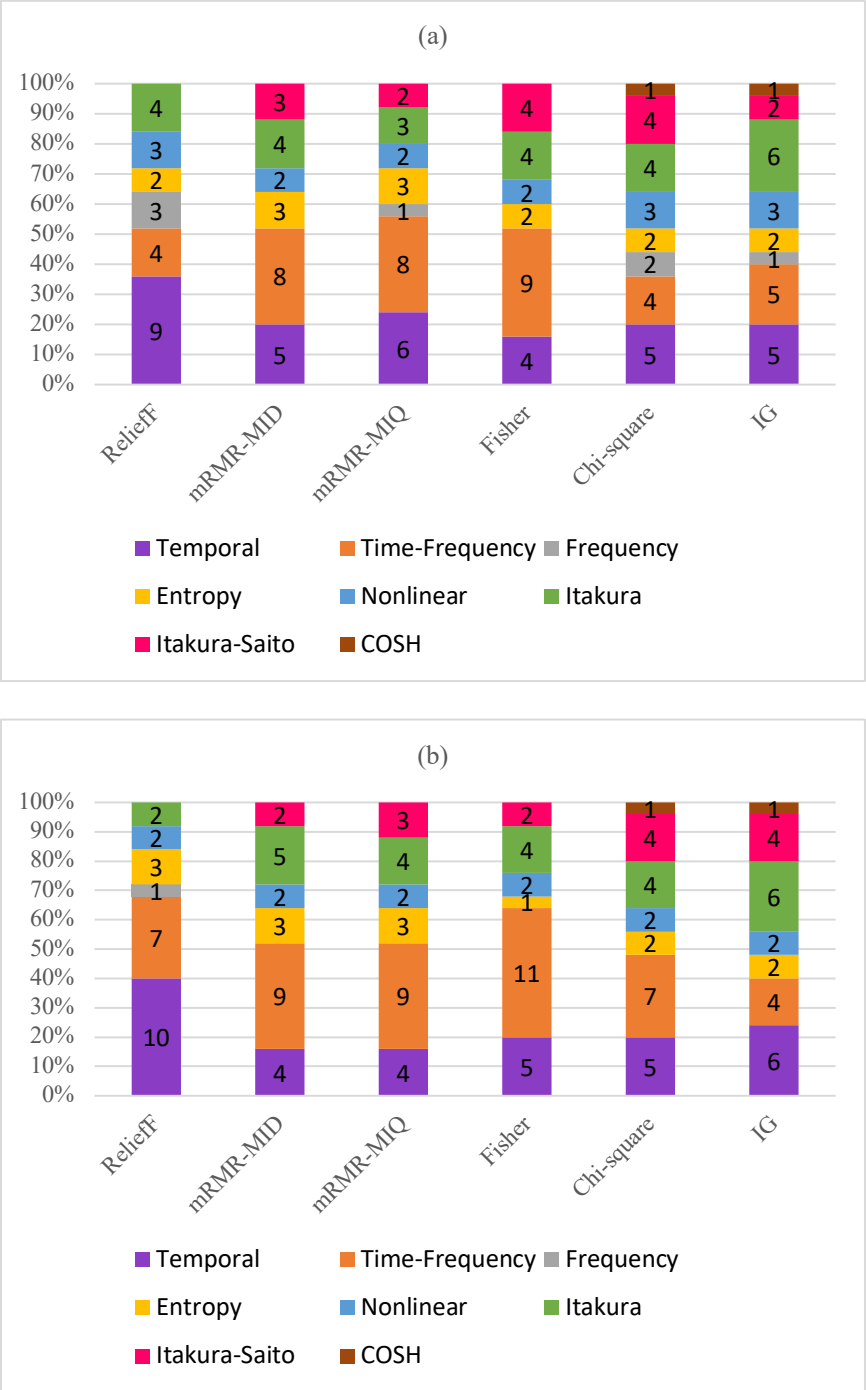

Figure 27. Graphical representation of feature ranking results for the total feature set (a) normalized with STD (b) normalized with Min-Max.

- There are some features never ranked in the top 25 by any of the methods. Examples of these features are mean curve length (**F17**) and mean Teager energy (**F20**).
- Tables 18-26 depict the 5-stage (Wake, REM, N1, N2 and N3) classification accuracy results along with the optimum number of features selected by the Vikor method for all three feature sets and three classifiers. The reliability of the results was validated by using 10 times repeated 10-fold cross validation method on the whole data from 10 healthy subjects. For each ranked list of features, created by one of the ranking methods, and each classifier, the overall classification accuracy, sensitivity and specificity were calculated for the top 25 features. Sensitivity (also called the true positive rate, the recall) measures the proportion of actual positives that are correctly identified as such. On the other hand, specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such. Analysing the results reveals that, starting with one feature, each additional feature typically leads to an increment in the classification accuracy.

However, at some point, the increment on the classification accuracy for each additional feature is not significant. Inspired by MCDM problems, the Vikor method was applied to the classification results for determining the optimal feature number that provides a satisfactory trade-off between the selected number of features and the classification accuracy. Accuracy and number of features were two conflicting criteria with the corresponding weights of 0.7 (w_1) and 0.3 (w_2), respectively, meaning that, in our sleep stage classification system, classification accuracy had priority over complexity. Figure 28 shows a sample of the Vikor method results for the features scaled by standardization method, ranked with ReliefF and

classified by kNN classifier. The utopian solution, shown with a black star, represents the ideal solution in which the accuracy is maximum, and the number of features is minimum. The selected point by the Vikor method in each case is the closest point of the Pareto front (the set of solutions) to the utopian solution considering the weights of the two criteria.

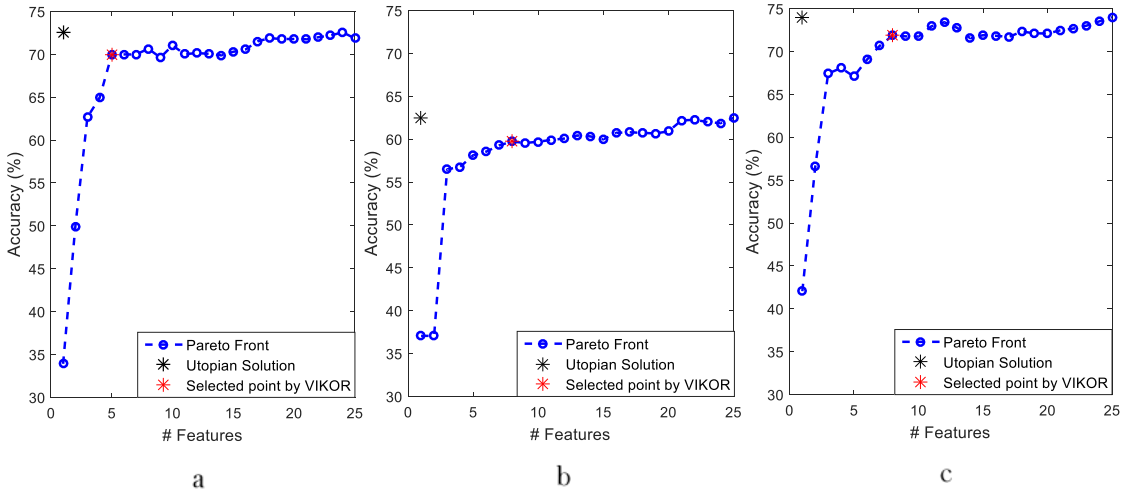

Figure 28. Optimum number of features selected by the VIKOR method for the (a) conventional, (b) distance-based, and (c) total feature sets.

Next, the assessment of the results related to the kNN classifier (Tables 18-20) will be discussed.

- The maximum enhancement in classification accuracy after adding the distance-based features to the conventional feature set occurred in mRMR-MID with Min-Max.
- For all three feature sets, the maximum accuracy, regardless of the feature normalization method, was achieved by mRMR-MID or mRMR-MIQ method. Seven and in one case eight features were selected by the Vikor method to achieve this accuracy. The Itakura distance of EEG-EOG spectral coefficients, Itakura-Saito distance of

EEG-EOG spectral coefficients, and Itakura distance of EMG AR coefficients are among these features.

- For all three feature sets, the minimum accuracy, regardless of the feature normalization method, was achieved by the Chi-square method.
- For most of the ranking methods, adding distance-based features to the conventional feature set improved the sensitivity and specificity of the classification.

Table 18. kNN classifier results for the conventional feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
#Features	5	6	10	7	9	8	8	8	7	8	8	8
#Neighbours	18	16	20	11	20	20	12	12	12	20	16	8
Sensitivity	72.8	72.9	75.6	72.5	73.7	71	71.5	72.7	71.3	74.6	73.1	72.9
Specificity	93.4	93.2	94	93.4	93.5	92.6	93.1	93.2	92.9	93.8	93.3	93.4
Accuracy	70	70.9	72.1	71.3	72.9	70.8	69.7	71.6	69	71.9	69.2	72.7

Table 19. kNN classifier results for the distance-based feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
#Features	8	11	6	6	6	5	11	12	10	8	10	8
#Neighbours	19	6	16	9	17	9	10	11	18	12	17	12
Sensitivity	64.3	61.7	62.5	65.6	64	63	64.3	63.3	63.9	60	64.7	61.5
Specificity	91.2	90.6	90.4	91.1	90.6	91	91.5	90.8	91	89.9	91	90.3
Accuracy	59.7	59	61.5	60.6	61.9	60	62	60	61	56.3	61.1	56.6

Table 20. kNN classifier results for the total feature set.

	ReliefF		mRMR-MID		mRMR-MIQ		Fisher		Chi-square		IG	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
#Features	8	8	8	7	7	7	7	6	7	8	8	10
#Neighbours	14	10	11	6	17	12	10	11	11	10	19	10
Sensitivity	75.1	73.3	74.1	77.4	75.1	75.3	76.5	73.4	72.3	70.6	74	75.4
Specificity	93.8	93.7	93.6	94.2	93.9	93.6	94.2	93.5	93	92.4	93.8	94.3
Accuracy	72	71	73.2	73	72.2	72.3	71.1	71	71	70	71	70.3

Next, the assessment of the results related to MLF neural network classifier (Tables 21-23) will be discussed.

- The maximum enhancement in classification accuracy after adding the distance-based features to the conventional feature set occurred in mRMR-MIQ with standardization.
- For all three feature sets, the maximum accuracy, regardless of feature normalization method, was achieved by the mRMR-MID or mRMR-MIQ method. Up to 11 features were selected by the Vikor method to achieve this accuracy. The Itakura distance of the EEG-EOG spectral coefficients, Itakura-Saito distance of the EEG-EOG spectral coefficients, and Itakura distance of the EMG AR coefficients are among these features.
- Compared to the results of the kNN classifier, the overall accuracy, sensitivity and specificity of MLF classifier is higher for three feature sets.

Table 21. MLF neural network classifier results for the conventional feature set.

	ReliefF		mRMR-MID		mRMR-MIQ		Fisher		Chi-square		IG	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
#Features	9	10	11	11	9	11	8	10	9	9	11	8
Sensitivity	72.6	77.7	75.9	78.3	74.9	76	73.9	74.6	73.4	76.9	73.6	75.4
Specificity	93.7	94.4	94	94.6	93.7	94	93.5	93.6	93.3	94.2	93.4	93.9
Accuracy	79	80	80	80.6	79	79.8	79.8	79.2	78.5	79.7	78.7	79.6

Table 22. MLF neural network classifier results for the distance-based feature set.

	ReliefF		mRMR-MID		mRMR-MIQ		Fisher		Chi-square		IG	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
# Features	9	7	7	7	13	7	11	13	15	15	15	15
Sensitivity	62.1	59.9	63.3	61.1	64.8	61.3	63.4	63.6	66.1	64	65.1	63
Specificity	90.5	90	90.9	90.2	91.1	90.3	90.8	90.9	91.5	90.5	91.2	90.7
Accuracy	74.3	72.1	75.2	74	75.6	74	75	74.2	75	73.1	75	73.1

Table 23. MLF neural network classifier results for the total feature set.

	ReliefF		mRMR-MID		mRMR-MIQ		Fisher		Chi-square		IG	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
# Features	8	9	9	10	8	11	9	9	9	14	8	10
Sensitivity	75.1	75.4	76.5	76.7	76.7	78.8	74	74.8	73.3	76.3	74	74.2
Specificity	93.8	93.8	94.1	94.3	94.2	94.7	93.5	93.7	93.3	94.1	93.5	93.5
Accuracy	79.5	79.2	80.2	79.9	80.2	80.4	79.2	79.1	79.2	79.5	79.2	78.5

Next, assessment of results related to the DSVM classifier (Tables 24-26) will be discussed.

- The maximum enhancement in classification accuracy after adding the distance-based features to the conventional feature set occurred in mRMR-MIQ with Min-Max.
- For all three feature sets, the maximum accuracy, regardless of the feature normalization method, was achieved by the mRMR-MID or mRMR-MIQ methods. Up to 13 features were selected by the Vikor method to achieve this accuracy. The Itakura distance of the EEG-EOG spectral coefficients, Itakura-Saito distance of the EEG-EOG spectral coefficients, and Itakura distance of the EMG AR coefficients are among these features.
- Considering that the overall performance of the DSVM classifier, including accuracy, sensitivity and specificity, is the highest among the classifiers used in this paper, it can be concluded that DSVM outperforms kNN and ANN classifiers in sleep stage classification.

Looking at the results for all the classifiers, the accuracy obtained by Min-Max is higher than standardization in most cases. Furthermore, the presence of the distance-based features among the selected features by the Vikor method shows their positive contribution to sleep stage classification.

Table 24. DSVM classifier results for the conventional feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
# Features	10	10	10	11	9	9	8	11	8	9	8	8
Sensitivity	79.2	74.4	80.1	78.5	79	76.3	77.2	76.6	73.2	78.4	76.3	75.7
Specificity	95.3	94.2	95.7	94.9	95.6	94.6	95.2	94.6	94.7	95.4	94.9	94.7
Accuracy	83.7	84.5	84.0	84.7	84.0	83.8	81.5	81.7	81.0	81.9	81.0	81.8

Table 25. DSVM classifier results for the distance-based feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
# Features	9	11	7	6	8	6	11	11	9	15	15	15
Sensitivity	61.1	60.6	70.1	63.6	70.3	60.7	64.1	58.3	62.3	62.9	68.5	64.4
Specificity	91.1	90.9	93.4	92.1	93.4	91.1	91.8	90.7	91.7	91.5	92.8	92.5
Accuracy	78.1	77.2	79.7	79.3	79.8	77.8	79.2	78.1	77.8	78.7	79.4	79.2

Table 26. DSVM classifier results for the total feature set.

	ReliefF		**mRMR-MID**		**mRMR-MIQ**		**Fisher**		**Chi-square**		**IG**	
	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max	STD	Min-Max
# Features	11	9	8	13	8	11	9	14	9	14	9	15
Sensitivity	79.3	76	81.6	79.8	80.6	80.5	75.1	76.3	75.3	73.8	77.5	76.5
Specificity	95.5	94.9	96.5	96.3	96.1	96	94.6	95.3	94.6	94.3	94.9	94.8
Accuracy	84.8	82.0	84.4	85.5	84.7	85.3	81.3	81.9	80.8	81.6	80.8	81.7

As mentioned before, to perform a comprehensive analysis and compare the discrimination ability of conventional and distance-based feature sets, independent t-tests were applied on the top 25 features of the total feature set (according to Table 17) with standardization and Min-Max methods. The significance level (α-value) for the t-test was chosen to be 0.05, which is a common value. Tables 27 and 28 present the results. In these tables, two categories of features are noticeable, namely “Discriminative” and “Redundant”. These categories are defined as:

- Discriminative: features with the highest discrimination ability between corresponding pairs of stages were included in this

category. From the perspective of the t-test results, features with the lowest p-value were categorized as “Discriminative” features.

- Redundant: features that cannot discriminate between corresponding pairs of stages were included in this category. From the perspective of the t-test results, features with a p-value of more than 0.05 were categorized as “Redundant” features.

Table 27. Discrimination ability analysis results for the top 25 features selected from the total feature set with standardization

	“Discriminative” Features	**“Redundant” Features**
Wake-REM	**F13**, F15, F18, F21, F53, F54, F55, F56.	F6, F31, F41, F61, F62, F67, F68.
Wake-N1	**F13**, F15, F18, F21, F25, F34, F45, F46.	F6, F29, F41, F49, F50.
Wake-N2	F9, F13, **F15**, F18, **F21**.	F6, F23, F30, F33.
Wake-N3	F9, F13, F15, **F18**, F65, F66.	F2, F6.
REM-N1	F13, F15, F18, F21, **F53**, **F54**, F55, F56.	F5, F6, F19, F22, F41.
REM-N2	F2, F4, F23, F26, **F34**, F35, F53, F54, F55, F56, F65, F66.	F6, F41, F51, F52, F73.
REM-N3	F2, **F4**, F5, F9, F11, F15, F18, F19, F21, F22, F23, F28, F29, F31, F36, F65, F66.	F10, F27, F36, F41, F46, F61, F62.
N1-N2	F4, F9, **F11**, F15, F18, F23, F29, F34, F35.	F6, F36, F45, F46, F55, F56.
N1-N3	F4, F5, F9, **F11**, F15, F16, F18, F19, F21, F22, F23, F28, F29, F30, F31, F33, F34, F35, F49, F50, F65, F66.	F26, F36, F39, F41.
N2-N3	F4, F5, F9, F11, F15, **F18**, F21, F29, F30, F31, F33, F34, F35, F46.	F25, F36, F37, F39, F41, F61, F62.

Table 28. Discrimination ability analysis results for the top 25 features selected from the total feature set with min-max

	"Discriminative" Features	**"Redundant" Features**
Wake-REM	F9, **F13**, F15, F18, F21, F53, F54.	F6, F31, F43, F47, F48, F67, F68.
Wake-N1	F13, F15, F18, F21, **F25**, F34, F45.	F6, F29, F41, F43, F47, F49, F50.
Wake-N2	F9, **F15**, F18, F21.	F7, F10, F30, F33, F43, F47.
Wake-N3	F9, **F15**, F18, F21, F29, F65, F66.	F2, F3, F6, F24, F43, F47.
REM-N1	F13, F15, F21, **F52**, **F53.**	F6, F22, F41, F43, F47, F51, F52, F71, F72, F73.
REM-N2	F2, F3, F10, F34, F35, F53, F54, **F65**, **F66**.	F11, F21, F32, F41, F43, F47.
REM-N3	F2, F3, F4, F9, F15, F18, F21, F34, F35, **F65**, **F66**.	F6, F10, F42, F43, F47.
N1-N2	F9, F13, F21.	F6, F12, F20, F25, F43, F45, F46, F47, F48, F51, F52, F73.
N1-N3	F4, F9, F15, **F18**, F21, F29, F30, F31, F32, F34, F35.	F6, F26, F43, F47.
N2-N3	F15, **F18**, F21.	F25, F41, F43, F47.

The features with highest discrimination ability (minimum *p*-value) are shown in bold. Assessment of the results in Tables 27 and 28 leads to the following observations:

- The minimum number of "Redundant" group features is related to the Wake-N3 pair with two features in the standardization method.
- The maximum number of "Redundant" group features is related to the N1-N2 pair with 11 features in the Min-Max method.
- The maximum number of "Discriminative" group features is related to the N1-N3 pair with 22 features in the standardization method.
- The minimum number of "Discriminative" group features is related to the N2-N3 pair with three features in the Min-Max method.

- There were some features in the Min-Max method that could not distinguish between any of the sleep stage pairs and were always categorized in the “Redundant” group, such as F43 and F47.
- There were some features that could always distinguish between any pair of sleep stages and were always categorized in the “Discriminative” group. For the standardization method, these features were: F4, F7, F9, F11, F13, F15, F18, F21, F28, F32, F34, F35, F44, F53, F54, F57, F58, F65 and F66 (19 features in total). The distance-based features constitute 31% of these features. For the Min-Max method, the features always categorized as “Discriminative” include: F4, F5, F9, F13, F15, F16, F18, F19, F27, F28, F34, F35, F36, F37, F39, F44, F53, F54, F57, F58, F65, F66, F69, and F70 (24 features in total). The distance-based features constitute 33% of these features
- Among distance-based features, the Itakura distance of EEG-EOG (F65 and F66) has the highest discrimination ability for both normalization methods.

5.2.2 Automatic EOG and EMG Artefact Removal Method for Sleep Stage Classification

Single channel sleep stage classification systems are often developed based on the signal acquired from one EEG channel. On the other hand, feature vector quality is dependant not only on the type of the features extracted, but also on the raw signal quality. It is crucial to be confident about the quality of the signal before applying any feature extraction or selection algorithm. EEG is usually contaminated with several artefacts such as power line noise, EMG, EOG, electrode movements, sweating noise,

etc. Therefore, removal or attenuation of the noise and unwanted signals is a prerequisite.

The basics for the artefact removing are diverse and are closely related to the specific application in which the algorithm is going to be used. A commonly used method for avoiding artefacts is the rejection of the contaminated segments of the recorded EEG [179]. This method although simple, results in huge data loss. Instead, denoising the contaminated EEG segments would not only preserve the amount of data, but also would probably contribute to the increase of accuracy in the automatic sleep stage classification [180].

We proposed a new method for EEG artefact removing for sleep stage classification. Rather than other works that used artificial noise, we used real EEG data contaminated with EOG and EMG for evaluating the proposed method. The artefact detection was performed by thresholding the EEG-EOG and EEG-EMG cross correlation coefficients. Then, the segments considered contaminated were denoised by normalized least-mean squares (NLMS) adaptive filtering technique. Using a single EEG channel, four sleep stages consisting of Awake, N1 + REM, N2 and N3 were classified.

5.2.2.1 Methodology

Figure 29 shows the block diagram of the sleep stage classification framework with the proposed EEG artefact removal scheme.

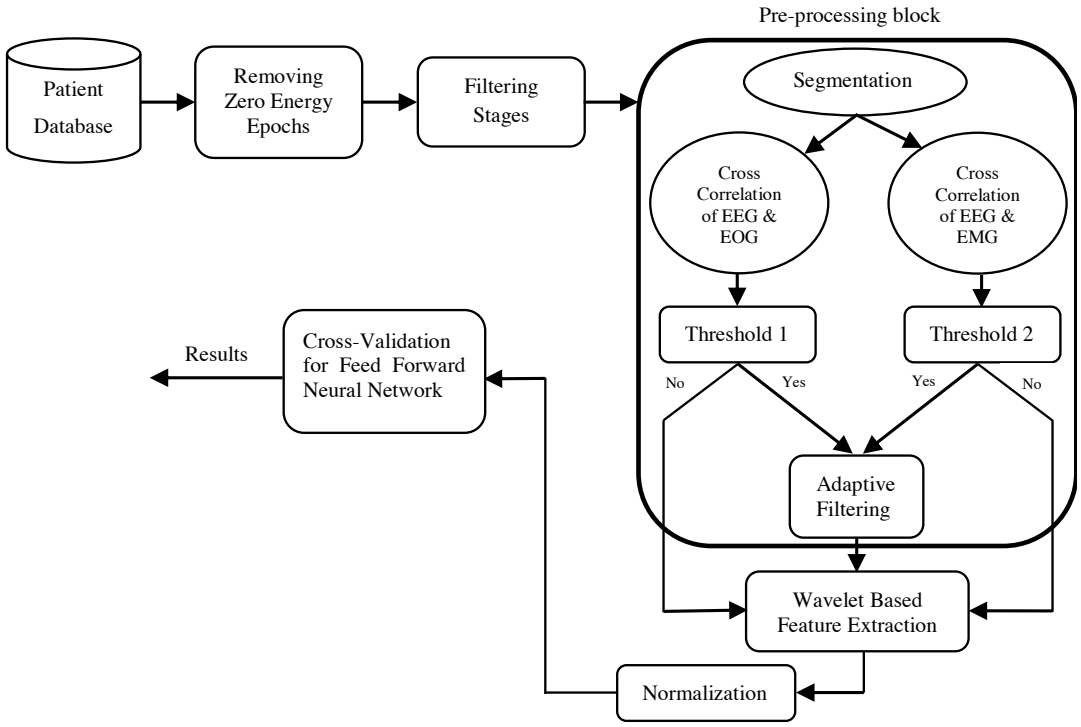

Figure 29. Block diagram of the sleep stage classification framework with the proposed EEG artefact removal scheme.

In this work, data was acquired from The Physionet Sleep-EDF database [Expanded] which includes records of 22 Caucasian males and females with the duration of nine hours. Pz-Oz EEG channel, horizontal EOG and submental chin EMG recordings of all the subjects were used for evaluation of the proposed method. In this study for reducing the artefacts, and guarantee the reliability of the classification results, all three steps of pre-processing, including band pass filtering, windowing and trimming, described in chapter 4 were applied to the selected PSG subset. For the WP-based filtering, Daubechies order 20 (db20) was used as mother wavelet.

Conventionally, it is assumed that the measured EEG is a linear combination of cerebral activity with one or more kinds of artefacts. Thus for detecting the EOG and EMG contamination, the filtered EEG, EOG and EMG recordings were divide into 1000-sample segments and then the cross

correlation of each EEG segment was calculated with the corresponding EOG and EMG segment. If the absolute value of the EEG-EOG cross correlation coefficients or EEG-EMG cross correlation coefficients was more than threshold 1 or threshold 2 respectively, the corresponding segment would be fed to an artefact removal block which was based on NLMS adaptive filtering. Adaptive filtering [181] has been extensively used in EEG artefact removal algorithms. It uses a recorded reference of the artefact (in our case horizontal EOG and submental chin EMG) to adjust a vector of weights that models the contamination according to an optimization algorithm.

On the other hand, if the thresholding conditions for cross correlation coefficients were not satisfied, the relevant EEG segment would be copied to the output without any change.

In order to perform sleep stage classification, the output of the pre-processing block was fed to feature extraction block. A WP tree with 7 decomposition levels and Daubechies order 2 (db2) mother wavelet was used for feature extraction. Different frequency bands of EEG including Delta, Theta, Alpha, spindle, Beta1 and Beta 2 were extracted from WP coefficients according to the scheme proposed in [52]. The following statistical features were calculated for each epoch using the WP coefficients:

- Energy of the WP coefficients for each frequency band (F22-F27 according to Table 7, chapter 4)
- Total Energy (F28 according to Table 7, chapter 4)
- Mean of the absolute values of WP coefficients for all frequency bands (F34 according to Table 7, chapter 4)
- Standard deviation of WP coefficients for all frequency bands (F35 according to Table 7, chapter 4)

- Energy ratio of various frequency bands (F29 to F33 according to Table 7, chapter 4)

Next, the extracted features were normalized to have zero mean and unit variance. In this study for classification of stages, MLF neural network was used. The two-layer feed forward network consisting of 14 input neurons, 12 hidden neurons and 4 output neurons for discrimination between the four sleep stages Wake, REM+N1, N2 and N3 was used. A sigmoid transfer function in the hidden layer and a linear transfer function in the output layer were selected. Levenberg-Marquardt training algorithm was chosen to train the network.

5.2.2.2 Results

The performance of the proposed method was assessed using the six subjects selected from the dataset. In the artefact detection stage, a threshold of 0.5 (*Threshold 1*) for EEG-EOG cross correlation coefficients and 0.25 (*Threshold 2*) for EEG-EMG cross correlation coefficients were selected. These thresholds were selected empirically considering the highest classification accuracy. Three different result validation approaches including subjective and objective methods were applied.

The cross-correlation coefficients for EEG-EOG and EEG-EMG which were detected by thresholding before and after applying the artefact removal algorithm are shown in Figure 30. A significant reduction in the correlation coefficients is noticeable after artefact removal.

Figures 31 and 32 illustrate the cancellation of EOG and EMG artefacts from contaminated EEG segments. It can be seen that the artefacts can be correctly eliminated without distorting the original EEG.

After the completion of the artefact removal stage, the data is fed to the feature extraction algorithm. For training MLF neural network, unlike the

conventional approaches in the literature, which all the existing stages to the neural network are imported, we used a quantity of training data selected out from each patient's data. This method is suitable for large databases helping on the reduction of the computational complexity of the classifier training stage.

To assess the effectiveness of our artefact removal algorithm, we studied the sleep stage classification accuracy for raw (after removing zero energy epochs), filtered and artefact removed data. Table 29 shows the results of statistical analysis for comparison of each stage and overall accuracy for all the above-mentioned data. The results are validated using repeated random sub-sampling method which is also known as Monte Carlo cross-validation technique. It is observed that there is an improvement in the performance of the classifier after filtering the data, but the best performance is achieved by applying the proposed artefact removal algorithm.

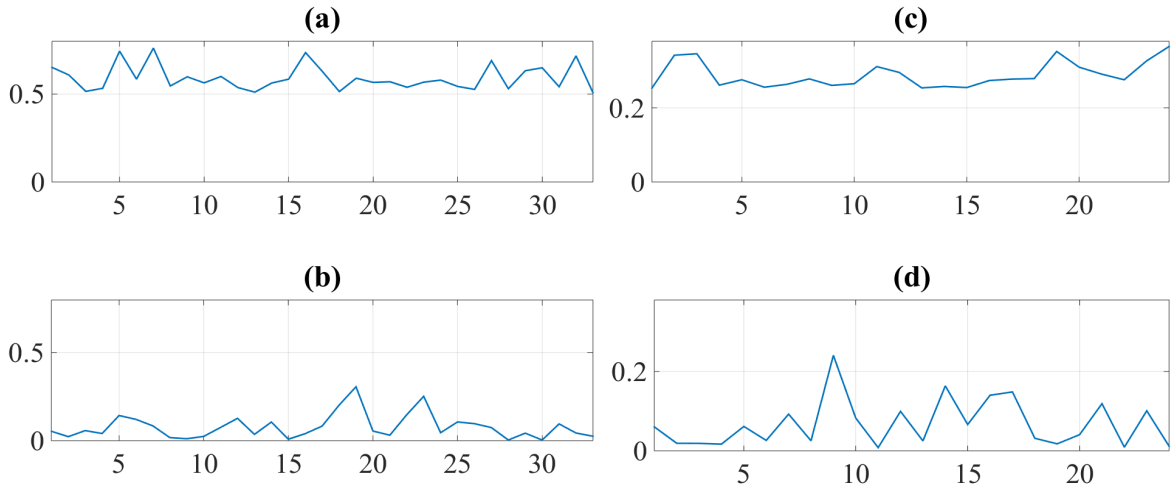

Figure 30. Absolute value of cross correlation coefficients, (a) EEG-EOG before artefact removal, (b) EEG-EOG after artefact removal, (c) EEG-EMG before artefact removal, (d) EEG-EMG after artefact removal algorithm.

Table 29. Results of the statistical analysis for comparison of each stage and overall accuracy.

	Wake (%)	REM + N1 (%)	N2 (%)	N3 (%)	Overall (%)
Raw	77.56	87.08	74.67	78.11	63.70
Filtered	79.44	78.75	83.26	90.74	70.60
Proposed Method	87.08	87.25	87.38	90.93	77.80

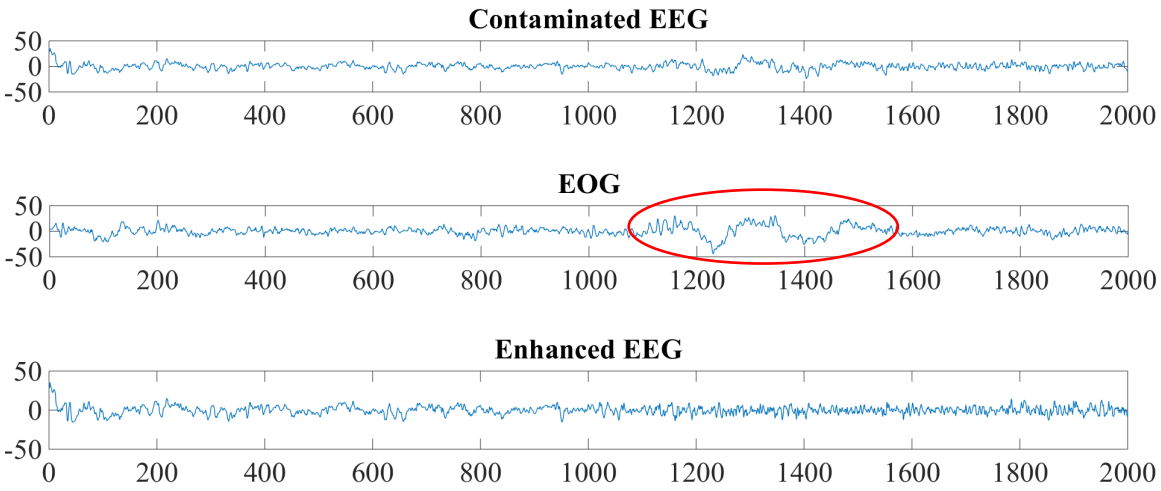

Figure 31. EOG artefact cancelation from contaminated EEG.

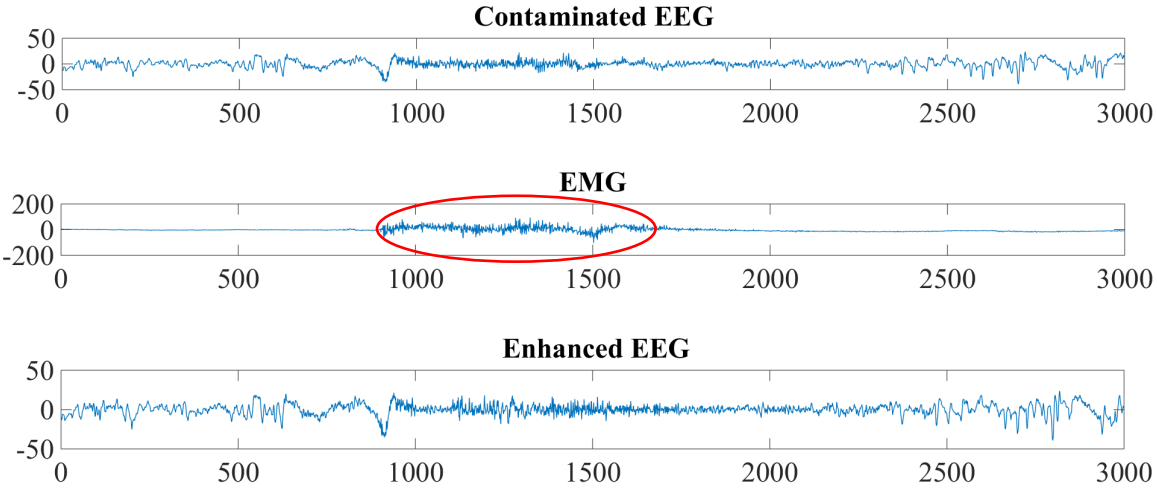

Figure 32. EMG artefact cancelation from contaminated EEG.

5.3 Summary

In this chapter, the four main contributions of this thesis work in feature extraction and selection were described. For each contribution, the corresponding experimental setup details and results were presented. In the next chapter, the obtained results will be interpreted and compared to the state of the art results.

Chapter 6

6. Discussion and Conclusion

As mentioned in chapter1, in this thesis the main goal was identifying a robust and reliable feature set that can lead to efficient classification of sleep stages. For achieving this goal, three types of contributions were introduced in the following areas: feature selection, feature extraction including feature vector quality enhancement. All three contributions are aligned with the proposed hypothesis presented in chapter 1.

In this chapter, the obtained results will be interpreted and compared to similar studies. Also, the significance and limitations of each work will be described. Finally, we will summarize the contributions of this thesis and discuss some suggestions for directions of future work.

6.1 Discussion

In this thesis two main contributions were made for the feature selection step of automatic sleep stage classification. First, two rank aggregation methods, namely Borda and RRA were applied to a set of 49 conventional features. Originally common in bioinformatics, rank aggregation methods are believed to be robust through the broad variety of classifiers and produce comparable classification accuracy to the individual feature selection methods. In our work, their performance was extensively compared to seven

different feature ranking methods using stability, similarity and accuracy criteria.

The stability analysis results (Figure 21 and Table 9) show that Fisher method has the highest stability and the CMIM method is the least stable one. Also, the stability of Chi-square and IG methods seemed very convergent. Although the stability of rank aggregation methods was comparable to the conventional feature ranking techniques, none of them could outperform the conventional methods. This result is reasonable, since both of the selected rank aggregation methods were calculated in a way that almost all of the ranking techniques affected them equally. Therefore, the archieved stability is an average of overall stability.

There existed a huge reduction in stability for MRMR_MID, MRMR_MIQ and ReliefF for three-feature subset. On the other hand, both MRMR methods were always 100% stable in selecting the first feature which was the Hurst Exponent. It means that the Hurst Exponent has the highest discrimination ability from the MRMR methods point of view. Also, the Fisher method had 100% stability for the three-feature and the five-feature subsets (ID, Hurst exponent, Petrosian fractal dimension as three-feature group and ID, Hurst exponent, Petrosian fractal dimension, zero-crossing rate and approximate entropy as five-feature group). Considering thirteen features, Fisher method was almost totally stable (99.92%). Finally, for twenty-nine features, IG outperformed other methods from the mean stability point of view.

In similarity analysis (Table 11), Chi-square and IG pair and MRMR-MID and MRMR-MIQ pair generated highly similar results. The similarity of MRMR methods can be explained by their similar theoretical background.

The average similarity of Borda and RRA with other methods was approximately 0.5 with the other methods. Regarding the aggregation characteristics it was predictable.

Table 10 illustrates the top 10 features selected by each method. As it can be seen, Itakura spectral distance (F36) always appeared in the top 10 for all the methods. In spite of the fact that different feature ranking methods have their own specific criteria for ranking the features, observing ISD in the top 10 list, means that ISD is a preferable feature for all the feature selection methods. In addition to ISD, there were some other features that can be considered most preferable. EEG ZCR (F18) is a simple, yet effective feature that is listed in top 10 by all methods except ReliefF. Following ZCR, Petrosian fractal dimension (F9), Hurst exponent (F21), WP feature (F22), approximate entropy (F13), spectral entropy (F11), and Hjorth mobility parameter (F15) were selected by at least five ranking methods to be included in top 10 list.

The optimum number of features for each method, selected by the Kneedle algorithm, is also presented in Table 10. For MLF neural network and kNN classifiers, a slight difference existed in the optimum number. Considering the maximum accuracy reached by different methods in their optimum points, the MRMR-MID method using kNN classifier outperformed all the others with seven selected features. For MLF neural network, both MRMR methods outperformed all the other methods with five features. None of the aggregation methods showed better performance than the rest of the feature ranking methods.

Considering the obtained results, although mRMR method outperformed others from the classification accuracy point of view, the most stable feature set was generated by Fisher. Moreover, CMIM method needed the minimum number of features (3 features) to reach its optimum accuracy. It can be

concluded that selection of the feature ranking method is dependent on the system requirements that one has, such as highest accuracy/stability or minimum computational complexity. Regarding the poor performance of the rank aggregation methods, it should be noted that only two of many available rank aggregation methods were evaluated in this work. Both of these methods evaluated, follow the concept of averaging the results from different methods and therefore generate results that are reflecting the characteristics of all methods from the best to the poorest.

Our second contribution in feature selection was the application of SSAE for feature transformation and dimensionality reduction in sleep stage classification. The main advantage of using a dimensionality reduction method like SSAE is that these kinds of methods are unsupervised and no information about groups is used in dimension reduction. In addition, because of its theoretical and mathematical structure which is related to deep learning, SSAE is able to learn and generate meaningful and efficient representation of the input feature set.

According to Table 12, It is noticeable that the combination of SSAE method and Softmax classifier outperformed the other two classifiers in terms of overall accuracy. Also, for the individual sleep stages, in most of the cases SSAE discriminated the stages better. In addition to the higher performance, SSAE provided a significant reduction in the dimension of the feature vector. Considering that the second layer of SSAE had 12 hidden units, it succeeded to decrease the dimension from 37 to 12, which means 67% reduction. Therefore, it is a powerful tool to generate more descriptive features from original feature vector.

However, it should be noted that dimension reduction methods such as PCA, KDR and SSAE impose a limitation to the overall system. This limitation arises from the fact that it is essential to keep and calculate all

the features in the classification and practical application steps, because these methods use all the feature vector to generate useful representations while this is not the case in feature ranking methods.

Regarding feature extraction, the main contribution of this thesis work was the application and evaluation of a distance-based set of features which were originally used in speech signal processing. The performance of the distance-based feature set along with 48 conventional temporals, frequency domain, time-frequency domain, non-linear, and entropy-based features were evaluated in sleep stage classification.

Similar features were removed from the feature sets by thresholding L1-norm between feature vectors. This step was advantageous because removing these features reduces the final feature vector dimensionality and enhances the stability of feature-ranking results. Moreover, according to the results of Table 14, this step led to an improvement in the classification accuracy. This improvement was expected since the existence of redundant features has no positive effects on the classification results and increases the computational complexity of the whole system. Regarding the threshold value, although in our work it was chosen empirically, it is better to use a systematic threshold search method for an optimium parameter selection.

After removing similar features, feature ranking was applied. According to the obtained results, from the conventional feature set, EEG zero-crossing rate was selected as the best feature by most of the ranking methods. In addition to the zero-crossing rate, Petrosian fractal dimension, Hjorth mobility parameter, and Hurst exponent were always among the top-ranked features. This validates the outstanding performance of these features already demonstrated in previous studies such as [25] and also our study on feature rank aggregation.

In [70], [72], it had been shown that the Itakura distance between EEG and EOG signals and also between a reference EEG epoch and other EEG epochs have meaningful variations in different sleep stages. In these studies, It was concluded that these measures can be used as useful features in automated sleep staging systems and our simulations confirmed this conclusion. According to the results, all the ranking methods listed EEG Itakura distance, EEG-EOG Itakura and Itakura-Saito distances in their top 25 features. Moreover, the features related to the similarity of a baseline EOG/EMG epoch to the rest of the EOG/EMG were always among the top 25 features.

The ranking results for the total feature set in Table 17 show that the top 25 features for all the ranking methods include features from both conventional and distance-based sets. This fact implies that a combination of features from different domains yields better results. According to this table, distance-based features occupy 28% of the top-ranked features.

To further investigate the contribution of distance-based features, three different classifiers, kNN, MLF neural network and DSVM, were used. Previous studies [30], [82] showed that combining different types of features, i.e. temporal, spectral, time-frequency domain and nonlinear, would lead to a satisfactory level of classification accuracy with a fewer number of features. In this work, we showed that using distance-based features together with conventional ones can further improve the performance of the sleep scoring system. This improvement is noticeable in the results of all three classifiers. According to the results of the Vikor method, 8-13 carefully selected measures from the total feature set were sufficient to reach, on average, 85% accuracy, and usually three of these features are from the distance-based category. The only method that listed conventional features higher in rank than distance-based features is the ReliefF method.

Specifically, with Min-Max normalization, this method had its first distance-based feature ranked 18th.

According to the literature [182], there has been a lack of discriminative features for distinguishing N1 stage from other sleep stages because neurophysiological signals of N1 and N2 have similarities with each other as well as other sleep stages [65]. For example, the PSG recordings show similar wave patterns in REM and N1 in EEG, both having low amplitude waves of 3-7 Hz [183]. Therefore, the accuracy obtained on the classification of the N1 stage is usually lower other stages. Especially, discriminating N1 from REM is challenging. To tackle this challenge and increase the discrimination ability of the overall system, other channels (EOG, EMG and ECG) along with EEG are usually used [66], [82], [133]. In this work, the ability of the features to discriminate between each pair of sleep stages was assessed using two-tailed student's t-test applied on the total feature set. The t-test results show that distance-based features outperform conventional features in discriminating between N1 and REM stages. According to Tables 27 and 28, the Itakura-Saito distance of EEG spectral coefficients (F52) and Itakura distance of EMG spectral and AR coefficients (F53 and F54) have outstanding performances in distinguishing N1 from REM stage, regardless of the feature normalization method. Therefore, these features can be appropriate choices to be included in the sleep stage classification feature set to increase the system's discrimination ability of the system. Regarding the effect of feature normalization on the overall performance, results show that the Min-Max method outperforms standardization. In other words, the accuracy achieved with the data normalized by Min-Max turned out to be higher than the accuracy achieved with standardization. To obtain a more general conclusion, the effect of feature normalization should be evaluated with different sleep databases.

Our last contribution was related to the enhancement of feature vector quality by adaptive removal of the EEG artefact. Specifically, in this thesis, we focused on the EMG and EOG artefacts on EEG signal. According to Figure 30, absolute cross correlation showed significant reduction after applying the proposed artefact removal technique. This enhancement was further confirmed by the classification accuracy results. According to Table 30, although filtering the signals according to AASM manual recommendations improved the accuracy, the major improvement was due to the artefact removal, especially in Wake, N2 and N3 stages.

Despite the obtained positive outcomes, it should be noted that the proposed method is more suitable for removing linear artefacts. In other words, since cross-correlation detects linear relationships between signals, it is not capable of detecting nonlinear correlations.

6.2 Conclusion and Future Work

Sleep quality is one of the most important measures of healthy life, especially considering the huge number of sleep-related disorders. Identifying sleep stages using multi-channel recordings like PSG signals is an effective way of assessing sleep quality. However, manual sleep stage classification is time-consuming, tedious and highly subjective. To overcome these hurdles, automatic sleep classification was proposed, in which pre-processing, feature extraction and classification are the three main steps. Proper feature extraction and selection play an important role in the automatic sleep scoring process and has undeniable effect on final classification results. Besides the significant amount of work done in this area, there are still challenges that need to be addressed. In this thesis, we tried to address some of these challenges by proposing solutions for feature selection, feature extraction and artefact removal of PSG signals. Also,

several different evaluation criteria were used to assess the effectiveness of the proposed methods. The following conclusions can be drawn from the obtained results:

- Regarding feature selection and considering that in this thesis, several feature ranking and rank aggregation methods were evaluated and compared, it can be concluded that MRMR methods outperformed other feature selection methods considering the evaluation criteria. However, the decision on the precise feature selection method depends on the system design requirements such as low computational complexity, high stability or high classification accuracy.
- In addition to conventional feature transformation and dimensionality reduction methods, novel methods such as SSAE were proposed in this thesis and showed promising performance.
- In addition to wide range of features used in automatic sleep stage classification, new and effective features such as distance-based features contribute positively to the classification performance.
- New Effective and loss-less enhancement of raw signal quality is crucial for achieving high final classification accuracy. The proposed adaptive artefact removal method allowed 14% enhancement in overall accuracy.
- Min-Max normalisation outperformed standardisation.

In this thesis, the evaluation of the sleep stage classification systems was done based on the hypnograms provided by the databases creators. The use of these hypnograms imposed some limitations to our work. For example, in ISRUC database the available hypnograms were created from the consensus of two experts on visual sleep scoring. There were some cases of interscorer

variability, especially on N1. Moreover, the database was pre-processed, and raw data was not available for possible change in pre-processing step.

Imbalanced data and few N1 stage epochs were other limitations of this thesis work. In normal human sleep hypnogram, different sleep stages are not presented equally, especially because there is always a shortage for N1 stage. Therefore, the stage-wise classification accuracy is usually low for N1 stage and this negatively affects the overall classification accuracy.

Future work for this thesis can include:

- Developing a selective aggregation method that incorporates only the most effective ranking methods will be desirable;
- The comparison of the SSAE-based feature transformation with conventional methods and parameter adjustment;
- Confirming the positive contributions of the distance-based features using other sleep datasets;
- Extending the applications of suggested adaptive artefact removal algorithm for nonlinear artefact;
- Developing a prototype for automatic sleep stage classification software

References

[1] S. Chokroverty, *Sleep Disorders Medicine: Basic Science, Technical Considerations, and Clinical Aspects*. Saunders/Elsevier, 2009.

[2] R. B. Berry *et al.*, *AASM - Manual for the Scoring of Sleep and Associated Events version 2.1*. 2014.

[3] T. L. Lee-Chiong, *Sleep: A Comprehensive Handbook*. Wiley, 2005.

[4] N. Sukhorukova, A. Stranieri, and B. Ofoghi, "Automatic sleep stage identification: difficulties and possible solutions," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 108, no. Hikm, pp. 39–44, 2010.

[5] H. Danker-Hopfe *et al.*, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard.," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, Mar. 2009.

[6] C. Pollak, M. J. Thorpy, and J. Yager, *The Encyclopedia of Sleep and Sleep Disorders*. Facts on File, 2010.

[7] B. Högl, C. L. Comella, and H. R. Smith, Eds., "Index," in *Sleep Medicine*, Cambridge: Cambridge University Press, 2008, pp. 256–270.

[8] C. L. Nunn, D. R. Samson, and A. D. Krystal, "Shining evolutionary light on human sleep and sleep disorders," *Evolution, Medicine and Public Health*, vol. 2016, no. 1. pp. 227–243, 2016.

[9] J. D. Geyer, P. R. Carney, and T. A. Payne, *Atlas of Polysomnography*. Lippincott Williams & Wilkins, 2010.

[10] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, and Y.-H. Wang, *Automatic Stage Scoring of Single-Channel Sleep EEG by Using Multiscale Entropy and Autoregressive Models*, vol. 61. 2012.

[11] B. Robertson, B. Marshall, and M. A. Carno, *Polysomnography for the Sleep Technologist: Instrumentation, Monitoring, and Related Procedures*. Elsevier Health Sciences, 2014.

[12] S. T.-B. Hamida and B. Ahmed, "Computer Based Sleep Staging: Challenges for the Future," in *2013 7th IEEE GCC Conference and Exhibition (GCC)*, 2013, pp. 280–285.

[13] "Sleep and sleep pharmacology | Clinical Gate." [Online]. Available: https://clinicalgate.com/sleep-and-sleep-pharmacology/. [Accessed: 23-Apr-2018].

[14] M. R. Nuwer *et al.*, "IFCN standards for digital recording of clinical EEG. International Federation of Clinical Neurophysiology.," *Electroencephalogr. Clin. Neurophysiol.*, vol. 106, no. 3, pp. 259–61, Mar. 1998.

[15] J. N. Acharya, A. Hani, J. Cheek, P. Thirumala, and T. N. Tsuchida, "American Clinical Neurophysiology Society Guideline 2," *J. Clin. Neurophysiol.*, vol. 33, no. 4, pp. 308–311, Aug. 2016.

[16] J. Kim, "A Comparative Study on Classification Methods of Sleep Stages by Using EEG," *J. Korea Multimed. Soc.*, vol. 17, no. 2, pp. 113–123, Feb. 2014.

[17] M. Peker, "A new approach for automatic sleep scoring: Combining Taguchi based complex-valued neural network and complex wavelet transform," *Comput. Methods Programs Biomed.*, vol. 129, pp. 203–216, Jun. 2016.

[18] A. Subasi, M. K. Kiymik, M. Akin, and O. Erogul, "Automatic recognition of vigilance state by using a wavelet-based artificial neural network," *Neural Comput. Appl.*, vol. 14, no. 1, pp. 45–55, Mar. 2005.

[19] M. E. Tagluk, N. Sezgin, and M. Akin, "Estimation of Sleep Stages by an Artificial Neural Network Employing EEG, EMG and EOG," *J. Med. Syst.*, vol. 34, no. 4, pp. 717–725, Aug. 2010.

[20] A. R. Hassan and M. I. H. Bhuiyan, "An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting," *Neurocomputing*, vol. 219, pp. 76–87, 2017.

[21] A. R. Hassan and M. I. H. Bhuiyan, "Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting," *Comput. Methods Programs Biomed.*, vol. 140, pp. 201–210, 2017.

[22] M. Diykh and Y. Li, "Complex networks approach for EEG signal sleep stages classification," *Expert Syst. Appl.*, vol. 63, pp. 241–248, 2016.

[23] M. Diykh, Y. Li, and P. Wen, "EEG sleep stages classification based on time domain features and structural graph similarity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 11, pp. 1159–1168, 2016.

[24] S. Mahvash Mohammadi, S. Kouchaki, M. Ghavami, and S. Sanei, "Improving time–frequency domain sleep EEG classification via singular spectrum analysis," *J. Neurosci. Methods*, vol. 273, pp. 96–106, 2016.

[25] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and Classification Algorithms," *J. Med. Syst.*, vol. 38, no. 3, p. 18, Mar. 2014.

[26] N. Burioka *et al.*, "Approximate entropy in the electroencephalogram during wake and sleep.," *Clin. EEG Neurosci.*, vol. 36, no. 1, pp. 21–24, 2005.

[27] M. Obayya and F. E. Z. Abou-Chadi, "Automatic classification of sleep stages using EEG records based on Fuzzy c-means (FCM) algorithm," in *2014 31st National Radio Science Conference (NRSC)*, 2014, pp. 265–272.

[28] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, "Classification of Sleep Stages Using Multi-wavelet Time Frequency Entropy and LDA," *Methods Inf. Med.*, vol. 49, no. 3, pp. 230–237,

Jan. 2010.

[29] L. J. Herrera *et al.*, "Combination of Heterogeneous EEG Feature Extraction Methods and Stacked Sequential Learning for Sleep Stage Classification," *Int. J. Neural Syst.*, vol. 23, no. 03, p. 1350012, Jun. 2013.

[30] M. Radha, G. Garcia-Molina, M. Poel, and G. Tononi, "Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 1876–1880.

[31] H. G. Jo, J. Y. Park, C. K. Lee, S. K. An, and S. K. Yoo, "Genetic fuzzy classifier for sleep stage identification," *Comput. Biol. Med.*, vol. 40, no. 7, pp. 629–634, 2010.

[32] L. J. Herrera, a. M. Mora, and C. M. Fernandes, "Symbolic Representation of the EEG for Sleep Stage Classification," in *11th International Conference on Intelligent Systems Design and Applications*, 2011, pp. 253–258.

[33] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Comput. Biol. Med.*, vol. 42, no. 12, pp. 1186–1195, Dec. 2012.

[34] S. Güneş, K. Polat, and Ş. Yosunkaya, "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7922–7928, Dec. 2010.

[35] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier," *Comput. Methods Programs Biomed.*, vol. 108, no. 1, pp. 10–19, Oct. 2012.

[36] Yi Li, Fan Yingle, Li Gu, and Tong Qinye, "Sleep stage classification based on EEG Hilbert-Huang transform," in *2009 4th IEEE Conference on Industrial*

Electronics and Applications, 2009, pp. 3676–3681.

[37] T. H. Sanders, M. McCurry, and M. a Clements, "Sleep stage classification with cross frequency coupling," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, vol. 2014, pp. 4579–4582.

[38] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Comput. Methods Programs Biomed.*, vol. 140, pp. 77–91, 2017.

[39] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, 2016.

[40] P. Tian *et al.*, "A hierarchical classification method for automatic sleep scoring using multiscale entropy features and proportion information of sleep architecture," *Biocybern. Biomed. Eng.*, vol. 37, no. 2, pp. 263–271, 2017.

[41] A. Ouanes and L. Rejeb, "A Hybrid Approach for Sleep Stages Classification," *Proc. 2016 Genet. Evol. Comput. Conf. - GECCO '16*, pp. 493–500, 2016.

[42] M. Peker, "An efficient sleep scoring system based on EEG signal using complex-valued machine learning algorithms," *Neurocomputing*, vol. 207, pp. 165–177, 2015.

[43] A. R. Hassan and M. I. Hassan Bhuiyan, "Automatic sleep scoring using statistical features in the EMD domain and ensemble methods," *Biocybern. Biomed. Eng.*, vol. 36, no. 1, pp. 248–255, 2016.

[44] Ö. F. Alçin, S. Siuly, V. Bajaj, Y. Guo, A. Şengu¨r, and Y. Zhang, "Multi-category EEG signal classification developing time-frequency texture

features based Fisher Vector encoding method," *Neurocomputing*, vol. 218, pp. 251–258, Dec. 2016.

[45] T. L. T. da Silveira, A. J. Kozakevicius, and C. R. Rodrigues, "Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain," *Med. Biol. Eng. Comput.*, vol. 55, no. 2, pp. 343–352, 2017.

[46] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, "Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation," *Entropy*, vol. 18, no. 9, p. 272, 2016.

[47] S. I. Dimitriadis, C. Salis, and D. Linden, "A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates," *Clin. Neurophysiol.*, vol. 129, no. 4, pp. 815–828, 2018.

[48] Y.-L. Hsu, Y.-T. Yang, J.-S. Wang, and C.-Y. Hsu, "Automatic sleep stage recurrent neural classifier using energy features of EEG signals," *Neurocomputing*, vol. 104, no. c, pp. 105–114, Mar. 2013.

[49] K. a. I. Aboalayon, H. T. Ocbagabir, and M. Faezipour, "Efficient sleep stage classification based on EEG signals," in *IEEE Long Island Systems, Applications and Technology (LISAT) Conference 2014*, 2014, pp. 1–6.

[50] T. Kayikcioglu, M. Maleki, and K. Eroglu, "Fast and accurate PLS-based classification of EEG sleep using single channel data," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7825–7830, Nov. 2015.

[51] V. Bajaj and R. B. Pachori, "Automatic classification of sleep stages based on the time-frequency image of EEG signals," *Comput. Methods Programs Biomed.*, vol. 112, no. 3, pp. 320–328, 2013.

[52] F. Ebrahimi, M. Mikaeili, E. Estrada, and H. Nazeran, "Automatic Sleep

Stage Classification Based on EEG Signals by Using Neural Networks and Wavelet Packet Coefficients," *2008 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2008, pp. 1151–1154, Aug. 2008.

[53] A. R. Hassan and M. I. H. Bhuiyan, "Computer-aided sleep staging using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and bootstrap aggregating," *Biomed. Signal Process. Control*, vol. 24, pp. 1–10, Feb. 2016.

[54] R. Acharya U., O. Faust, N. Kannathal, T. Chua, and S. Laxminarayan, "Non-linear analysis of EEG signals at various sleep stages," *Comput. Methods Programs Biomed.*, vol. 80, no. 1, pp. 37–45, Oct. 2005.

[55] J. Fell, J. Röschke, K. Mann, and C. Schäffner, "Discrimination of sleep stages: A comparison between spectral and nonlinear EEG measures," *Electroencephalography and Clinical Neurophysiology*, vol. 98, no. 5. pp. 401–410, 1996.

[56] R. Kaplan, Y. Wang, K. Loparo, M. Kelly, and R. Bootzin, "Performance evaluation of an automated single-channel sleep–wake detection algorithm," *Nat. Sci. Sleep*, vol. 6, p. 113, Oct. 2014.

[57] A. Pasieczna and J. Korczak, "Classification Algorithms in Sleep Detection—A Comparative Study," 2016, vol. 8, pp. 113–120.

[58] M. Elmessidi, S. T. Ben Hamida, B. Ahmed, and T. Penzel, "Accurate automatic identification of slow wave sleep using a single electro-oculogram channel," *Middle East Conf. Biomed. Eng. MECBME*, pp. 232–235, 2014.

[59] J. Virkkala, J. Toppila, P. Maasilta, and A. Bachour, "Electro-oculography-based detection of sleep-wake in sleep apnea patients," *Sleep Breath.*, vol. 19, no. 3, pp. 785–789, Sep. 2015.

[60] S. M. Isa, I. Wasito, A. M. Arymurthy, and A. Noviyanto, "Kernel

Dimensionality Reduction on Sleep Stage Classification using ECG Signal," *Int. J. Comput. Sci. Issues*, vol. 8, no. 1, pp. 1178–1181, 2011.

[61] A. Noviyanto and A. M. Arymurthy, "Sleep stages classification based on temporal pattern recognition in neural network approach," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–6.

[62] M. Adnane, Z. Jiang, and Z. Yan, "Sleep–wake stages classification and sleep efficiency estimation using single-lead electrocardiogram," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1401–1413, 2012.

[63] Q. K. Le, Q. D. K. Truong, and V. T. Vo, "A tool for analysis and classification of sleep stages," *2011 Int. Conf. Adv. Technol. Commun. (ATC 2011)*, no. Atc, pp. 307–310, 2011.

[64] S. Khalighi, T. Sousa, and U. Nunes, "Adaptive Automatic Sleep Stage Classification under Covariate Shift," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 2259–2262.

[65] S. Khalighi, T. Sousa, G. Pires, and U. Nunes, "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels," *Expert Syst. Appl.*, vol. 40, no. 17, pp. 7046–7059, 2013.

[66] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, and F. Chapotot, "Feature Selection for Sleep/Wake Stages Classification Using Data Driven Methods," *Biomed. Signal Process. Control*, vol. 2, no. 3, pp. 171–179, Jul. 2007.

[67] T. Lajnef *et al.*, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *J. Neurosci. Methods*, vol. 250, pp. 94–105, 2015.

[68] T. Sousa, A. Cruz, S. Khalighi, G. Pires, and U. Nunes, "A two-step automatic sleep stage classification method with dubious range detection,"

Comput. Biol. Med., vol. 59, pp. 42–53, Apr. 2015.

[69] S. Khalighi, T. Sousa, D. Oliveira, G. Pires, and U. Nunes, "Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, no. July, pp. 3306–3309.

[70] F. Ebrahimi, M. Mikaili, E. Estrada, and H. Nazeran, "Assessment of Itakura Distance as a Valuable Feature for Computer-aided Classification of Sleep Stages," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, vol. 2007, pp. 3300–3303.

[71] L. Chen, Y. Zhao, J. Zhang, and J. Zou, "Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7344–7355, Nov. 2015.

[72] E. Estrada, P. Nava, H. Nazeran, K. Behbehani, J. Burk, and E. Lucas, "Itakura Distance: A Useful Similarity Measure between EEG and EOG Signals in Computer-aided Classification of Sleep Stages.," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2, pp. 1189–1192, 2005.

[73] T. Willemen *et al.*, "An Evaluation of Cardiorespiratory and Movement Features With Respect to Sleep-Stage Classification," *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 2, pp. 661–669, Mar. 2014.

[74] V. C. Figueroa Helland *et al.*, "Investigation of an Automatic Sleep Stage Classification by Means of Multiscorer Hypnogram," *Methods Inf. Med.*, vol. 49, no. 5, pp. 467–472, 2010.

[75] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiol. Meas.*, vol. 36, no. 10, pp. 2027–2040, Oct. 2015.

[76] F. Chapotot and G. Becq, "Automated Sleep-Wake Staging Combining Robust Feature Extraction, Artificial Neural Network Classification, and Flexible Decision Rules," *Int. J. Adapt. Control Signal Process.*, vol. 24, no. 5, pp. 409–423, 2009.

[77] G. Becq, S. Charbonnier, F. Chapotot, A. Buguet, L. Bourdon, and P. Baconnier, "Comparison Between Five Classifiers for Automatic Scoring of Human Sleep Recordings," in *Classification and Clustering for Knowledge Discovery*, vol. 127, 2005, pp. 113–127.

[78] S. F. Liang, C. E. Kuo, F. Z. Shaw, Y. H. Chen, C. H. Hsu, and J. Y. Chen, "Combination of expert knowledge and a genetic fuzzy inference system for automatic sleep staging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 10, pp. 2108–2118, 2016.

[79] J. Foussier, P. Fonseca, X. Long, and S. Leonhardt, "Automatic Feature Selection for Sleep/Wake Classification with Small Data Sets," *6th Int. Conf. Bioinforma. Model. Methods Algorithms*, pp. 1–7, 2013.

[80] M. Čić, J. Šoda, and M. Bonković, "Automatic classification of infant sleep based on instantaneous frequencies in a single-channel EEG signal," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2110–2117, Dec. 2013.

[81] I. Zhovna and I. D. Shallom, "Automatic detection and classification of sleep stages by multichannel EEG signal modeling," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, vol. 2008, no. c, pp. 2665–2668.

[82] A. Krakovská and K. Mezeiová, "Automatic sleep scoring: A search for an optimal combination of measures," *Artif. Intell. Med.*, vol. 53, no. 1, pp. 25–33, Sep. 2011.

[83] J. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, D. Cuesta-

Frau, E. Cirugeda-Roldán, and D. Peluffo, "Automatic Sleep Stages Classification Using EEG Entropy Features and Unsupervised Pattern Analysis Techniques," *Entropy*, vol. 16, no. 12, pp. 6573–6589, Dec. 2014.

[84] L. G. Doroshenkov, V. A. Konyshev, and S. V. Selishchev, "Classification of human sleep stages based on EEG processing using hidden Markov models," *Biomed. Eng. (NY).*, vol. 41, no. 1, pp. 25–28, Jan. 2007.

[85] E. Estrada, H. Nazeran, F. Ebrahimi, and M. Mikaeili, "EEG signal features for computer-aided sleep stage detection," in *2009 4th International IEEE/EMBS Conference on Neural Engineering*, 2009, pp. 669–672.

[86] J. Shi, X. Liu, Y. Li, Q. Zhang, Y. Li, and S. Ying, "Multi-channel EEG-based sleep stage classification with joint collaborative representation and multiple kernel learning," *J. Neurosci. Methods*, vol. 254, pp. 94–101, Oct. 2015.

[87] H. Simões, G. Pires, U. Nunes, and V. Silva, "Feature Extraction and Selection for Automatic Sleep Staging Using EEG," in *ICINCO*, 2010, pp. 128–133.

[88] Z. Liu, J. Sun, Y. Zhang, and P. Rolfe, "Sleep staging from the EEG signal using multi-domain feature extraction," *Biomed. Signal Process. Control*, vol. 30, pp. 86–97, 2016.

[89] J. Virkkala, J. Hasan, A. Värri, S.-L. Himanen, and K. Müller, "Automatic sleep stage classification using two-channel electro-oculography," *J. Neurosci. Methods*, vol. 166, no. 1, pp. 109–115, Oct. 2007.

[90] S. Liang *et al.*, "Development of an EOG-Based Automatic Sleep-Monitoring Eye Mask," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 11, pp. 2977–2985, Nov. 2015.

[91] K. Šušmáková and A. Krakovská, "Discrimination ability of individual measures used in sleep stages classification," *Artif. Intell. Med.*, vol. 44, no. 3,

pp. 261–277, Nov. 2008.

[92] A. Piryatinska, W. A. Woyczynski, M. S. Scher, and K. A. Loparo, "Optimal channel selection for analysis of EEG-sleep patterns of neonates," *Comput. Methods Programs Biomed.*, vol. 106, no. 1, pp. 14–26, 2012.

[93] T. K. Padma Shri and N. Sriraam, "Comparison of t-test ranking with PCA and SEPCOR feature selection for wake and stage 1 sleep pattern recognition in multichannel electroencephalograms," *Biomed. Signal Process. Control*, vol. 31, pp. 499–512, 2017.

[94] A. Garcés Correa, L. Orosco, and E. Laciar, "Automatic detection of drowsiness in EEG records based on multimodal analysis," *Med. Eng. Phys.*, vol. 36, no. 2, pp. 244–249, Feb. 2014.

[95] S. Özşen, "Classification of sleep stages using class-dependent sequential feature selection and artificial neural network," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1239–1250, Oct. 2013.

[96] V. Bajaj and R. B. Pachori, "Automatic classification of sleep stages based on the time-frequency image of EEG signals," *Comput. Methods Programs Biomed.*, vol. 112, no. 3, pp. 320–328, Dec. 2013.

[97] A. R. Hassan and A. Subasi, "A decision support system for automated identification of sleep stages from single-channel EEG signals," *Knowledge-Based Syst.*, vol. 128, pp. 115–124, 2017.

[98] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed Neural Network Approach for Temporal Sleep Stage Classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, 2018.

[99] R. Sharma, R. B. Pachori, and A. Upadhyay, "Automatic sleep stages classification based on iterative filtering of electroencephalogram signals," *Neural Comput. Appl.*, vol. 28, no. 10, pp. 2959–2978, 2017.

[100] D. Görür, U. H. Halıcı, G. Ongun, F. Özgen, and K. Leblebicioğlu, "Sleep Spindles Detection Using Autoregressive Modeling," *Proc. ICANN/ICONIP*, 2003.

[101] H. Bozdogan, "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, Sep. 1987.

[102] R. J. Bhansali, "The Criterion Autoregressive Transfer function of PARZEN," *J. Time Ser. Anal.*, vol. 7, no. 2, pp. 79–104, Mar. 1986.

[103] B. Hjorth, "EEG Analysis Based on Time Domain Properties," *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, 1970.

[104] S. Najdi, A. A. Gharbali, and J. M. Fonseca, "A Comparison of Feature Ranking and Rank Aggregation Techniques in Automatic Sleep Stage Classification Based on Polysomnographic Signals," in *4th International Conference, IWBBIO*, 2016, pp. 230–241.

[105] P. Memar and F. Faradji, "A Novel Multi-Class EEG-Based Sleep Stage Classification System," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, 2018.

[106] M. Dursun, S. Gunes, S. Ozsen, and S. Yosunkaya, "Comparison of Artificial Immune Clustering with Fuzzy C-means Clustering in the sleep stage classification problem," in *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 2012, pp. 1–4.

[107] R. Broberg and R. Lewis, "Classification of epileptoid oscillations in EEG using Shannon's entropy amplitude probability distribution," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8821, pp. 247–252.

[108] K. Alsharabi, S. Ibrahim, R. Djemal, and A. Alsuwailem, "A DWT-entropy-

ANN based architecture for epilepsy diagnosis using EEG signals," *2nd Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2016*, pp. 288–291, 2016.

[109] F. Karimzadeh, R. Boostani, E. Seraj, and R. Sameni, "A Distributed Classification Procedure for Automatic Sleep Stage Scoring Based on Instantaneous Electroencephalogram Phase and Envelope Features," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 362–370, 2018.

[110] A. Renyi, "On Measures of Entropy and Information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1961, pp. 547–561.

[111] R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft, "Clustering using Renyi's entropy," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, 2003, vol. 1, pp. 523–528.

[112] C. Bandt and B. Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," *Phys. Rev. Lett.*, vol. 88, no. 17, p. 174102, Apr. 2002.

[113] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, "A Regularity Statistic for Medical Data Analysis.," *J. Clin. Monit.*, vol. 7, no. 4, pp. 335–45, Oct. 1991.

[114] K. K. Ho *et al.*, "Predicting Survival in Heart Failure Case and Control Subjects by Use of Fully Automated Methods for Deriving Nonlinear and Conventional Indices of Heart Rate Dynamics.," *Circulation*, vol. 96, no. 3, pp. 842–8, Aug. 1997.

[115] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy.," *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, no. 6, pp. H2039-49, Jun. 2000.

[116] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale Entropy Analysis of

Complex Physiologic Time Series," *Phys. Rev. Lett.*, vol. 89, no. 6, p. 068102, Jul. 2002.

[117] A. Di Ieva, F. Grizzi, H. Jelinek, A. J. Pellionisz, and G. A. Losa, "Fractals in the Neurosciences, Part I: General Principles and Basic Neurosciences," *Neurosci.*, vol. 20, no. 4, pp. 403–417, Aug. 2014.

[118] M. J. Katz, "Fractals and the Analysis of Waveforms," *Comput. Biol. Med.*, vol. 18, no. 3, pp. 145–156, Jan. 1988.

[119] M. Carrozzi, A. Accardo, and F. Bouquet, "Analysis of sleep-stage characteristics in full-term newborns by means of spectral and fractal parameters.," *Sleep*, vol. 27, no. 7, pp. 1384–93, Nov. 2004.

[120] W. Klonowski, E. Olejarczyk, and R. Stepien, "Sleep-EEG Analysis Using Higuchi's Fractal Dimension," in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA '05)*, 2005, pp. 222–225.

[121] N. Sriraam, B. R. Purnima, K. Uma, and T. K. Padmashri, "Hurst exponents based detection of wake-sleep — A pilot study," in *International Conference on Circuits, Communication, Control and Computing*, 2014, pp. 118–121.

[122] B. Weiss, Z. Clemens, R. Bódizs, Z. Vágó, and P. Halász, "Spatio-temporal analysis of monofractal and multifractal properties of the human sleep EEG," *J. Neurosci. Methods*, vol. 185, no. 1, pp. 116–124, Dec. 2009.

[123] J. Röschke, J. Fell, and P. Beckmann, "The calculation of the first positive Lyapunov exponent in sleep EEG data," *Electroencephalogr. Clin. Neurophysiol.*, vol. 86, no. 5, pp. 348–352, May 1993.

[124] C. J. Stam, *Nonlinear Brain Dynamics*. Nova Science Publishers, 2006.

[125] A. Lempel and J. Ziv, "On the Complexity of Finite Sequences," *IEEE Trans.*

Inf. Theory, vol. 22, no. 1, pp. 75–81, Jan. 1976.

[126] M. W. Rivolta, M. Migliorini, M. Aktaruzzaman, R. Sassi, and A. M. Bianchi, "Effects of the series length on Lempel-Ziv Complexity during sleep," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, no. 20133, pp. 693–696.

[127] D. Abásolo, S. Simons, R. Morgado da Silva, G. Tononi, and V. V Vyazovskiy, "Lempel-Ziv complexity of cortical activity during sleep and waking in rats," *J. Neurophysiol.*, vol. 113, no. 7, pp. 2742–2752, Apr. 2015.

[128] E. Estrada, H. Nazeran, P. Nava, K. Behbehani, J. Burk, and E. Lucas, "EEG feature extraction for classification of sleep stages.," in *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2004, vol. 1, pp. 196–199.

[129] B. Iser, W. Minker, and G. Schmidt, "Bandwidth extension of speech signals," in *Lecture Notes in Electrical Engineering*, 2008, vol. 13 LNEE, pp. 1–182.

[130] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, Feb. 2010.

[131] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–8.

[132] K. Pearson, *On Lines and Planes of Closest Fit to Systems of Points in Space*. University College, 1901.

[133] M. Rempe, W. Clegern, and J. Wisor, "An automated sleep-state classification algorithm for quantifying sleep timing and sleep-dependent

dynamics of electroencephalographic and cerebral metabolic parameters," *Nat. Sci. Sleep*, vol. 7, p. 85, Sep. 2015.

[134] Z. Yu, C. Kuo, C. Chou, C.-T. Yen, and F. Chang, "A machine learning approach to classify vigilance states in rats," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10153–10160, Aug. 2011.

[135] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel Dimensionality Reduction for Supervised Learning," *J. Mach. Learn. Res.*, vol. 5, no. 1, pp. 73–99, 2004.

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

[137] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," in *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, 2005, vol. 3, no. 2, pp. 523–528.

[138] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Springer US, 1998.

[139] T. Gasser, P. Bächer, and J. Möcks, "Transformations towards the normal distribution of broad band spectral parameters of the EEG.," *Electroencephalogr. Clin. Neurophysiol.*, vol. 53, no. 1, pp. 119–24, Jan. 1982.

[140] PhysioNet, "The Sleep-EDF Database [Expanded]," 2015. [Online]. Available: https://physionet.org/physiobank/database/sleep-edfx/. [Accessed: 01-Feb-2017].

[141] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, Feb. 2016.

[142] A. Rechtschaffen and A. Kales, "A manual of standardized techniques and scoring system for sleep stages of human subjects," *Washington, D.C. U.S. Gov. Print. Off.*, vol. NIH Public, p. 12, 1968.

[143] A. B. Wiltschko, G. J. Gage, and J. D. Berke, "Wavelet Filtering Before Spike Detection Preserves Waveform Shape and Enhances Single-Unit Discrimination.," *J. Neurosci. Methods*, vol. 173, no. 1, pp. 34–40, Aug. 2008.

[144] X. Kong, N. Thakor, and V. Goel, "Characterization of EEG signal changes via Itakura distance," in *Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society*, 1995, pp. 873–874.

[145] M. M. Deza and E. Deza, *Encyclopedia of distances*. 2009.

[146] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB." 2005.

[147] K. Kira and L. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *AAAI*, 1992, pp. 129–134.

[148] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1–2, pp. 23–69, 2003.

[149] C. Ding, H. Peng, and H, "Minimum redundancy feature selection from microarray gene expression data.," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[150] G. Guo, D. Neagu, and M. T. D. Cronin, "A Study on Feature Selection for Toxicity Prediction," Springer, Berlin, Heidelberg, 2005, pp. 31–34.

[151] Q. Gu, Z. Li, and J. Han, "Generalized Fisher Score for Feature Selection," 2012.

[152] H. L. H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, 1995, pp. 5–8.

[153] J. R. Quinlan, "C4.5: Programs for Machine Learning," Mar. 1993.

[154] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.

[155] R. Wald, T. M. Khoshgoftaar, and D. Dittman, "Mean Aggregation versus Robust Rank Aggregation for Ensemble Gene Selection," in *2012 11th International Conference on Machine Learning and Applications*, 2012, pp. 63–69.

[156] S. Lin, "Rank aggregation methods," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 5, pp. 555–570, Sep. 2010.

[157] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust Rank Aggregation for Gene List Integration and Meta-Analysis," *Bioinformatics*, vol. 28, no. 4, pp. 573–580, Jan. 2012.

[158] A. Ng, J. Ngiam, C. Foo, Y. Mai, and C. Suen, "UFLDL Tutorial," *http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial*, 2010.

[159] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[160] P. Prakash and A. S. K. Rao, *R deep learning cookbook : solve complex neural net problems with TensorFlow, H2O and MXNet*. Packt Publishing, 2017.

[161] D. T. Larose, "k-Nearest Neighbor Algorithm," in *Discovering Knowledge in Data: An Introduction to Data Mining*, 2004, pp. 90–106.

[162] M. Cilimkovic, "Neural Networks and Back Propagation Algorithm," *Fett.Tu-Sofia.Bg*, 2010.

[163] D. Svozil, V. Kvasnička, and J. Pospíchal, "Introduction to multi-layer feed-forward neural networks," in *Chemometrics and Intelligent Laboratory Systems*, 1997, vol. 39, no. 1, pp. 43–62.

[164] I. Steinwart and A. Christmann, *Support Vector Machines*, vol. 13, no. 4. 2010.

[165] K. Benabdeslem and Y. Bennani, "Dendogram based SVM for multi-class classification," in *28th International Conference on Information Technology Interfaces, 2006.*, 2006, pp. 173–178.

[166] L. Duckstein and S. Opricovic, "Multiobjective optimization in river basin development," *Water Resour. Res.*, vol. 16, no. 1, pp. 14–20, 1980.

[167] S. Opricovic and G. H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *Eur. J. Oper. Res.*, vol. 156, no. 2, pp. 445–455, 2004.

[168] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, May 2007.

[169] S. A. Imtiaz and E. Rodriguez-Villegas, "Recommendations for Performance Assessment of Automatic Sleep Staging Algorithms," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, vol. 2014, pp. 5044–5047.

[170] R. Mundry and J. Fischer, "Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: Examples from Animal Behaviour," *Animal Behaviour*, vol. 56, no. 1. pp. 256–259, 1998.

[171] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.

[172] S. A. Imtiaz and E. Rodriguez-Villegas, "Recommendations for performance assessment of automatic sleep staging algorithms.," in *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual*

Conference, 2014, vol. 2014, pp. 5044–7.

[173] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.

[174] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013.

[175] S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature ranking and rank aggregation for automatic sleep stage classification: a comparative study," *Biomed. Eng. Online*, vol. 16, no. S1, p. 78, Aug. 2017.

[176] F. Takahashi and S. Abe, "Decision-tree-based multiclass support vector machines," in *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.*, 2002, vol. 3, pp. 1418–1422.

[177] G. Madzarov, D. Gjorgjevikj, and I. Chorbev, "A Multi-class SVM Classifier Utilizing Binary Decision Tree," *Informatica*, vol. 33, no. 2. 2009.

[178] M. Bala and R. K. Agrawal, "Optimal Decision Tree Based Multi-class Support Vector Machine," *Informatica*, vol. 35, no. 2, 2011.

[179] S. Devuyst, T. Dutoit, T. Ravet, P. Stenuit, M. Kerkhofs, and E. Stanus, "Automatic Processing of EEG-EOG-EMG Artifacts in Sleep Stage Classification," in *IFMBE Proceedings*, 2009, vol. 23, pp. 146–150.

[180] R. J. Croft, J. S. Chandler, R. J. Barry, N. R. Cooper, and A. R. Clarke, "EOG correction: A comparison of four methods," *Psychophysiology*, vol. 42, no. 1, pp. 16–24, Jan. 2005.

[181] S. S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.

[182] P. Anderer *et al.*, "An E-Health Solution for Automatic Sleep Classification

according to Rechtschaffen and Kales: Validation Study of the Somnolyzer 24 × 7 Utilizing the Siesta Database," *Neuropsychobiology*, vol. 51, no. 3, pp. 115–133, May 2005.

[183] R. B. Berry *et al.*, "AASM - Manual for the Scoring of Sleep and Associated Events version 2.1." 2014.

Annex **List of Publications Related to the Proposed Work**

Publications in International Journals

1	S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature ranking and rank aggregation for automatic sleep stage classification: a comparative study," *Biomed. Eng. Online*, vol. 16, no. S1, p. 78, Aug. 2017.
2	A. A. Gharbali, S. Najdi, and J. M. Fonseca, "Investigating the contribution of distance-based features to automatic sleep stage classification," Comput. Biol. Med., vol. 96, pp. 8–23, May 2018.

Publications in International Conferences Proceedings

1	S. Najdi, A. A. Gharbali, and J. M. Fonseca, "A Comparison of Feature Ranking and Rank Aggregation Techniques in Automatic Sleep Stage Classification Based on Polysomnographic Signals," in *4th International Conference, IWBBIO*, 2016, pp. 230–241.
2	A. A. Gharbali, J. M. Fonseca, S. Najdi, and T. Y. Rezaii, "Automatic EOG and EMG Artifact Removal Method for Sleep Stage Classification," in *7th IFIP Advanced Doctoral Conference on Technological Innovation for Cyber-Physical Systems*, 2016, pp. 142–150.
3	S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature Transformation Based on Stacked Sparse Autoencoders for Sleep Stage Classification," in Technological Innovation for Smart Systems, 2017, pp. 191–200.