Classification of Sleep Stages Using Machine Learning Methods

Timplalexis Christos

SID: 3308170026

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2018

THESSALONIKI – GREECE

Classification of Sleep Stages Using Machine Learning Methods

Timplalexis Christos

SID: 12345678

Supervisor: Prof. Konstantinos Diamantaras

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2018

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in Data Science at the International Hellenic University.

Sleep is a necessary part in every human's life. Effective diagnosis and treatment of patients with sleep-related complaints is currently an urgent and heavily researched topic in the healthcare community. Sleep stage classification was introduced almost 50 years ago and manual approaches are sometimes used until today. Automatic sleep stage classification using machine learning can increase consistency and reliability, assisting experts at diagnosing sleep related health problems. In this dissertation, multiple classifiers are tested on 3 different datasets of healthy and patient subjects. The optimal algorithm achieves accuracy over 90% for the healthy subjects' dataset. In the results, the difference between the EEG patterns of healthy and patient subjects is highlighted. It is finally concluded that, using mixed datasets from healthy people and patients with minor sleep disorders, decent classification accuracies can be achieved. In addition to that, algorithms can generalize better as they can be used for a larger number of people.

I would first like to thank my thesis supervisor Professor Konstantinos Diamantaras for giving me the chance to elaborate on such an interesting field of research. His guidance and technical tips were valuable and crucial for the completion of this study. I would also like to thank the thesis co-supervisor Assistant Professor Ioanna Chouvarda who provided her valuable experience as an expert on the sleep stage classification problem. My deepest gratitude goes to my family for their support, encouragement and patience throughout all the years of my studies.

Contents

1 Introduction

1.1 Problem Statement

Sleep is fundamental for physical health and good quality of life. The understanding of the complex and dynamic phenomenon of sleep has always been a field of research for many scientists.

One of the most common methods that are used for sleep research is “sleep stage classification”. This method has been used for decades and it is still considered essential for the understanding of sleep architecture [1]. Sleep stage classification is one of the most critical steps in effective diagnosis and the treatment of sleep-related disorders. Classic approaches involve trained human sleep scorers, utilizing a manual scoring technique, according to some certain standards. Visual inspection undertaken by sleep experts is a time-consuming and burdensome task. Over the years, researchers have been trying to find more efficient, reliable and accurate methods for the classification process. Computer-assisted sleep stage classification systems are nowadays considered essential for both sleep-related disorders diagnosis and sleep monitoring. Over the years, the models that have been developed have been successfully applied to the sleep scoring problem. Even with that success, the research is ongoing aiming at improving these automated classification schemes.

1.2 Thesis Objectives

For years, researchers have been attempting to create automated sleep stage classifiers that can perform as well as human scorers. These efforts have seen some success using methods such as neuro-fuzzy systems (NFS), support vector machines (SVM) or artificial neural networks (ANN). The objective of this thesis is to suggest a sleep stage classification algorithm utilizing with the best possible way the existing features suggested by the literature, while introducing

some novelties on the feature extraction process, that result into a model with high classification accuracy. The algorithm is evaluated according to the new sleep scoring standards suggested by AASM [2], and is evaluated after choosing the appropriate metrics. The manual scoring is considered as a "golden standard" and everything is calculated with respect to that. Different models of classifiers are developed and they are compared to each other. All of the above methods are tested on three different datasets, including subjects with different sleeping habits (healthy-patients). Another pursuit of this thesis is to comparatively evaluate the suggested algorithm on these different datasets and draw conclusions about its generalization performance.

1.3 Thesis Composition

The rest of the thesis is structured as follows:

In **Chapter 2** the background theory which is necessary for the comprehension of the problem is analyzed. The importance of the researched problem is highlighted from medical aspect.

In **Chapter 3** the literature review related to the problem of sleep stage classification is summarized. Shallow and deep learning approaches are presented while different feature extraction approaches are pointed out.

In **Chapter 4** the data and methods used in this thesis are analyzed. Acquisition of data and data preprocessing methods are introduced. Feature extraction process is described and the selection of the classification algorithms is performed.

In **Chapter 5** the results obtained from testing the algorithms on the selected datasets, are presented.

In **Chapter 6** the results are discussed, conclusions are drawn and the direction of future studies is suggested.

2 Background Theory

2.1 Sleep Overview

Sleep is a natural condition of body and mind which is part of every person's life. Most people spend at least one third of their lives at a sleeping condition. The maintenance of health, wellbeing, homeostasis, memory and cognitive performance is closely related to good sleep [3,4].

Deprivation of sleep increases the risk of serious health problems such as heart diseases, obesity, diabetes and weakness of immunity system. After the pain, sleep disturbances are the second most frequent indicator of illness. Moreover, many restorative functions of the human body such as memory consolidation, mental restoration, mood and behavior are affected by the quality of a human's sleep [5].

The invention of electroencephalography in the 1930s is considered as the landmark for the beginning of the modern sleep research. In 1937, it was observed for the first time that sleep is not a homogeneous procedure, but it is comprised of some discrete stages deriving from the EEG [6]. In 1953, the rapid-eye-movement (REM) stage of sleep was observed. During REM sleep, rapid, binocularly symmetrical eye movements occur, EEG pattern is similar to the one observed during wakefulness and respiratory and heart rate are increased contrary to the other sleep stages. REM sleep is also closely related to dreaming as people awakened from REM sleep usually reported dreams [**7**]. According to [8], overnight recordings of EEG and EOG (electroculogram) specified the cyclic pattern of REM-NREM sleep which lasts about 90-100 minutes and this cycle typically occurs 4-5 times during every night. The same research also divided NREM sleep into four stages, stage 1 – stage 4, ranging from the lightest to the deepest sleep stage.

People were always trying to explain why sleep is so important for human life. Many theories answer adequately to many aspects of this problem, but there is no widely accepted theory able to explain all the phenomena related to sleep till now. According to [9] some of the most popular theories about sleep importance are the following:

1) Conservation of energy: It has been noticed that human energy consumption is increased during sleep deprivation while the basal metabolism is decreased 5-25% during sleep [10].
2) Restoration of tissues and growth: During the first hours of sleep growth hormone excretion, cell mitosis and protein synthesis are increased. In the time of growth or after a more labored day the amount of NREM sleep is increased. However, this theory is criticized as some scientists believe that cell mitosis occurs a few hours after food intake and the decreasing metabolic rate is in discrepancy with the protein synthesis that needs higher energy cost.
3) Thermoregulation: Sleep is believed to decrease the temperature as experiments with long-term sleep deprived rats recorded a temperature increase in about 10 degrees [11].
4) Regulation of emotions: Sleep deprivation is closely connected with disturbances of emotional behavior. Clinical observations have shown that depressed patients have a lower duration of NREM sleep.
5) Neural maturation: It is assumed that during REM sleep the maturation of brain and myelinization of nerve fibers proceed, since it has been noticed that children in about 6 months of prenatal phase spend 80% of sleep in REM sleep, while in young adults this percentage decreases to 25% [10].
6) Memory and learning: There is an information transfer between cortex and hippocampus during the sleep that realizes the fixation of memory traces or during REM sleep the insignificant bindings are abolished [12]. The learning process is related with this reprocessing of information. Several studies have shown the improvement of perceptual or motor task performance after sleep [13, 14]. The improvement is due to sleep and not due to time interval or circadian factors.

2.2 Sleep Stages

The most general categorization of sleep includes the stages of wakefulness, REM sleep and non-REM sleep. Apart from that, non-REM sleep is furtherly divided into 4 sub-stages. Stage 1 represents light sleep and moving towards stage 4 the sleep gets deeper. The scoring of

sleep stages is done according to [15], which is the proposal of a committee led by A. Rechtschaffen and A.Kales aiming at standardizing the process of recording and scoring sleep stages. It was published in 1968, trying to increase the comparability of the results reported by different researchers. This standard became widely accepted and the terminology suggested along with the scoring system are often used until today.

Rechtschaffen and Kales scoring system suggests scoring sleep stages by epochs. The duration of the epoch is set at 20 or 30 seconds (nowadays 30 second epochs are used in most recordings). The duration of the epoch does not change during the whole recording. When two stages occur during one epoch, the one that takes up the largest portion of the epoch should be scored as the stage of the epoch.

Stage W: It corresponds to a waking state. Alpha activity and low voltage, mixed frequency EEG are mainly present at this stage. High tonic EMG and eye blinks on the EOG are usually found in this stage.

Movement Time (MT): This scoring is assigned to epochs in which the EEG and EOG recordings are obscured for more than half of the epoch by muscle tension of amplifier blocking artifacts associated with movement of the subject. Movement time is not counted neither with sleep nor with wake time, unless the researcher specifies otherwise. Movement time should not be confused with discrete body movements which are physiological events that could happen during any sleep stage. Movement arousal is also distinguished in this stage and it is defined as any increase in EMG on any channel, which is accompanied by a change in pattern on any additional channel. Movement arousal is not used as an epoch score, but it is used as an aid in the scoring of sleep stages as it indicates that the continuity of a sleep stage has been disturbed in some way.

Stage 1: It is characterized by low voltage, mixed frequency EEG with a prominence of activity at 2-7 Hz. Vertex sharp waves may occur reaching an amplitude of 200μV. Stage 1 often occurs during the transition from W to other sleep stages or after body movements during sleep. During nocturnal sleep stage 1 is relatively short, as it occurs for only 1 – 7 minutes. Moreover, slow eye movements lasting for several seconds may appear during this stage. Scoring of this stage requires absolute absence of clearly defined K-complexes and sleep spindles. Traces of low voltage activity may appear when stage 2 approaches.

Stage 2: The presence of sleep spindles and/or K-complexes in combination with the absence of sufficient high amplitude define this stage. The presence of a spindle is registered if its duration is at least 0.5 seconds. K-complexes are EEG waveforms with a well delineated negative sharp wave, followed by a positive component with a total duration that exceeds 0.5 seconds. Sleep spindles occur in 12-14 frequency range. Sleep spindles and K-complexes are transient phenomena, so large periods of time may occur between these events without a sleep stage change. If the time interval between spindles and/or K-complexes is less than 3 minutes, then the intervening epochs are scored as stage 2, if there is no indication of movement arousal. If the time interval is more than 3 minutes, then the epochs are scored as stage 1 even if they do not contain movement arousal.

Stage 3: An epoch is labeled as stage 3 when the EEG recording consists of waves of 2Hz or slower for at least 20% but not more than 50% of the epoch. The amplitude of these waves is greater than 75μV peak-to-peak. Moreover, in this stage sleep spindles or K-complexes may occur.

Stage 4: This stage is defined by an EEG record in which more than 50% of the epoch consists of waves of 2Hz or slower and their amplitudes are greater than 75μV peak-to-peak. Even though, high amplitude may be present at half of the epoch (or a little more), the majority of stage 4 epochs are dominated by this activity. Intervals of lower amplitude and faster activity are present just for a few seconds. Sleep spindles and K-complexes could also occur in this stage.

REM: The main characteristics of REM stage are the simultaneous appearance of relatively low voltage, mixed frequency EEG activity and episodic REMs. The EEG pattern is closer to the one described in stage 1, differentiated by the absence of vertex sharp waves. In REM sleep, distinctive saw-tooth waves may appear in vertex and frontal regions combined with bursts of rapid eye movements. Alpha waves activity is prominent during REM with the frequency being 1-2Hz slower than wakefulness. Sleep spindles and K-complexes are completely absent, as it happens in stage 1.

For almost 40 years the only widely accepted method for sleep scoring was R&K method even though it was criticized as the rules often resulted in different scoring among experts. The American Academy of Sleep Medicine provided some new guidelines in 2007 (updated in 2012) regarding terminology, recording method and scoring rules [2]. The most important change was

considered the change in terminology, as the sleep stages were decreased from 7 to 5 and their annotations changed. R&K S1-S4 stages were changed to N1-N3 with N3 reflecting slow waves sleep (R&K stages S3 and S4). REM sleep is annotated as R and movement time stage is abolished. According to [16], following the rules for scoring, there is not a perfect match between stages S1-N1, S2-N2, (S3+S4)-N3. It is estimated that AASM rules decreased S2 by 4.9% while light sleep (S1) and deep sleep (S3+S4) were increased by 2.8% and 2.4% respectively. REM stage and total sleep time are not affected significantly.

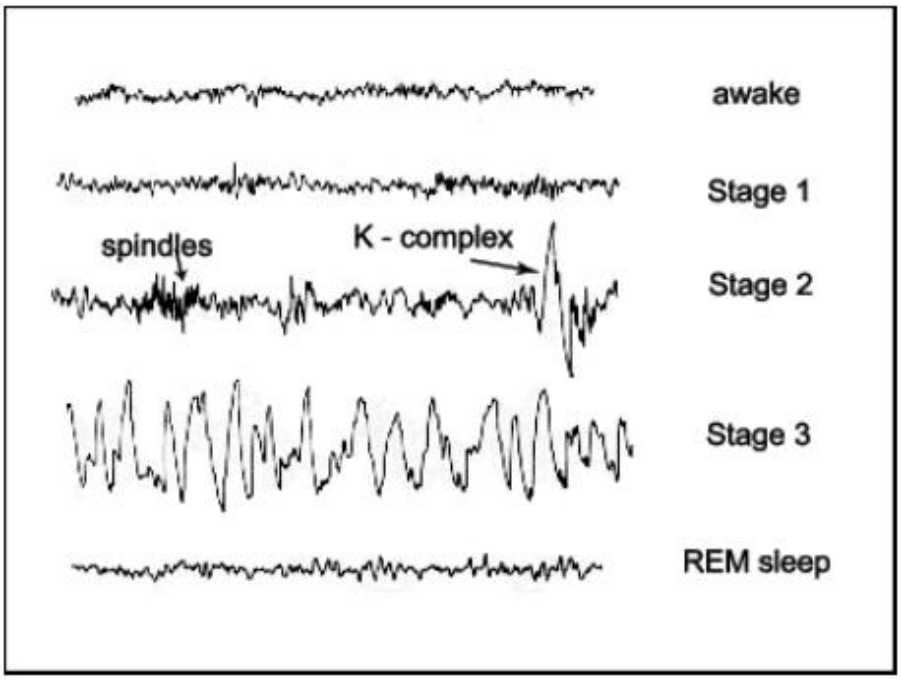

Figure 1 Sleep stages EEG waveforms

2.3 Electroencephalogram and Polysomnography

An electroencephalogram (EEG) is a recording of the voltages generated by the brain activity. Electrodes are attached to the scalp. Wires attach these electrodes to a machine, which records the electrical impulses. The results are either printed out or displayed on a computer screen and are used for the diagnosis of epilepsy, sleep disorders, brain tumors etc.

Polysomnography (PSG) is a type of sleep study used for the diagnosis of sleep disorders such as sleep apnea, narcolepsy, idiopathic hypersomnia, REM behavior disorder and parasomnia. Polysomnography is a process that makes a continuous recording during sleep measuring several physiological variables. Except for EEG, electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG) and respiration traces are measured. The polysomnography recordings are studied by experts who analyze sleep quality in both healthy and patient subjects. The analysis is done using sleep staging defined by certain standards (R&K, AASM). A sample PSG recording of 30 seconds, using Polyman software, is displayed below:

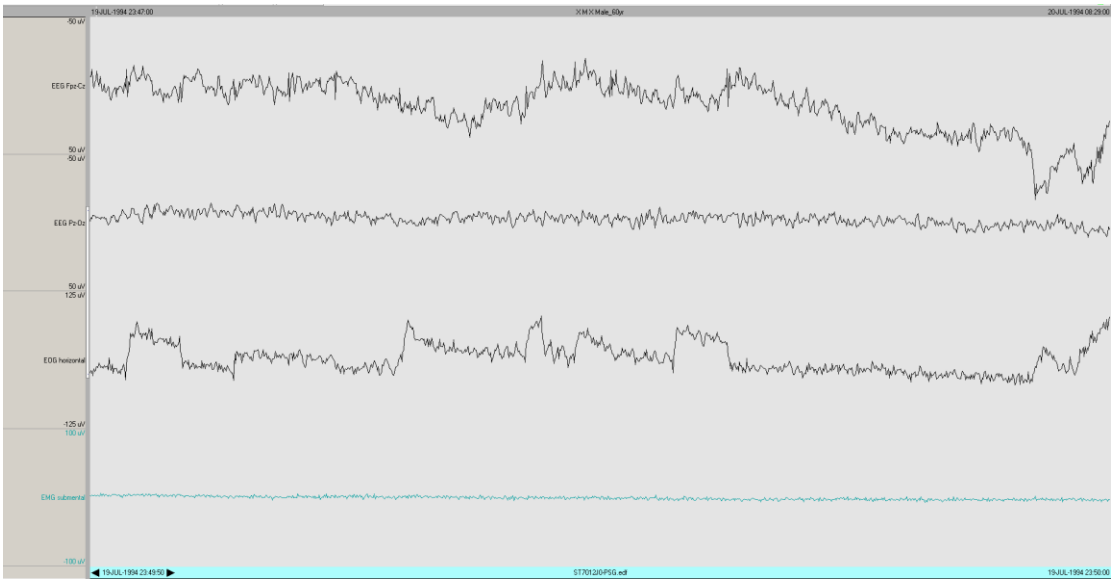

Figure 2 PSG waveform on Polyman

2.3.1 The 10-20 Electrodes Positioning System

The 10-20 system is an internationally recognized method which is used for the standardization of the electrodes positioning on the human scalp during EEG recording [17]. The method ensures that the spacing among electrodes is equal while the placement is adjusted to every subject as it is proportional to the subject's scalp shape and size. The numbers 10 and 20 denote that the distance between two electrodes is either 10% or 20% of the total front-back or right-left distance of the skull. All of the brain regions are covered. Letters are used to identify

the lobe location and numbers for the hemisphere location. The lobe locations are Frontal (F), Temporal (T), Parietal (P), Occipital (O) and Central (C) even though no central lobe exists and it is used only for identification purposes. The ‘z’ (zero) refers to an electrode placed on the mid line and finally even (2,4,6,8) and odd (1,3,5,7) numbers refer to electrodes positioned on the right and left hemisphere respectively. The electrode placement can be seen in the figure below:

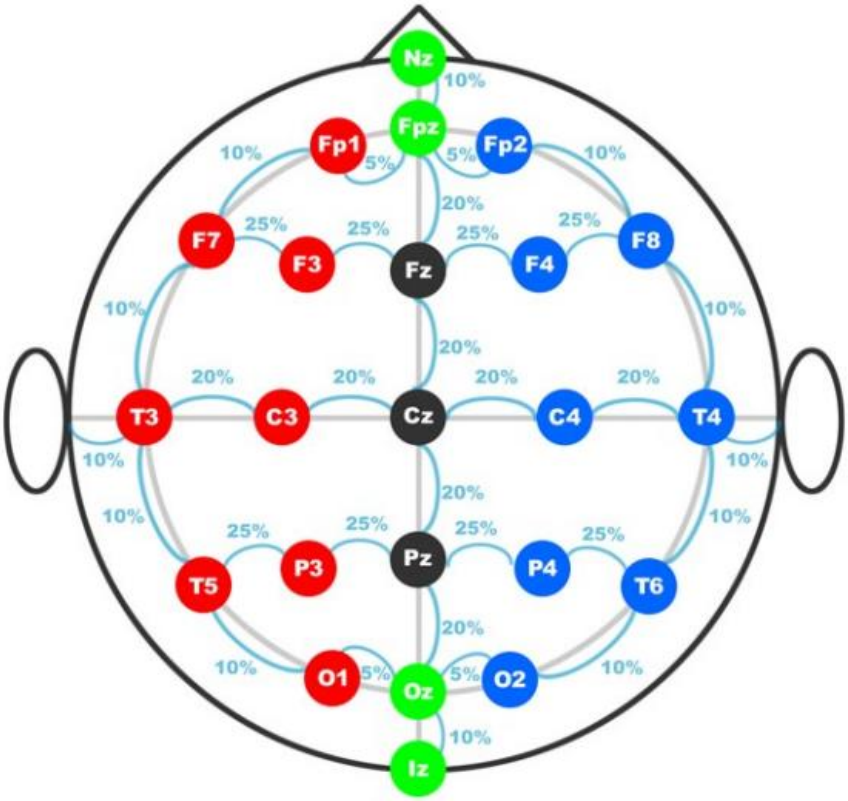

Figure 3 The 10-20 electrodes positioning system

2.3.2 Manual Sleep Scoring

Manual sleep scoring is considered a really demanding process. It is done by certified experts and according to very specific rules. The scorers must assign a sleep stage to the subject every 30 seconds for as long as the subject is sleeping (7-8 hours). All of the PSG recordings (EEG, EOG, EMG, ECG, respiratory recording) have to be taken carefully into consideration before the sleep stage is decided. Apart from that, every human may have some minor differences to his/her PSG recording depending on his/her age, health condition, sleep condition

etc. The PSG may even differ from day to day for the same subject, but polysomnography is a process that cannot be easily repeated because of its cost, effort and inconvenience of the patient. All the above, denote that it is almost impossible to have 100% reliable sleep scoring. In fact, according to [18] in the best of circumstances the experts agree at a percentage of 90%. The results of sleep scoring among 3 expert raters at two different groups of PSGs are presented below:

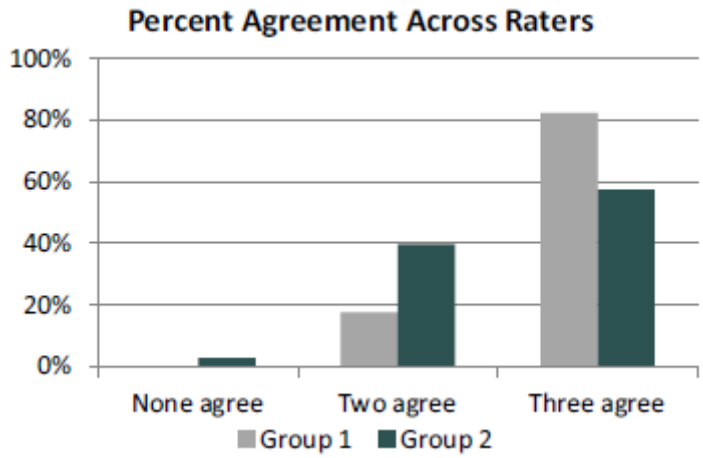

Figure 4 Agreement across 3 human scorers on sleep staging

It becomes obvious that human sleep scoring is done in a subjective way and the results depend on the rater's interpretation of the rules. Automatic sleep staging is a way to overcome some of these issues. The authors of [18] develop an automatic sleep staging algorithm and try to compare the results with manual scoring. It is found that the auto-staging algorithm was comparable to manual scoring for all stages combined (Wake, N2, N3) and the agreement slightly decreased for REM sleep. It is also observed that patients with insomnia and psychiatric disorders who were using medications, had a more difficult recognition of their N3 and REM stages which resulted in lower agreement between raters and auto-staging. Contrary to that, patients with Obstructive Sleep Apnea (OSA) have easier to recognize EEG signal patterns which leads to a partial agreement between manual and automatic scoring, however the use of more than one human experts using majority voting, increased significantly the sensitivity for all of the sleep stages.

3 Literature Review

During the research it was found that sleep stage classification using machine learning has been extensively studied by many researchers. Acquisition of sleep data is not always feasible, as they are considered sensitive, medical data. As a consequence, most of the researches, including this one, use data available at online data repositories (e.g. Physionet). The volume of data is also a problem due to the EEG signal sampling rate that creates huge hypnogram files. In most cases, a quite small number of sleep subjects is being used, even less than 10. This mainly applies for researches that were carried out some years earlier, since computing power nowadays allows the easier processing of big hypnograms files. Data preprocessing is a demanding process as hypnograms datasets tend to be noisy and moreover, data must be brought to a specific form in order to proceed with the feature extraction. There is a big variety of features and data transformations that has already been tested. The classification problem has been approached by miscellaneous classifiers, according to the literature, but it still remains some opportunity for further investigation mainly for deep learning methods.

3.1 Shallow Learning Approaches

Shallow learning methods follow the typical path of data preprocessing and feature extraction before feeding the data into a classifier.

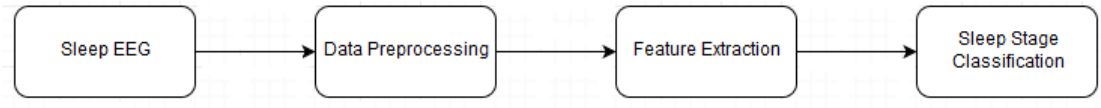

Figure 5 Typical process for shallow learning approaches

More specifically, in [19] Physionet repository is selected, using data from both male and female healthy subjects aged between 21 and 35 years. Hidden Markov Model is used for classification trying to correlate the transitions between sleep stages. After studying the accuracy of predicting each class (stage) separately, it is deduced that some classes are predicted more easily than others. Stage 1 is underrepresented in the dataset and its accuracy is below 50% while

stage 4 and REM reach 91% and 86% respectively. The same problem regarding the accuracy of sleep stage 1 is met at [20]. In this case, SVM classifier is used while most of the features are statistical measures of the Pz-Oz electrode such as absolute mean-value, kurtosis, skewness, standard deviation. Basic statistical measures are typically used by all sleep stage classification models. In [21], the authors suggest some additional features such as spectral energy band, central frequency, bandwidth and Itakura Distance. Conclusively, Itakura distance and central frequency are characterized as promising for sleep stage classification tasks. A different approach is also noticed at [22], which is one of the few researches that uses data from actual patients (male and female patients with apnea). The authors included signals from EEG, EOG and EMG. All of the signals were converted to frequency domain using FFT and Delta, Theta, Alpha, Beta and Gamma wave frequencies were extracted. Regarding the classifier, MLP optimal number of hidden layers and learning rate were estimated after multiple experiments. MLP outperformed other classifiers, but its optimum performance is still considered low, so deep learning approaches are suggested. Conversion from time domain to frequency domain is widely used as a preprocessing tool used for feature extraction. In [23], FFT and wavelet transform (WT) in EEG signals are compared. It is found that in the spectral analysis, WT is more efficient than FFT, mainly due to the fact that EEG signals are non-stationary, so small changes may not be realized by FFT and the analysis may change depending on the length of data. A comparison among conditional random fields (CRF), HMMs and Bayesian linear discriminants using the AASM scoring standard is attempted in [24]. A total of 443 recordings were used using healthy subjects and patients diagnosed with apnea. They were separated at 3 datasets, the first one containing 428 healthy or apnea subjects and the second and third one comprising of 6 and 9 healthy subjects respectively. It is proven that, CRF classifiers are superior to the others and they moreover provide moderate sleep stage classification results for patients with apnea, outperforming earlier work. A common finding with other researches is the fact that the sensitivity of N1 stage needs to be improved. The authors of [25] use single channel EEG recording, introducing a new time domain feature named Statistical Behavior of Local Extrema (SBLE). This feature provides a profound understanding of the hidden dynamics of EEG signals by quantifying and symbolizing its local extrema information. The feature extraction process comprises of 6 steps: identification of local extrema, finding amplitude intervals, symbol assignment, definition of micro patterns, definition of macro patterns and finally feature

extraction. The proposed features are: 1) number of local maxima in each amplitude interval, 2) number of local minima in each amplitude interval, 3) number of each macro pattern. The above process essentially transforms the values of the EEG time-series signal into a finite set of symbols, with each symbol being a member of a predefined symbolic set. The proposed classification method using SVM, achieved an accuracy of 91.8% for 5-stage classification at the Sleep-EDF database. It is highlighted that only Fpz-Cz channel was used for feature extraction and it outperformed Pz-Oz, contrary to what most researches have shown. Furthermore, SBLE feature contributed to yielding outstanding results, even when a small part of the whole dataset was used for training which means that the proposed method can accelerate the process of training the classifier and reduce the computational cost. The novelty of [26] is the emphasis given on the feature selection process. EEG signal is decomposed into 8 sub-bands and then 13 features are extracted for each sub-band. The statistical significance of the features is examined using the Kruskal-Wallis test, discarding the ones with low significance. Then the best remaining features are selected using mRMR algorithm. Random forest is used for classification achieving an average accuracy of 98.5% for the 2-stage classification problem. The method is tested at 3 Physionet datasets: Sleep-EDF, UCDDB and Expanded Sleep-EDF. Conclusively, the authors believe that an implementation of their method to a portable device could be possible, mainly because of the use of only one EEG channel and due to the method's low computational cost. A research that focuses on the non-linearity and non-stationarity of EEG data is described in [27]. The fact that, non-linear signal processing is more efficient than time domain analysis and power spectral analysis, is supported by relevant literature. The analysis of EEG signals is done using higher order spectra (HOS) that are spectral representations of higher order moments or cumulants of a signal. The Gaussian Mixture Model is used for classification reaching an accuracy of 88.7% at 6-stage classification using 2 EEG channels from St Vincent's University Hospital dataset. The high accuracy obtained denotes that HOS analysis can reveal information that is hidden due to the EEG signal's non-stationarity. The majority of the researches related to sleep stage classification use time, frequency or time-frequency features. In [28], the problem is tackled in a totally different way as graph domain features are extracted from a single channel (Pz-Oz). More specifically, each segment of the signal is mapped into a visibility graph (VG) and a horizontal visibility graph (HVG), without any frequency domain preprocessing. Then a difference visibility graph (DVG) is constructed for each EEG segment. Finally, 7

distinguishable degree distribution values are selected for each DVG as representative features, which are fed into an SVM classifier. The results are impressive, as an accuracy score of 89% is achieved for 10-fold cross-validation with 14963 epochs of EEGs. By the time the paper was published, this was by far the best score for Sleep-EDF dataset using more than 10000 epochs. In [29], emphasis is given in building a model as simple as possible so that it can be easily deployed on a wearable device. Sleep-EDF dataset is chosen and only one EEG channel is used for feature extraction. The signal is firstly segmented into 30s epochs and then spectral features are computed for each epoch. Less useful features are discarded through a feature selection process. Boosted trees algorithm is selected for classification as it is considered a computationally non-intensive algorithm. The results show 82.03% accuracy for 5-stage classification which is generally a poor result, but the whole method seems more plausible to be implemented to a real-life application. In [30], the authors decided to use 5 EEG channels taking into consideration the cross correlation information that exists between channels. Most researches use up to 2 EEG channels, so 5 channels were expected to provide some extra information and boost sleep stage classification accuracy. It was finally found that this method reached 93.2% average accuracy for 2 stage classification considering 4 stages in total (S1, S2, S3+S4, REM). This accuracy was considered decent when the paper was published, but leads to the conclusion that adding more channels to the analysis does not significantly improve classification accuracy. On the contrary, it adds unnecessary computational cost and could probably lead to overfitting. A totally different set of features is tested in [31]. Distance based features are compared with conventionally used features for the sleep stages classification problem, achieving remarkable accuracy. Only healthy subjects were used for the evaluation of the proposed method. The C3-A2 EEG channel which is proposed by the AASM rules was used for feature extraction along with EOG, EMG and ECG signals. The conventional feature vector consisted of 48 features that were found to be the most common in the literature. 31 distance-based features were extracted, all of which were based on Itakura distance and COSH diastance.

3.2 Deep Learning Approaches

Artificial Neural Networks (ANN) and deep learning in general is a solution that is tried out at almost every type of machine learning problem. The results are often impressive, nevertheless the problem of classification of sleep stages does not seem to have any significant improvement at the prediction accuracy. One of the first classification methods that used neural networks is described in [32]. The study was published in 1999 and to the authors' knowledge this is the first research that used Wavelet Transform for feature extraction. This is considered a milestone, as Wavelet Transform later proved to be really effective at the sleep stage classification problem, usually outscoring methods that used Fourier Transform for feature extraction. A feedforward neural network was used for classification, which was comprised of 13 neurons in the input layer, 10 neurons in the hidden layer fully connected to the first layer and 6 neurons in the output layer. The structure of the network is indicative of the available computational power that could be used at that time. The classification results compared to those of a human expert reached a 70-80% agreement. Apparently, the results are poor but the methods used were novel for that time and paved the way for future studies. In [33] sleep data from PhysioBank database were used and Pz-Oz EEG channel was selected for the extraction of feature vectors from the signal's wavelet coefficients. Time-frequency analysis (wavelet transform) is suitable for EEG recordings, because the signals are dynamic, sometimes transient and mostly non-stationary. The feature vectors were classified using an MLP ANN with one hidden layer and traingdx algorithm. It was observed that the optimum number of neurons for the hidden layer was 8. After binary classification at a total of 4 stages (Wake, S1+REM, S2, SWS) the results indicated 93 ± 4% accuracy. This accuracy score is considered common when one class (sleep stage) is classified against all the rest. In [34], a recurrent neural classifier is presented, attempting 5-stage classification to 8 recordings obtained from the Sleep-EDF database, using energy features that derived from the EEG signal of the Fpz-Cz channel. The classification of the proposed neural classifier was tested against a feedforward neural network and a probabilistic neural network. It was found that the recurrent classifier was by far more accurate, reaching an accuracy of 87%, while the other two classifiers classified correctly approximately 81% of the samples. However, it is recognized that the discrimination of stage N1 from REM sleep needs to be improved. Coming to more recent researches, in [35] the use of

complex-valued convolutional neural network is examined. The authors argue that the construction of handcrafted features for sleep stage classification is a process that requires domain knowledge from experienced experts and besides that, it is time-consuming. Their method initially converts real-valued EEG signal into complex numbers. Then the transformed signal is fed into the network which is essentially a multi-layer perceptron with a special topology containing more than one hidden layers. A 6-layer network is used, reaching 90.8% accuracy which outperforms real-valued CNNs. The problem of low classification accuracy for stage S1 is also noticed by the authors. Another approach that also uses CNNs is presented in [36]. PSGs from Sleep-EDFx database were used for this study and 6-stage classification was attempted. Fractional Discrete Fourier Transform (F-DFT) is used to fully utilize the local frequency domain information of EEG signals. Wavelet Transform is also used in an effort to depict the low frequency structure information of local signals and better classify stages 3 and 4. A 3-dimensional signal is constructed from EEG signal, F-DFT signal and WT signal. The signal is fed into the CNN where the feature extraction process happens automatically. The total accuracy for the 6-stage classification is 90.11%.

4 Data and Methods

4.1 Data Acquisition

For the purpose of this thesis, the datasets were obtained from Physionet online data repository [37]. Physionet offers free access to a collection of recorded physiological signals and open-source software related to the signals' processing. There are several polysomnographic sleep recordings libraries, however the Sleep-EDF Database [Expanded] was used for this research [38]. The database was greatly expanded in March 2018 and now contains 197 PSG recordings in total that were obtained from 2 different studies. The first study was conducted in 1987-1991 and had as a target to study the effects of age in sleep quality. 153 recordings are from this study and the subjects were healthy Caucasians aged 25-101. Files are named in the form SC4ssNEO-PSG.edf where ss is the subject number and N is the night. The remaining 44 PSGs were obtained in 1994 in a study related to the effects of temazepam on sleep. 22 subjects (Caucasian males and females) that had mild difficulty falling asleep but were otherwise healthy participated to this research. Each subject was recorded for 2 nights, one of which was after temazepam intake and the other after placebo intake. Files are named in the form `ST7`**`ssN`**`J0-PSG.edf` where **ss** is the subject number, and **N** is the night.

All of the PSGs contain EEG, EOG, EMG, respiration and body temperature recordings. Sleep scoring is conducted manually by well-trained experts according to Rechtschaffen and Kales manual, [15] but Fpz-Cz and Pz-Oz EEG channels are used even though C4-A1 and C3-A2 are suggested by the manual. EEG, EOG and EMG are sampled at 100Hz. The PSG files are formatted in EDF while the hypnograms (annotations) are in EDF+.

For the purpose of this study it was decided to include 3 groups of subjects. The first one contains totally healthy subjects that their PSG was recorded under no medication. The second one contains patients that have mild difficulty falling asleep, during their placebo intake nights and the third one also contains patients with difficulty falling asleep but the recording was done under temazepam effect. 4 PSGs were used for each group.

Table 1 PSG recordings for each group of subjects

Group	PSGs
Group A: Healthy subjects	SC4001E0, SC4002E0, SC4011E0, SC4012E0
Group B: Mild difficulty falling asleep – no medication	ST7011J0, ST7022J0, ST7041J0, ST7052J0
Group C: Mild difficulty falling asleep – temazepam intake	ST7012J0, ST7081J0, ST7152J0, ST7222J0

4.1.1 Temazepam Effects

Temazepam belongs to the class of medications called benzodiazepines and it is used for the treatment of short-term sleeping problems (difficulty falling asleep, frequent awakenings, early morning awakenings). Temazepam's plasma half-life is 8.5 ± 1.5 h and maximum plasma levels are reached 0.6-0.8 h after oral administration [39]. Benzodiazepines are commonly used medications that alter sleep spindles during NREM sleep. The effects of temazepam on human EEG have been studied by several researchers in the past. In [40], twenty healthy males aged 21-26 years with regular sleeping habits participated in the study. The subjects' EEG was recorder both for placebo and temazepam intake nights. It was found that compared to the placebo condition, temazepam significantly reduced the interval between lights-off and the first occurrence of stage 2 NREM sleep. Moreover, total sleep time was significantly longer in the temazepam condition and comparing the first 6 hours of sleep for the two nights, it was noticed that temazepam significantly reduced REM sleep but it did not reduce slow-waves sleep or stage 4 NREM sleep (R&K scoring). A similar study [41], which was also conducted with healthy volunteers on placebo and medication nights, detected changes in the recorded EEGs, using mean power density spectra and t tests. The authors concluded that the changes of the EEG pattern include a decrease in power in the 7 to 12 Hz frequency region and an increase in power in the 12 to 25 Hz region. The distinction between placebo and temazepam nights was absolutely clear. Lastly, a much newer (2015) study is presented, where sleep was monitored using 256-

channel hdEEG [42] while prior investigations have typically utilized 1 or 2 central channels in their analyses. 18 healthy adults participated in this study recording EEGs on two separate nights, after placebo and temazepam intake. Spectral analysis and spindle waveform detection of sleep EEG data were performed for each participant night. Sleep staging did not have significant differences between placebo and temazepam nights, however spindle duration and amplitude were significantly increased for temazepam nights. According to the topographic analysis, spindle duration was increased in the vast majority of channels except for a few centrolateral derivations.

4.2 Data Preprocessing

The European Data Format (EDF) is a data format that was created in 1990 for the easiest storage and exchange of multichannel biological and physical signals. It is the most popular data format for EEG and PSG recordings. An extension of EDF, developed in 2002, is EDF+. The newer type of files also contains interrupted recordings, annotations, stimuli and events. All existing EDF viewers are also compatible with EDF+.

The PSG recordings that were downloaded for this study were EDF files and their respective hypnograms were EDF+. Some certain steps of preprocessing were followed for each file, so that it could result into a dataset that could be read from python in order to proceed with the feature extraction process.

- PSG

The first step was using the appropriate software for opening EDF files. “Polyman” was used for this purpose, which is an open-source software available at Physionet webpage. PSG files are loaded on Polyman and the user has the opportunity to plot one or many of the channels of the PSG recording.

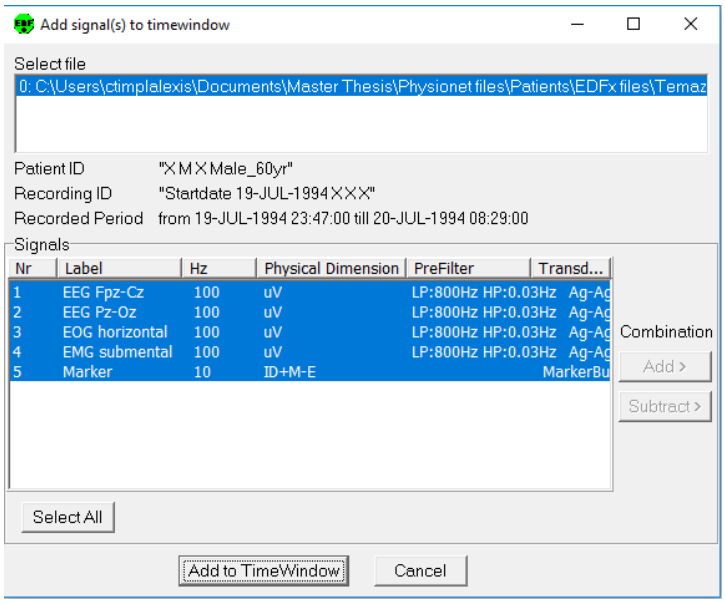

Figure 6 Adding PSG signals to Polyman

The selected signal was converted to ASCII and finally the file which was created was converted to .csv so that it could be read from python. The signals that were used to the final datasets are the following:

- **EEG Fpz-Cz**
- **EEG Pz-Oz**
- **EOG Horizontal**

- Hypnogram

Hypnograms had EDF+ format and EDF Browser was used for their easier manipulation. This software is also open-source and available at Physionet webpage. When hypnogram files were loaded the expert annotations for the corresponding PSG appeared.

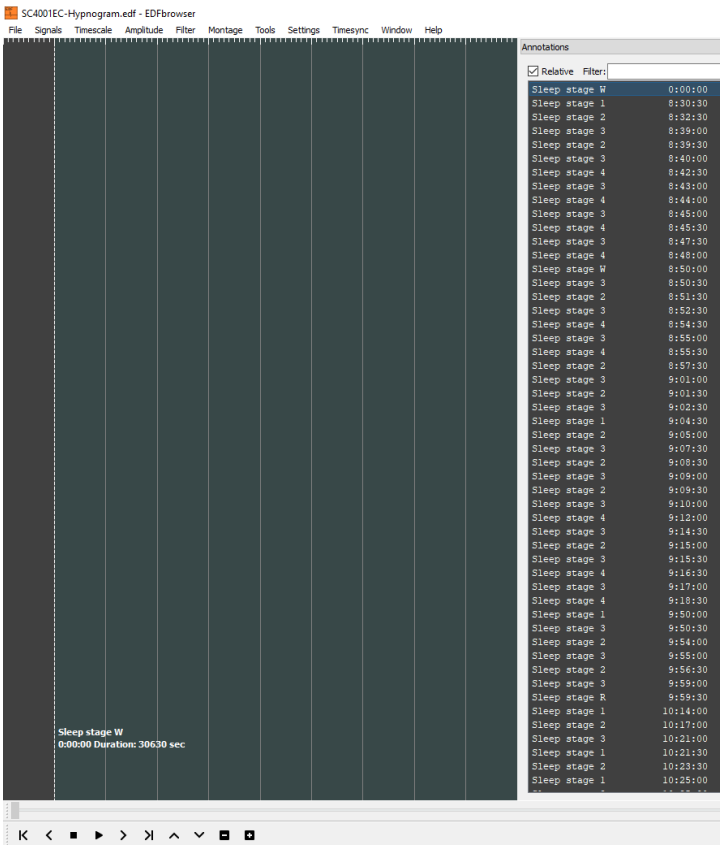

Figure 7 Reading hypnogram annotations (EDF+ files) on EDF Browser

Then, by selecting the option "Export EDF to ASCII" a .txt file was created that contained the annotations, the onset and the time duration of each sleep stage. Finally, this .txt was also converted to .csv.

Examining the annotations file, it becomes clear that some parts of the signal need to be removed. Movement time is a stage that has to be discarded. Moreover, most of the recordings start and/or finish with many consecutive hours of "Awake" stage for the subject. Keeping these parts of the signals would create imbalance to the sleep stages of the dataset, so they were also discarded. Apparently, the same parts of the recording were discarded for all of the channels that were used for the feature extraction, so that time correspondence was kept among them.

4.3 Selection of Scoring Standard

For the purpose of this research, the sleep scoring standard of the American Academy of Sleep Medicine was used [2]. This standard is much more recent than R&K rules and most of the

studies published the latest years follow this standard. At this point, an inconsistency between AASM scoring and the available sleep datasets needs to be pointed out. As it has been mentioned before, there are not many polysomnographic recordings publicly available, as medical data are considered sensitive. As a matter of fact, most of the PSG datasets are quite old and the scoring from the experts is done according to the old R&K rules. So, recent studies usually convert sleep stages from R&K to AASM simply by adding stages S3 + S4 of slow wave sleep, creating stage N3. This conversion cannot be considered totally accurate, as the new rules have changed the overall duration of every sleep stage during a normal night's sleep. In addition to that, the new rules suggest sampling frequency of 500Hz while most datasets (including the one used in this study) include signals sampled at 100Hz. Moreover, the proposed EEG channels are F4-M1, C4-M1, O2-M1 and backup channels F3-M2, C3-M2 and O1-M2. Sleep EDF-x EEG signals that were analyzed in this study are from channels Fpz-Cz and Pz-Oz.

Despite the above inconsistencies it has been noticed that many recent studies select to follow the AASM scoring rules even if the dataset is annotated otherwise. Consequently, AASM manual is also chosen for this study which means that samples are classified at epochs of 30 seconds or 3000 data points (f=100Hz) using 5 stages for classification (W, N1, N2, N3, R).

4.4 Feature Engineering

After the dataset is cleaned, raw data must be transformed into meaningful features that represent the information carried by the signal.

4.4.1 Feature Extraction

The nature of the signal should be well understood, in order to create features that describe raw data with the best possible way. PSG signals are non-stationary, which means that the signal's statistics change over time. This means that analysis of the signal on the time domain is not sufficient. Time domain features, frequency domain features, time-frequency domain features, entropy features and non-linear features are extensively used in the literature [43] as they reveal different aspects of the EEG, EMG and EOG signals.

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.

Figure 8 Conventional features extracted from PSG recordings

The features that were extracted from the PSG signals are presented below. For every feature, each instance was calculated at 30 second epochs, so 3000 data points were used for every instance.

- Arithmetic Mean: This feature was used both for EEG signals (2 channels) and EOG. The mean electric potential of an epoch is calculated. The data points of an epoch (electric potentials) were added and then divided by 3000.

$$\boldsymbol{E(X) = \mu = \frac{\sum_{i=1}^{i=N} x_i}{N}}$$

- Variance: It was also used for both EEG and EOG signals. Variance measures how much an epoch's data points spread out from their average value.

$$\boldsymbol{Var(X) = E[(X - \mu)^2]}$$

- Skewness: Is is a measure of the degree of asymmetry of a distribution. The skewness of a normal distribution is zero, while positive and negative skewness indicates that data are skewed right and left respectively. Skewness is a higher-order-statistics measure (third moment).

$$\boldsymbol{skewness = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^3 / N}{s^3}}$$

- Kurtosis: It is the peakedness or flatness of the graph of a frequency distribution especially with respect to the concentration of values near the mean as compared with the normal distribution. Kurtosis is a higher-order-statistics measure (fourth moment).

$$\boldsymbol{kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^4 / N}{s^4} - 3}$$

- Entropy: It is a way to measure the randomness of an epoch's data points. Entropy describes the lack of order or predictability. High entropy denotes a stochastic process that does not form a specific pattern.

$$\boldsymbol{entropy} = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{p_i} * \boldsymbol{log}(\boldsymbol{p_i})$$

- EEG Frequency Bands: EEG waveforms are classified into five different frequency bands. These bands are components of the overall EEG waveform captured from one electrode. Band information can be extracted when a mathematical transformation like Fourier Transform is used, in order to convert the signal from time domain to frequency domain. Frequency bands have been successfully used as a feature to many machine learning problems related to EEG analysis, from classification of sleep stages [44] to epilepsy detection, human emotion recognition [45] and cognitive performance [46].

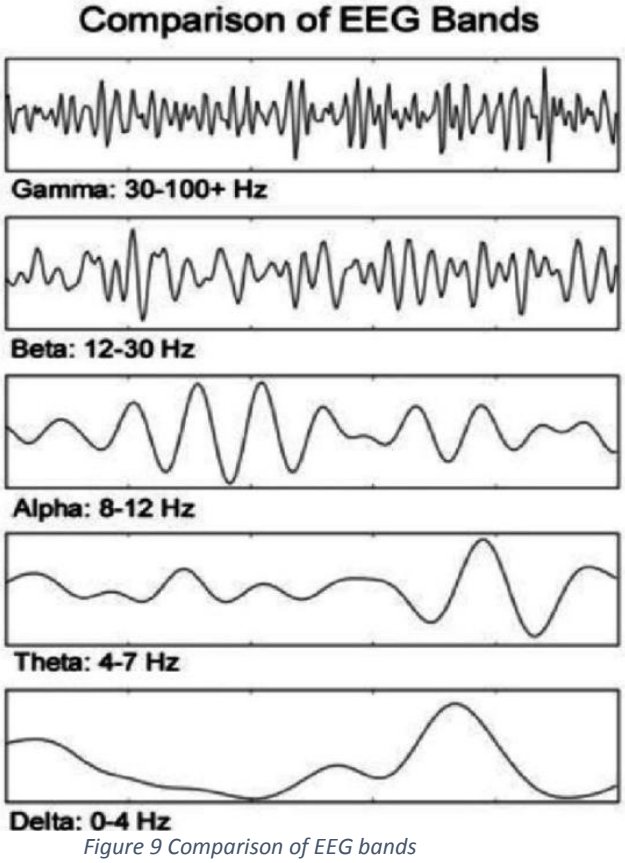

Figure 9 Comparison of EEG bands

After the signal is converted to the frequency domain, features can be extracted. The mean value and the average power of each band (delta, theta, alpha, beta, gamma) were extracted for every epoch.

$$\boldsymbol{average\ power} = \frac{1}{N}\sum_{n=1}^{N} |x(n)|^2$$

- Power spectral Density: It is used to describe how the power of a signal or a time-series is distributed over frequency. The power is defined as the squared value of the signal. The unit of PSD is energy per frequency and its computation is done directly from FFT.

- Petrosian Fractal Dimension: Fractal dimension is a ratio that provides a statistical index of complexity, comparing how detail in a pattern changes depending on the scale at which it is measured. Petrosian Fractal Dimension is a feature extracted from PyEEG library [47] which is an open-source python module for EEG/MEG feature extraction.

All of the above features were included in the final model. However, some other features that eventually did not improve the model's accuracy were tried.

- Hjorth parameters: This is a feature that is widely used on EEG feature extraction [48, 49, 50]. Hjorth mobility and complexity are also computed from PyEEG library.

$$\boldsymbol{Mobility} = \sqrt{\frac{\boldsymbol{var}\left(\frac{\boldsymbol{dy(t)}}{\boldsymbol{dt}}\right)}{\boldsymbol{var}\big(\boldsymbol{y(t)}\big)}}$$

$$\boldsymbol{Complexity} = \frac{\boldsymbol{Mobility}\left(\frac{\boldsymbol{dy(t)}}{\boldsymbol{dt}}\right)}{\boldsymbol{Mobility}\big(\boldsymbol{y(t)}\big)}$$

- Wavelet analysis features: Wavelet Transform (WT) solves Fourier Transform's disadvantage, which is that is does not provide any localization in time. In WT, some scaled (stretched and compressed versions) of a mother wavelet are correlated with the original signal, revealing details at different scales. For non-stationary signals, such as EEG signals, WT which is a time-frequency analysis method is often used with great results [32, 51]. In this study, some statistical features were extracted from the approximation and the detail coefficients of WT, but they did not contribute towards the improvement of the model's accuracy.

- Number of zero crossings: This feature counts the occurrences of a signal value change from positive to negative, or from negative to positive. It has been used in electroencephalography both for sleep stage classification [52] and epileptic seizures detection [53].

Another method which was tried, but was finally rejected is the filtering of the raw signal. PSG signals contain a lot of noise because of the subject's movements or the electrodes' contact with the scalp. The idea of denoising the signal has been used in the past and according to [54, 55] Savitzky-Golay filter has been successfully tested for EEG signal processing. In the present study, when raw signal was initially filtered with a Savitzky-Golay filter and the features were extracted afterwards, the model's accuracy dropped by 3-4%. This means that some useful information of the signal was lost, trying to smoothen the noise.

4.4.2 PSG Signals as Time-Series

The EEG and EOG signals that were used for feature extraction, can be considered as time-series data. The signals' discrete data points form a sequence which is ordered in time. This ordering creates dependence between adjacent points of the signal that could possibly provide some extra information to our model and increase its predictive accuracy. This concept has been implemented in the literature, mostly at deep learning models that used LSTM layers [56, 57]. In this study, a different approach is used, as the existing features are shifted in time. More specifically, the 36 features that are initially extracted are shifted one time-step backward and

one time-step forward. Finally, the proposed model ends up with 108 features (3*36). Time-shifting for one feature is explained at the tables below:

Table 2 Backward time shifting

Time	Original Feature Extracted	Time-Shifted Backward
t=1	X_1	X_1
t=2	X_2	X_1
t=3	X_3	X_2
t=4	X_4	X_3
t=5	X_5	X_4

Table 3 Forward time-shifting

Time	Original Feature Extracted	Time-Shifted Forward
t=1	X_1	X_2
t=2	X_2	X_3
t=3	X_3	X_4
t=4	X_4	X_5
t=5	X_5	X_5

The novelty of the current study is focused on the above feature transformation that increased the model's accuracy by 3-4% depending on the dataset that it was tested. Moreover, it is an approach that did not significantly increase the time complexity of the model.

4.5 Classification Methods

After the feature extraction process, the model was tested with many different classifiers (linear, decision trees, boosting) in order to study the behavior of each classifier and select the best one. Gaussian Naïve Bayes and Support Vector Machines (SVM) provided the worst results, out of all the classifiers that were used. SVM was expected to be more efficient as it has been successfully used in the literature [20]. Low accuracy may have occurred as a result of suboptimal tuning of the algorithm's parameters. Logistic Regression and Random Forest algorithms achieved decent results that approach the optimal solutions. The performance of Logistic Regression is remarkable as it is a computationally cheap algorithm that runs very fast. The best results were obtained with Extremely Randomized Trees, Gradient Boosting and XGBoost which are analyzed below:

- Extremely Randomized Trees: During the nineties, the systematic study of tree-based models revealed their high variance. During the same period, the idea of combining multiple models in order to reduce variance and bias appeared. In 1994, Leo Breiman proposed "Bagging" [58] which can be considered a particular instance of a broader family of ensemble methods that are called randomization methods. These methods introduce randomization into the learning algorithm, providing a different version of the original learning sample at every run. Extremely Randomized Trees is a tree-based ensemble method, presented in 2006, that is suitable for both classification and regression problems. The algorithm selects splits, both attribute and cut-point, totally or partially at random [59]. The Extra-Trees algorithm builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure, but has two important differences with other tree-based ensemble methods. Firstly, it splits nodes by choosing cut-points totally at random and secondly it uses the whole learning sample to grow the tree. When the algorithm was introduced it was found that among other ensemble methods, Extra-Trees significantly reduce variance and moderately increase bias, providing the best tradeoff between bias and variance [59].
- Gradient Boosting: Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. The main idea is that the subsequent predictors

learn from the errors of the previous ones. Therefore, unlike bootstrap, the observations have unequal probability of appearing in the next models, and more specifically, observations with the highest error appear most. Gradient Boosting is built on the above rationale, producing a prediction model (suitable for classification and regression) using an ensemble of weak prediction models, typically decision trees. The algorithm tries to reach a local minimum of the loss function. An analogy with gradient descent process is that gradient descent tries to update its parameter along the negative gradient direction whereas gradient boosting adds a new function to the model that moves along the negative gradient direction.

- XGBoost: The name XGBoost stands for "eXtreme Gradient Boosting". It is a scalable and accurate implementation of gradient boosting and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources for tree boosting algorithms. The implementation of XGBoost offers several advanced features for model tuning, computing environments and algorithm enhancement. It is capable of performing the three main forms of gradient boosting (Gradient Boosting (GB), Stochastic GB and Regularized GB) and it is robust enough to support fine tuning and addition of regularization parameters.

The tuning of the above algorithms is done taking into consideration that all of them are ensemble methods which are robust to overfitting. Regarding the Extra Trees Classifier 250 estimators (number of trees in the forest) are selected and 'gini index' is used for the measurement of the quality of a split. The estimators of Gradient Boosting (number of boosting stages to perform) is set to 1200, which is found to be the optimal point, as further increase does not improve the classifier's accuracy but increases its complexity. The learning rate of XGBoost is found to be optimal at 0.21. Since all three algorithms performed equally well, an ensemble between them was attempted. The three algorithms were combined using a voting classifier. The voting parameter is set to 'hard', which means that majority rule voting is implemented. Voting classifier resulted at higher classification accuracy comparing to each classifier separately.

5 Results

5.1 5-stage Classification

In this section the three datasets are classified according to AASM guidelines. The distribution of each sleep stage across the datasets is depicted in the graph below. N2 stage is dominant at all cases, as it was expected, while light and deep sleep follow the expected characteristics when comparing healthy and patient datasets, as healthier subjects tend to have a bigger percentage of deep sleep and a smaller percentage of light sleep. The pattern formed at the "Awake" stage is created randomly and has to do with the dataset cleaning, where large portions of "Awake" stage are discarded from the signal. There is class imbalance at all datasets, which will be taken into consideration at the commenting of the results. This imbalance led to the selection of the F-score metric, along with accuracy at the experiments that follow. F-score is interpreted as a weighted average of precision and recall. In the multi-class case, it is the average of the F-score of each class.

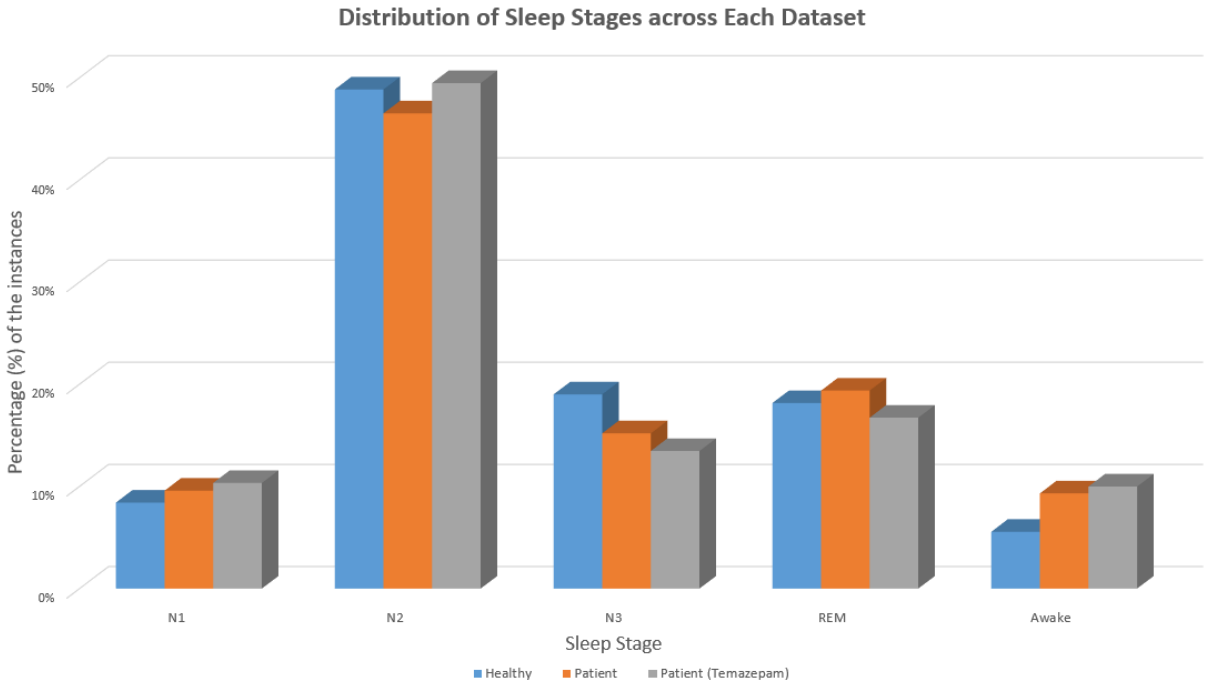

Figure 10 Percentage of each sleep stage across all datasets

5.1.1 Healthy Subjects

The classification algorithm was firstly implemented on the healthy dataset which, as expected, provided the highest accuracy and f-score.

Table 2 Healthy Dataset Classification

5–stage classification – Healthy subjects				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.79	0.66	0.72	64
Sleep Stage N2	0.94	0.95	0.94	375
Sleep Stage N3	0.92	0.95	0.94	154
REM	0.86	0.87	0.87	123
Awake	0.87	0.85	0.86	40
Avg/Total	0.91	0.91	0.91	756

Accuracy = 90.87%

F-score = 0.866

5.1.2 Patient Subjects (no medication)

The patient subjects' dataset has as a result a decrease on the accuracy and f-score.

Table 3 Patients dataset classification (no medication)

5–stage classification – Patient subjects (no medication)				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.69	0.59	0.64	70
Sleep Stage N2	0.88	0.92	0.9	378
Sleep Stage N3	0.88	0.8	0.84	128
REM	0.96	0.93	0.95	163
Awake	0.82	0.92	0.87	64
Avg/Total	0.87	0.88	0.87	803

Accuracy = 87.55%

F-score = 0.838

5.1.3 Patient Subjects (Temazepam)

The accuracy remains similar to the previous case.

Table 4 Patients dataset classification (Temazepam)

5–stage classification – Patient subjects (Temazepam medication)				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.6	0.62	0.61	64
Sleep Stage N2	0.9	0.93	0.91	390
Sleep Stage N3	0.91	0.83	0.87	109
REM	0.9	0.84	0.87	137
Awake	0.92	0.91	0.91	99
Avg/Total	0.88	0.88	0.88	799

Accuracy = 87.61%

F-score = 0.836

5.1.4 Combining Datasets

In the results that follow, it was attempted to train the model with data from the healthy subjects' dataset and then try to test the accuracy on datasets with patient subjects (with or without medication). As expected, the accuracy dramatically falls as healthy and patient subjects have totally different characteristics.

Table 5 Classification results training on healthy and testing on patient (no mediaction) subjects

5–stage classification – Train Healthy/Test Patients				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.25	0.34	0.28	385
Sleep Stage N2	0.64	0.79	0.71	1867
Sleep Stage N3	0.71	0.01	0.02	610
REM	1	0.03	0.06	778
Awake	0.26	0.79	0.39	374
Avg/Total	0.65	0.48	0.41	4014

Accuracy = 48.13%

F-score = 0.292

Table 6 Classification results training on healthy and testing on patient (Temazepam) subjects

5–stage classification – Train Healthy/Test Temazepam				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.26	0.37	0.31	413
Sleep Stage N2	0.76	0.83	0.79	1976
Sleep Stage N3	0	0	0	538
REM	0.8	0.17	0.28	668
Awake	0.27	0.76	0.4	399
Avg/Total	0.56	0.55	0.51	3994

Accuracy = 55.1%

F-score = 0.354

The next table presents the results of training the model with patient subjects without medication and then testing it to patients with temazepam. The accuracy in this case is significantly better as it reaches 75%.

Table 7 Classification results training on patients without medication and testing on patients with Temazepam

5–stage classification – Train Patients/Test Temazepam				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.56	0.31	0.4	413
Sleep Stage N2	0.83	0.83	0.83	1976
Sleep Stage N3	0.91	0.76	0.83	538
REM	0.52	0.76	0.62	668
Awake	0.8	0.76	0.78	399
Avg/Total	0.76	0.75	0.75	3994

Accuracy = 74.81%

F-score = 0.691

In the last experiment of the 5-stage classification it was attempted to merge the 3 datasets into a single one. The algorithm was trained and tested on this common dataset. The results were impressive as both accuracy and f-score were high, which means that the proposed model can predict the sleep stages of healthy and patient subjects equally well.

Table 8 Classification results using all datasets for training and testing

5–stage classification – All datasets combined				
	Precision	Recall	F1-score	Support
Sleep Stage N1	0.67	0.62	0.65	207
Sleep Stage N2	0.91	0.95	0.93	1150
Sleep Stage N3	0.93	0.86	0.89	384
REM	0.9	0.9	0.9	408
Awake	0.84	0.82	0.83	208
Avg/Total	0.88	0.89	0.89	2357

Accuracy = 88.63%

F-score = 0.84

5.2 2-stage Classification

In this section, each sleep stage is classified versus all others. This type of classification may be used when wake stage needs to be distinguished from the sleep stages, or REM sleep to be differentiated from NREM sleep, or deep sleep percentage to be extracted from the rest of the sleep.

The accuracies obtained from 2-class classification are obviously better than those obtained from multi-class classification. However, the problem of poor detection of stage N1 that is pointed out in many studies is also found here.

5.2.1 Healthy Subjects

Table 9 2-stage classification results for healthy subjects

Sleep Stage	Accuracy	F-score
N1	**93.78%**	**0.72**
N2	**93.78%**	**0.938**
N3	**97.22%**	**0.956**
REM	**95.1%**	**0.906**
Awake	**98.02%**	**0.892**

5.2.2 Patient Subjects (no medication)

Table 10 2-stage classification results for patient subjects without medication

Sleep Stage	Accuracy	F-score
N1	**93.28%**	**0.675**
N2	**89.41%**	**0.894**
N3	**95.77%**	**0.916**
REM	**96.76%**	**0.948**
Awake	**97.63%**	**0.918**

5.2.3 Patient Subjects (Temazepam)

Table 11 2-stage classification results for patient subjects with Temazepam intake

Sleep Stage	Accuracy	F-score
N1	**93.37%**	**0.663**
N2	**88.24%**	**0.882**
N3	**95.99%**	**0.909**
REM	**93.74%**	**0.875**
Awake	**96.87%**	**0.921**

6 Discussion

This section contains the discussion regarding the methods and subsequent observed results, as well as some suggestions for future pursuits in this area.

6.1 Analysis of Methods

The methods used in this thesis were chosen taking into consideration the state-of-the-art methods found in the literature about the sleep stage classification problem. The choices that were made regarding the parameters and implementation of the methods definitely had an impact on the results.

The available EEG channels at the recordings of the Sleep EDF-x Database are FPz-Cz and Pz-Oz. The channels suggested by AASM scoring rules are C4-M1, F4-M1 and O2-M1. This fact may have prevented the detection of certain patterns that would be available from these channels. According to [2], some researchers support that k-complexes are detected more easily in the frontal lobe derivations. In general, most of the existing available PSG datasets fail to meet the standards defined by AASM rules. As a consequence, the vast majority of the studies conducted so far are not compliant with AASM rules.

The features that were used for the classification models of this thesis, were chosen empirically after studying the relevant literature. The trial and error process was followed, adding and removing features, in order to achieve the optimal model. Principal component analysis was attempted at the final model but it failed to provide better classification accuracy. The technique that gave the biggest boost to the model's accuracy was the incorporation of the time sequence of the instances, described in 3.4.2. The logic behind this technique was that, during training, every instance of the EEG is considered independent from its past or future values. As it turned out an instance can be classified with greater accuracy when the previous and the next value of the signal is known. This creates a restriction for our model, since it could not be applied at a real-time sleep scoring application.

6.2 Analysis of Results

With a first look at the results it is obvious that there is some imbalance among the datasets. This is considered normal, as according to normal sleep pattern the sleep stages are not evenly distributed to a night's sleep. Stage 2 is dominant reaching almost 50% of the whole sleep duration. The discrimination between the healthy and the patients' dataset becomes clear by observing that healthy subjects tend to have a greater duration of stage 3 deep sleep.

At the 5-stage classification problem, the algorithm achieves the highest accuracy (90.8%) with the healthy subjects' dataset. Obviously, the EEGs from healthy people form patterns that are distinguished more easily and the algorithm is able to perform the classification task more efficiently. The f-score reaches 0.868 which is a sign that the classes are not predicted equally correctly due to imbalance, but this deviation was not considered very important, so class balancing techniques were not used. The accuracy score remains high when the algorithm is trained on the patients' dataset (without medication intake). Apparently, people with sleep disorders have similar EEGs, so this model was trained at sleep patterns for this type of patients achieving accuracy close to 87.5%. Following the same procedure for the Temazepam dataset the accuracy remains similar at 87.6%. The conclusion which is drawn is that sleep disorders and medication intake can distort the EEG signal and make the sleep stage classification task more difficult.

The next experiment had as a target to find out the difference between EEG signals from healthy and patient subjects. It was decided to train the model on the healthy dataset and then try to generate predictions for the others. Accuracy dramatically dropped to 48.1% and 55.1% respectively for the two patient datasets. Training on the patients' datasets and testing on the temazepam, increased the accuracy to 74.8%. It becomes clear that healthy and patient subjects have totally different EEG signal patterns and it is not possible for a model to get trained on one of them and then try to make predictions for the other.

Combining all datasets to one, was an idea towards the direction of creating one model that can predict sleep stages correctly for a larger group of people. When the algorithm was trained using all datasets, it generated predictions with an accuracy of 88.63% and an f-score of 0.84. This result is considered decent and the accuracy is satisfying, taking into consideration that the model might be "tricked" as it is trained on different datasets that are not quite similar. A

model that generalizes better can have multiple benefits. For example, the implementation of the algorithm to a wearable device used for sleep scoring would be very useful if it could be used by people who are healthy or people who face some difficulties with their sleeping habits or even use some hypnotics like temazepam.

The last experiment was about 2-stage classification. It is deduced that wake stage can be distinguished from sleep stages with a high accuracy, although there is a larger gap between accuracy and f-score as the wake stage is highly underrepresented. REM and NREM sleep seem to be easily differentiated at all datasets with a percentage above 90% which is confident. N1 stage remains problematic as it has already been pointed out by many studies. Even though the class is predicted with high accuracy, f-score is much lower. A more thorough analysis of this result can be seen at the following confusion matrix of the experiment 4.1.1. The majority of the misclassified instances that belong to class N1 are falsely classified as REM. This is justified by the fact that sleep spindles and k-complexes do not exist neither at N1 stage nor at REM stage, so these stages are difficult to distinguish.

Table 12 Confusion matrix of the experiment 4.1.1

Actual \ Predicted	N1	N2	N3	REM	W
N1	47	4	0	11	2
N2	5	359	9	2	0
N3	0	12	142	0	0
REM	5	11	0	106	1
W	3	1	0	2	34

6.3 Conclusions and Future Work

Automatic sleep scoring algorithms should ideally have confident results not only for healthy but also for patient subjects. Comparing the experiments that were conducted in this study it is found that the optimal choice is training the model with data from subjects with different EEG patterns. The tradeoff between the algorithm's accuracy and its capability of implementation at both healthy and patient subjects seems to be beneficial as accuracy is slightly decreased.

The sleep scoring problem has already reached at a high level of accuracy by many researchers. Most of the challenges have already been addressed but there still remain some slight issues. First of all, the diversity between the datasets used for the different studies, produces results that are not easily compared. As it is natural, an algorithm that achieves a high score on one dataset, may achieve a much different score on a different dataset depending on the subjects whose PSGs are being studied. The use of different channels by the datasets can also be a problem that prevents the comparison of the results. The creation of public datasets following specific guidelines e.g. AASM scoring rules would solve the above issue for future studies. Secondly, the feature engineering part of the models used for sleep staging seems to be exhaustively studied. There are many published studies that use different approaches (time, frequency, time-frequency domain) and this field seems to be adequately covered. An appropriate combination of features could probably create a model that would outperform the current ones. Finally, the majority of the classification algorithms have already been tested including deep learning methods. Deep learning has been proven less effective than expected at the sleep stage classification problem. Multi-layer models with huge computational cost, achieve similar accuracies with simple SVM or tree models, as it was shown in the literature review.

Since models have reached accuracies almost similar to human raters, the improvement of human scoring errors should be studied. Improving the ground truth of the datasets used, will eventually improve the predictions of the algorithms. According to [60], all efforts for automatic scoring are limited by the accuracy provided by the visual classification of sleep stages. It is expected that this accuracy will improve and consequently allow the successful development of computer based sleep scoring.

Even though there are many studies regarding automatic sleep scoring, there are much fewer real-life applications that the algorithms are actually used. Future researchers should be focused on the development of those applications and on the issues that could probably arise. Hospitals or sleep clinics equipped with automatic sleep scorers could faster and more easily diagnose sleep related issues to patients, since a human scorer (doctor) would no longer be necessary. In addition to that, many patients could have their own portable scoring device. Most of the algorithms used for scoring use 1 or 2 EEG channels, so the process of attaching multiple scalp electrodes can be avoided.

Bibliography

[1] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. A. Keenan, M. H. Kryger, T. Penzel, M. R. Pressman, and C. Iber. The visual scoring of sleep in adults. Journal Of Clinical Sleep Medicine, 3(2): 121–131, 2007

[2] Berry, Richard B., et al. "The AASM manual for the scoring of sleep and associated events." *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* (2012).

[3] A. A. Borb´ely and P. Achermann. Sleep homeostasis and models of sleep regulation. *J. Biol. Rhythms.,* 14(6):559-570, 1999] [J. M. Krueger, D. M. Rector, S. Roy, H. P. Van Dongen, G. Belenky, and J. Panksepp. Sleep as a fundamental property of neuronal assemblies. *Nat. Rev. Neurosci.,* 9(12):910-919, 2008

[4] R. Stickgold. Sleep-dependent memory consolidation. *Nature,* 437(7063):1272–1278, 2005

[5] M. Zokaeinikoo, 2016, "Automatic Sleep Stages Classification", Master Thesis, University of Tennessee, Knoxville

[6] A.L. Loomis, E.N. Harvey, and G.A Hobart. Cerebral States During Sleep, as Studied by Human Brain Potentials. J.Exp. Psychol., 21:127–144, 1937

[7] E. Aserinsky and N. Kleitman. Regularly Occuring Periods of Eye Motility, and Concomitant Phenomena, During Sleep. Science, 118:273–274, 1953

[8] W.C. Dement and N. Kleitman. Cyclic Variations in EEG During Sleep and their Relation to Eye Movements. Body Motility and Dreaming. Electroencephalogr. Clin. Neurophysiol., 9:673–390, 1957

[9] Šušmáková, Kristina. "Human sleep and sleep EEG." *Measurement science review* 4.2 (2004): 59-74

[10] S. Nevšímalová, K. Šonka, et al. Poruchy spánku a bdění. Maxdorf/Jessenius, Praha, 1997

[11] A. Rechtschaffen and J. M. Siegel. Sleep and Dreaming. In: Principles of Neuroscience. Edited by E. R. Kandel, J. H. Swartz and T. M. Jessel. McGraw-Hill, New York, 2000

[12] R. Stickgold. Sleep: off-line memory reprocessing. Trends in Cognitive Sciences, 2, No. 12:484–492, 1998

[13] R. Stickgold, James L. T., and J. A. Hobson. Visual discrimination learning requires sleep after training. Nature Neurosci.,3:1237–1238, 2000

[14] R. Huber, Ghilardi M.F., M. Massimini, and G. Tononi. Local sleep and learning. Nature, 430 (6995):78–81, 2004

[15] Rechtschaffen, Allan. "A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects." *Brain information service* (1968)

[16] Moser, Doris, et al. "Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters." *Sleep* 32.2 (2009): 139-149

[17] https://www.trans-cranial.com/local/manuals/10_20_pos_man_v1_0_pdf.pdf

[18] Stepnowsky, Carl, et al. "Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters." *Sleep medicine* 14.11 (2013): 1199-1207

[19] Doroshenkov, L. G., V. A. Konyshev, and S. V. Selishchev. "Classification of human sleep stages based on EEG processing using hidden Markov models." *Biomedical Engineering* 41.1 (2007): 25-28

[20] Alickovic, Emina, and Abdulhamit Subasi. "Ensemble SVM method for automatic sleep stage classification." *IEEE Transactions on Instrumentation and Measurement* 67.6 (2018): 1258-1265

[21] Estrada, E., et al. "EEG feature extraction for classification of sleep stages." *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*. Vol. 1. IEEE, 2004

[22] Yulita, Intan Nurma, et al. "Multi-Layer Perceptron for Sleep Stage Classification." *Journal of Physics: Conference Series*. Vol. 1028. No. 1. IOP Publishing, 2018

[23] Akin, Mehmet. "Comparison of wavelet transform and FFT methods in the analysis of EEG signals." *Journal of medical systems* 26.3 (2002): 241-247

[24] Fonseca, Pedro, et al. "A comparison of probabilistic classifiers for sleep stage classification." *Physiological measurement* 39.5 (2018): 055001

[25] Seifpour, Saman, et al. "A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal." *Expert Systems with Applications* 104 (2018): 277-293

[26] Memar, Pejman, and Farhad Faradji. "A novel multi-class EEG-based sleep stage classification system." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.1 (2018): 84-95

[27] Acharya, U. Rajendra, et al. "Analysis and automatic identification of sleep stages using higher order spectra." *International journal of neural systems* 20.06 (2010): 509-521

[28] Zhu, Guohun, Yan Li, and Peng Paul Wen. "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal." *REM* 806 (2014): 803

[29] Hassan, Ahnaf Rashik, Syed Khairul Bashar, and Mohammed Imamul Hassan Bhuiyan. "Automatic classification of sleep stages from single-channel electroencephalogram." *India Conference (INDICON), 2015 Annual IEEE*. IEEE, 2015

[30] Zhovna, Inna, and Ilan D. Shallom. "Automatic detection and classification of sleep stages by multichannel EEG signal modeling." *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008

[31] Gharbali, Ali Abdollahi, Shirin Najdi, and José Manuel Fonseca. "Investigating the contribution of distance-based features to automatic sleep stage classification." *Computers in biology and medicine* 96 (2018): 8-23

[32] Oropesa, Edgar, Hans L. Cycon, and Marc Jobert. "Sleep stage classification using wavelet transform and neural network." *International computer science institute* (1999)

[33] Ebrahimi, Farideh, et al. "Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients." *Engineering in medicine and biology society, 2008. EMBS 2008. 30th annual international conference of the IEEE*. IEEE, 2008

[34] Hsu, Yu-Liang, et al. "Automatic sleep stage recurrent neural classifier using energy features of EEG signals." *Neurocomputing* 104 (2013): 105-114

[35] Zhang, Junming, and Yan Wu. "Automatic sleep stage classification of single-channel EEG by using complex-valued convolutional neural network." *Biomedical Engineering/Biomedizinische Technik* 63.2 (2018): 177-190

[36] Liu, Nan, et al. "Learning a convolutional neural network for sleep stage classification." *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on*. IEEE, 2017

[37] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a

New Research Resource for Complex Physiologic Signals. *Circulation* 101(23):e215-e220 [Circulation Electronic Pages;**http://circ.ahajournals.org/content/101/23/e215.full**]; 2000 (June 13). PMID: 10851218; doi: 10.1161/01.CIR.101.23.e215

[38] Kemp, Bob, et al. "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG." *IEEE Transactions on Biomedical Engineering* 47.9 (2000): 1185-1194

[39] Fuccella, L. M. "Bioavailability of temazepam in soft gelatin capsules." *British journal of clinical pharmacology* 8.Supplement s1 (1979): 31S-35S

[40] Dijk, D. J., et al. "Effects of seganserin, a 5-HT2 antagonist, and temazepam on human sleep stages and EEG power spectra." *Eur J Pharmacol* 171.2-3 (1989): 207-18

[41] Janssen, F. H., et al. "Short-term effects of Temazepam in the EEGs of healthy volunteers." *Neuropsychobiology* 22.2 (1989): 72-76

[42] Plante, D. T., et al. "Effects of oral temazepam on sleep spindles during non-rapid eye movement sleep: a high-density EEG investigation." *European Neuropsychopharmacology* 25.10 (2015): 1600-1610

[43] Gharbali, Ali Abdollahi, Shirin Najdi, and José Manuel Fonseca. "Investigating the contribution of distance-based features to automatic sleep stage classification." *Computers in biology and medicine* 96 (2018): 8-23

[44] Estrada, E., et al. "EEG feature extraction for classification of sleep stages." *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*. Vol. 1. IEEE, 2004

[45] Davidson, Richard J. "Affective neuroscience and psychophysiology: Toward a synthesis." *Psychophysiology* 40.5 (2003): 655-665

[46] Klimesch, Wolfgang. "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis." *Brain research reviews* 29.2-3 (1999): 169-195

[47] Bao, Forrest Sheng, Xin Liu, and Christina Zhang. "PyEEG: an open source python module for EEG/MEG feature extraction." *Computational intelligence and neuroscience* 2011 (2011)

[48] Oh, Seung-Hyeon, Yu-Ri Lee, and Hyoung-Nam Kim. "A novel EEG feature extraction method using Hjorth parameter." *International Journal of Electronics and Electrical Engineering*2.2 (2014): 106-110.

[49] Türk, Ömer, et al. "Classification of mental task EEG records using Hjorth parameters." *Signal Processing and Communications Applications Conference (SIU), 2017 25th*. IEEE, 2017

[50] Cecchin, Thierry, et al. "Seizure lateralization in scalp EEG using Hjorth parameters." *Clinical Neurophysiology* 121.3 (2010): 290-300

[51] Shaker, Maan M. "EEG waves classifier using wavelet transform and Fourier transform." *brain* 2 (2006): 3

[52] Welch, A. J., P. C. Richardson, and J. N. Mockford. "Classification of sleep stage with period analysis features derived from the EEG." *Aviation, space, and environmental medicine* 49.2 (1978): 409-414

[53] Elgohary, Sahar, Seif Eldawlatly, and Mahmoud I. Khalil. "Epileptic seizure prediction using zero-crossings analysis of EEG wavelet detail coefficients." *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2016 IEEE Conference on*. IEEE, 2016

[54] Fraiwan, L., N. Khaswaneh, and Khaldon Y. Lweesy. "Automatic sleep stage scoring with wavelet packets based on single EEG recording." *World Academy of Science, Engineering and Technology* 54 (2009): 485-488

[55] Acharya, Deepshikha, et al. "Application of adaptive Savitzky–Golay filter for EEG signal processing." *Perspectives in science*8 (2016): 677-679

[56] S.J. Kern, 2017, "Automatic sleep stage classification using convolutional neural networks with long short-term memory", Master's Thesis, Radboud University, Nijmegen

[57] Yang, Yang, Xiangwei Zheng, and Feng Yuan. "A Study on Automatic Sleep Stage Classification Based on CNN-LSTM." *Proceedings of the 3rd International Conference on Crowd Science and Engineering*. ACM, 2018

[58] Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140

[59] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." *Machine learning* 63.1 (2006): 3-42

[60] Penzel, T., et al. "Sleep recording and sleep analysis methods." *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*. Springer, Berlin, Heidelberg, 2009

Figures

Figure 1: Šušmáková, Kristina. "Human sleep and sleep EEG." *Measurement science review* 4.2 (2004): 59-74

Figure 3: *https://www.trans-cranial.com/local/manuals/10_20_pos_man_v1_0_pdf.pdf*

Figure 4: Stepnowsky, Carl, et al. "Scoring accuracy of automated sleep staging from a bipolar electrooocular recording compared to manual scoring by multiple raters." *Sleep medicine* 14.11 (2013): 1199-1207

Figure 8: Gharbali, Ali Abdollahi, Shirin Najdi, and José Manuel Fonseca. "Investigating the contribution of distance-based features to automatic sleep stage classification." *Computers in biology and medicine* 96 (2018): 8-23

Figure 9: Nacy, Somer M., et al. "Controlling a Servo Motor Using EEG Signals from the Primary Motor Cortex." *American Journal of Biomedical Engineering* 6.5 (2016): 139-146