Classification of Sleep Stages for Healthy Subjects and Patients With Minor Sleep Disorders

Christos Timplalexis
School of Science and Technology
International Hellenic University
Thessaloniki, Greece
c.timplalexis@ihu.edu.gr

Konstantinos Diamantaras
Department of Information and Electronic Engineering
International Hellenic University
Thessaloniki, Greece
k.diamantaras@ihu.edu.gr

Ioanna Chouvarda
School of Medicine
Aristotle University of Thessaloniki
Thessaloniki, Greece
ioannach@auth.gr or 0000-0001-8915-6658

***Abstract*—Sleep stage classification is one of the most critical steps in the effective diagnosis and treatment of sleep-related disorders. Classic approaches involve trained human sleep scorers, utilizing a manual scoring technique, according to certain standards. This paper examines the implementation of an algorithm for the automation of the sleep scoring process. EEG recordings data are acquired from three different groups comprising of healthy subjects and people with minor sleep disorders. A mixture of time domain and frequency domain features are extracted. Temporal feature changes are utilized in order to capture contextual information of the EEG signal. Multiple classifiers are tested, culminating in a voting classifier, achieving a maximum accuracy of 90.8% for the healthy subjects' group. The main novelty introduced by the proposed solution is the algorithm's high accuracy when tested on a mixed dataset of healthy and patient subjects. The promising capabilities that derive from the successful implementation of this solution are discussed in the conclusions.**

***Index Terms*—sleep stage classification, sleep scoring, PSG, EEG**

I. Introduction

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. Sleep is fundamental for physical health and good quality of life. 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 [1]. The maintenance of health, wellbeing, homeostasis, memory and cognitive performance are also closely related to good sleep [2], [3].

Sleepers cyclically pass through different sleep stages that are defined by rules based on Rechtschaffen and Kales' (R&K) recommendations [4], or a more recent guideline developed by the American Academy of Sleep Medicine (AASM) [5], which was used in the current study. According to it, there are 5 different sleep stages (Wake (W), Stage 1 (N1), Stage 2 (N2), Stage 3 (N3) and REM) with the typical length of each five stages cycle being 90 to 110 minutes [6], [7].

The dominant characteristics of each stage are analyzed below [8], [9]:

- W stage is characterized by alpha frequency bands as well as frequent eye movements.
- N1 corresponds to light sleep. It is characterized by alpha or faster frequency bands occupying more than 50% of the epoch, while theta activity and slow eye movements are evident.
- In N2, the eyes stop moving, the brain waves become slower and sleep spindles or k-complexes are noted.
- N3 corresponds to deep sleep, where no eye movement and muscle activity exist, while delta activity is detected in over 20% of the epoch length.
- In REM stage the breathing rate increases and the eyes move rapidly.

Sleep stage classification was introduced almost 50 years ago and manual approaches are sometimes used until today. 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 polysomnography (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. Experts' experience is also a crucial factor that leads to a considerable percentage of disagreement between sleep scorers [10], [11]. In fact, according to [12] in the best of circumstances the experts agree at a percentage of 90%. Researchers have been attempting to propose methods that automate the sleep stage classification process by utilizing signal processing and machine learning techniques that are able to increase consistency and reliability, assisting experts at diagnosing sleep related health problems. The automation of sleep scoring makes the process much faster and most importantly it does not require to occupy a human expert. The current paper proposes an algorithm for automatic sleep stage classification, emphasizing on the algorithm's performance when tested on either healthy or patient (minor sleep disorder) subjects.

2471-7819/19/$31.00 ©2019 IEEE
DOI 10.1109/BIBE.2019.00068

This paper is organized as follows: The related work is presented in Section II, while Section III describes the data and methods that were implemented. Section IV presents and analyzes the experimental results. Section V discusses what has been presented in the paper and proposes potential future work. Finally, conclusions are drawn in Section VI.

II. Related Work

Automated sleep stage scoring is a topic extensively studied in the literature especially in recent years, given the computational power capabilities that derive from the development in the field of machine learning and deep learning. Acquisition of sleep data is not always feasible due to privacy issues, so most of the studies use data found at online data repositories (e.g. Physionet). The volume of the data is also an issue, since data sampling usually ranges between 100-256 Hz and a whole sleep recording lasts for about 8 hours. As a result of the big data volume, in most cases, a quite small number of sleep recordings is being used, even less than 10.

Starting with shallow machine learning approaches, in [13] 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 [14]. 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 [15], 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 [16], 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 [17], 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 [18]. A total of 443 recordings were used using healthy subjects and patients diagnosed with apnea. 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 novelty of [9] 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. In [19], 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 [20], 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. 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.

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 major improvement compared to classic machine learning approaches. One of the first classification methods that used neural networks is described in [21]. The study was published in 1999 and to the authors' knowledge this is the first research that used Wavelet 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 [22], 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 [23] 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 [24]. 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 deep sleep. A 3-dimensional signal is constructed from EEG signal, F-DFT signal and WT signal. The signal is fed into the CNN where the total accuracy for the 6-stage classification is 90.11%.

III. Data and Methods

A. Data Acquisition and Preprocessing

For the purpose of this study, data were obtained from Physionet online data repository [25]. There are several polysomnographic sleep recording libraries, however the Sleep-EDF Database [Expanded] was used for this study [26]. The database 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. 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 in this research. Each subject was recorded for 2 nights, one of which was after temazepam intake and the other after placebo intake.

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, [4] although 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+. R&K scoring is converted into AASM scoring for the needs of the current study.

It was decided to split the subjects into 3 different groups (Table I). 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 subjects (sleeps) were used for each group.

TABLE I
Subject Groups Used In The Study

Group	Subjects
Group A	Healthy Subjects
Group B	Mild difficulty falling asleep - no medication
Group C	Mild difficulty falling asleep - temazepam intake

Temazepam belongs to the class of medications called benzodiazepines and it is used for the treatment of short-term sleeping problems. The effects of temazepam on human EEG have been studied by several researchers in the past. In [27], 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 [28], 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. It is consequently deduced that the separation between placebo and drug intake nights is meaningful, as temazepam changes the EEG characteristics and sleep structure.

In the current study, the sleep scoring standard of the American Academy of Sleep Medicine was used [5]. This standard is 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. 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 nights 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).

A preprocessing 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 [29] [30] 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 models accuracy dropped by 3-4%. This means that some useful information of the signal was lost, trying to smoothen the noise.

B. Feature Engineering

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, consequently 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 [31] as they reveal different aspects of the EEG, EMG and EOG signals. 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.

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

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

$$Var(X) = \frac{\sum_{i=1}^{i=N} \left(X_i - E\left(X\right)\right)^2}{N} \tag{2}$$

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).

$$Skew = \frac{1}{N} \sum_{i=1}^{i=N} \left[\frac{X_i - E\left(X\right)}{\sigma}\right]^3 \tag{3}$$

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).

$$Kurt = \frac{1}{N} \sum_{i=1}^{i=N} \left[\frac{X_i - E\left(X\right)}{\sigma}\right]^4 \tag{4}$$

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.

$$Entropy = -\sum_{i=1}^{i=N} p_i \log(p_i) \tag{5}$$

EEG Frequency Bands: EEG waveforms are classified into five different frequency bands (Fig. 1). 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 [15] to epilepsy detection, human emotion recognition [32] and cognitive performance [33].

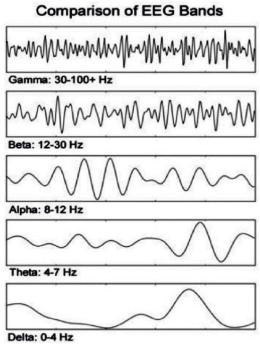

Fig. 1. Comparison of EEG bands

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.

$$AveragePower = \frac{1}{N} \sum_{i=1}^{i=N} |x(n)|^2 \tag{6}$$

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 [34] which is an open-source python module for EEG/MEG feature extraction.

2) Time dependence of PSG recordings: The EEG and EOG signals that were utilized for feature extraction comprise of discrete data points that are 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 [35] [36]. In this study, a similar approach is used, as the existing features are shifted in time. More specifically, for each feature F_i, three new features are created: $F_i - 1$, F_i, $F_i + 1$, where i is the timestamp of the feature. Finally, the proposed model ends up with 108 features. The described process is explained in fig.2.

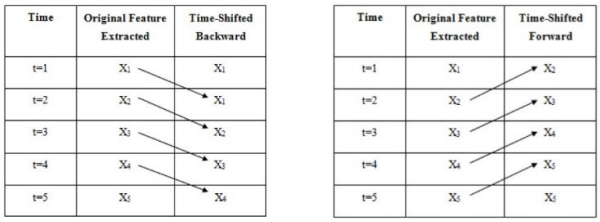

Fig. 2. Backward and forward feature time-shifting

C. Classification Algorithm

Before proceeding to the classification algorithm, the features are scaled to reduce intersubject variability, regarding their magnitude, range and units. Since the data distribution is not Gaussian, the min-max scaler is selected, scaling the data in the range [0,1]. The formula followed by min-max scaler is shown in eq. 7:

$$y_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{7}$$

Another step that led to a more accurate model is feature selection. It is commonly used to reduce overfitting, reduce training time and improve accuracy in general since misleading information is ignored. The score function that determined the most relevant features is chi-squared. Chi-square test measures dependence between stochastic variables, so the use of this function reveals the features that are most likely to be independent of class and therefore irrelevant for classification.

Finally, the model's performance was evaluated by being tested with many different classifiers. As expected, Gaussian Naive Bayes and Logistic Regression provided relatively poor results, compared to other more sophisticated methods. SVM also achieved low accuracy, even though it has been successfully used on the sleep stage scoring problem [14]. LSTM neural network configuration was also tried but it did not provide optimal results and was also computationally expensive. The three algorithms that achieved the highest accuracy were: Extremely Randomized (Extra) Tree Classifier, Gradient Boosting and XGBoost.

The Extra trees algorithm builds an ensemble of unpruned decision 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 [37].

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 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 is a scalable and accurate implementation of Gradient Boosting, built to enhance the model's performance and computational speed. XGBoosts main difference with Gradient Boosting is that it uses a more regularized model formalization to control overfitting, resulting in better performance.

Grid search is used to find the optimal parameters for the tuning of each model. The three top classifiers belong to the family of tree-based methods. Since all models result in high accuracy a voting classifier was tried in order to test whether the combination of these three models could provide even better results. Hard voting was applied which means that the majority rule voting was implemented with the final voting classifier resulting in higher classification accuracy comparing to each model separately. The algorithm's flow chart is shown in Fig.3.

IV. Results

In this section, sleep stage classification results are presented, studying the following cases:

- The algorithm is tested on groups A, B and C separately.
- The algorithm is trained on one group and tested on another group, trying to verify differences on EEG characteristics among the three groups.

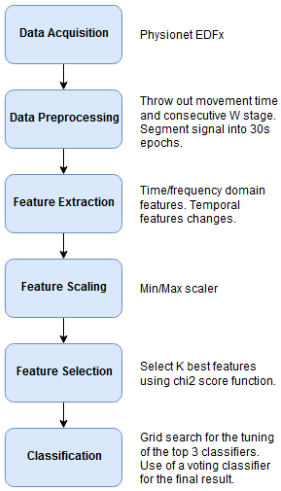

Fig. 3. Algorithm flow chart

- The algorithm is tested on an aggregated dataset of the three groups, aiming at providing a solution appropriate for healthy and patient subjects.

A. *Separate Groups*

The sleep stages of the three study groups are classified according to AASM standards. The distribution of each sleep stage across the groups is depicted in Fig. 4. 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 groups, as healthier subjects tend to have a bigger percentage of deep sleep and a smaller percentage of light sleep. The sleep stage imbalance is explained as according to normal sleep pattern the stages are not evenly distributed to a nights sleep. Stage 2 is dominant reaching almost 50% of the whole sleep duration. The discrimination between the healthy and the patients groups becomes clear by observing that healthy subjects tend to have a greater duration of stage 3 deep sleep. The class imbalance led to the selection of the F-score metric, along with the 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.

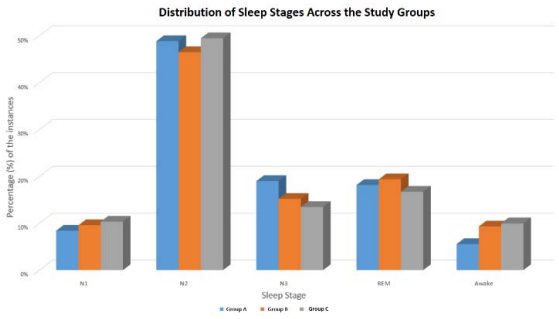

Fig. 4. Percentage of sleep stage across all study groups

The 5-stage classification results (accuracy, f-score) are presented in Table II, for each group separately:

TABLE II
ACCURACY AND F-SCORE FOR EACH GROUP

Group	Accuracy	F-score
Group A (Healthy)	90.87%	0.860
Group B (Placebo)	87.42%	0.835
Group C (Temazepam)	87.85%	0.839

Tables III, IV, V display a more detailed representation of the evaluation metrics of each sleep stage.

TABLE III
EVALUATION METRICS FOR GROUP A

	Precision	Recall	F-score	Support
Sleep stage N1	0.77	0.62	0.69	64
Sleep stage N2	0.93	0.96	0.95	375
Sleep stage N3	0.93	0.95	0.94	154
REM	0.88	0.87	0.87	123
Awake	0.85	0.85	0.85	40
Avg/Total	0.91	0.91	0.91	756

TABLE IV
EVALUATION METRICS FOR GROUP B

	Precision	Recall	F-score	Support
Sleep stage N1	0.70	0.56	0.62	70
Sleep stage N2	0.88	0.91	0.90	378
Sleep stage N3	0.87	0.82	0.84	128
REM	0.96	0.93	0.94	163
Awake	0.80	0.95	0.87	64
Avg/Total	0.87	0.87	0.87	803

TABLE V
EVALUATION METRICS FOR GROUP C

	Precision	Recall	F-score	Support
Sleep stage N1	0.61	0.64	0.63	64
Sleep stage N2	0.90	0.94	0.92	390
Sleep stage N3	0.90	0.82	0.86	109
REM	0.91	0.86	0.89	137
Awake	0.92	0.90	0.91	99
Avg/Total	0.88	0.88	0.88	799

At the 5-stage classification problem, the algorithm achieves the highest accuracy (90.87%) when tested on healthy subjects (group A). 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 0.86 which is a sign that the classes are not predicted with the same success due to imbalance. The accuracy drops when sleep staging is done for group B and group C, as sleep disorders and temazepam intake seem to distort the EEG signal in a way that drops the predictive accuracy approximately by 3%. At all cases, it is confirmed that stage N1 is not predicted with the same success as the other sleep stages, which is a finding that is supported by

most of the studies found in the literature. Moreover, it was observed that stage N1 is usually confused with REM stage, which is a reasonable hypothesis as these two stages share common characteristics such as eye movements.

B. Training and Testing on Different Groups

Trying to point out the differences between the three study groups, it was attempted to train the algorithm on one group and then test it on the others. When Group A was used for training, the accuracy dropped dramatically, showing that sleep staging training dataset may set limits to the algorithm's implementation. The accuracy also drops when training is done on Group B and testing on Group C, but since both groups include subjects with sleep disorders the score is higher comparing to the first case. The poor results obtained by this experiment lead us to the conclusion that healthy and patient subjects have different EEG characteristics so training and testing on different study groups is not possible.

TABLE VI
ACCURACY AND F-SCORE WHEN TRAINING AND TESTING IS DONE ON DIFFERENT GROUPS

Group	Accuracy	F-score
Train on Group A/Test on Group B	48.13%	0.292
Train on Group A/Test on Group C	55.10%	0.354
Train on Group B/Test on Group C	74.81%	0.691

C. Aggregated Dataset

The final experiment included the merging of the three study groups on a single dataset. Then, the suggested algorithm was trained and tested on the aggregated dataset. The successful implementation of the method is significant considering that real-world applications should be applicable to a large part of the population and not only to a single group (e.g. healthy subjects). The features used at the predictive model combined with the voting classifier seem to capture the variations between study groups that have different EEG waveforms, maintaining predictive accuracy at high levels as seen in Tables VII and VIII.

TABLE VII
ACCURACY AND F-SCORE OF THE AGGREGATED DATASET

Group	Accuracy	F-score
Group A + Group B + Group C	88.88%	0.847

TABLE VIII
EVALUATION METRICS FOR THE AGGREGATED DATASET

	Precision	Recall	F-score	Support
Sleep stage N1	0.69	0.67	0.68	207
Sleep stage N2	0.91	0.95	0.93	1150
Sleep stage N3	0.92	0.87	0.89	384
REM	0.90	0.89	0.90	408
Awake	0.85	0.82	0.84	208
Avg/Total	0.89	0.89	0.89	2357

V. DISCUSSION

The aim of the present study was to propose a reliable sleep stage classification algorithm to deal with the difficulties that arise on the manual sleep scoring process. A series of novelties are introduced towards this direction. Starting with the feature extraction process, time and frequency domain features carrying information for the non-stationary EEG signal were selected, while overfitting was avoided. Temporal feature changes were incorporated which boosted the model's accuracy, revealing the high time-dependence among the data points of the EEG and EOG signals. The voting classifier was also a key factor for the effectiveness of the suggested solution. Using majority voting among the top three classifiers finally improved even more the model's accuracy, avoiding the extreme increase in the algorithm's complexity and execution time. The major novelty of the proposed algorithm is derived from the experimental results that showcase the successful implementation of the algorithm on all of the study groups. Most of the studies found in the literature only use data from healthy subjects, even though the main application of sleep stage classification is its diagnostic aspect. The current algorithm's successful evaluation on healthy subjects and patients with or without benzodiazepines intake, makes it applicable to sleep staging diagnoses.

Even though there is a plethora of studies regarding the sleep stage classification problem, there still remain some issues that need to be addressed. First of all, the diversity between the datasets used for the different studies, leads to results that are not easily compared. The use of different EEG channels by the datasets can also be a problem that prevents the comparison of the results. Secondly, the majority of machine learning and deep learning approaches have already been tested and the models' accuracy reaches the accuracy achieved by human experts. The possible improvement of human scoring errors should be studied as improving the ground truth of the datasets will eventually improve the performance of the classification models. Finally, 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 of 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.

VI. CONCLUSIONS

In this work, an alternative in the sleep stage classification problem is presented. It constitutes a solution that utilizes a mixture of appropriate time and frequency domain features incorporating also contextual EEG information. An ensemble of tree based classifiers is used for the final classification prediction. The proposed solution achieves results that rank

among the state-of-the-art models found in the literature. Moreover, the suggested algorithm provides accurate results when tested on both healthy and patient (minor difficulty falling asleep with or without drug intake) subjects. The main limitation of the current study is its appropriateness for real-time sleep scoring, due to the use of temporal features changes, which would require a delay of at least 30 seconds.

REFERENCES

[1] M. Zokaeinikoo, "Automatic sleep stages classification," Master's thesis, University of Tennessee, Knoxville, 2016.

[2] A. A. Borb and P. Achermann, "Sleep homeostasis and models of sleep regulation," *Journal of biological rhythms*, vol. 14, no. 6, pp. 559–570, 1999.

[3] R. Stickgold, "Sleep-dependent memory consolidation," *Nature*, vol. 437, no. 7063, p. 1272, 2005.

[4] A. Kales, Anthony Rechtschaffen, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. United States: Bethesda, Md., U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968.

[5] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. Vaughn *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*, vol. 176, 2012.

[6] P. Tian, J. Hu, J. Qi, X. Ye, D. Che, Y. Ding, and Y. Peng, "A hierarchical classification method for automatic sleep scoring using multiscale entropy features and proportion information of sleep architecture," *Biocybernetics and Biomedical Engineering*, vol. 37, no. 2, pp. 263–271, 2017.

[7] M. A. Carskadon, W. C. Dement *et al.*, "Normal human sleep: an overview," *Principles and practice of sleep medicine*, vol. 4, pp. 13–23, 2005.

[8] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Computer methods and programs in biomedicine*, vol. 140, pp. 77–91, 2017.

[9] P. Memar and F. Faradji, "A novel multi-class eeg-based sleep stage classification system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 84–95, 2017.

[10] A. Malhotra, M. Younes, S. T. Kuna, R. Benca, C. A. Kushida, J. Walsh, A. Hanlon, B. Staley, A. I. Pack, and G. W. Pien, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, 2013.

[11] N. A. Collop, "Scoring variability between polysomnography technologists in different sleep laboratories," *Sleep medicine*, vol. 3, no. 1, pp. 43–47, 2002.

[12] C. Stepnowsky, D. Levendowski, D. Popovic, I. Ayappa, and D. M. Rapoport, "Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters," *Sleep medicine*, vol. 14, no. 11, pp. 1199–1207, 2013.

[13] L. Doroshenkov, V. Konyshev, and S. Selishchev, "Classification of human sleep stages based on eeg processing using hidden markov models," *Biomedical Engineering*, vol. 41, no. 1, pp. 25–28, 2007.

[14] E. Alickovic and A. Subasi, "Ensemble svm method for automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, 2018.

[15] E. Estrada, H. Nazeran, P. Nava, K. Behbehani, J. Burk, and E. Lucas, "Eeg feature extraction for classification of sleep stages," in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1. IEEE, 2004, pp. 196–199.

[16] I. N. Yulita, R. Rosadi, S. Purwani, and M. Suryani, "Multi-layer perceptron for sleep stage classification," in *Journal of Physics: Conference Series*, vol. 1028, no. 1. IOP Publishing, 2018, p. 012212.

[17] M. Akin, "Comparison of wavelet transform and fft methods in the analysis of eeg signals," *Journal of medical systems*, vol. 26, no. 3, pp. 241–247, 2002.

[18] P. Fonseca, N. Den Teuling, X. Long, and R. M. Aarts, "A comparison of probabilistic classifiers for sleep stage classification," *Physiological measurement*, vol. 39, no. 5, p. 055001, 2018.

[19] G. Zhu, Y. Li, and P. P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal," *IEEE journal of biomedical and health informatics*, vol. 18, no. 6, pp. 1813–1821, 2014.

[20] A. R. Hassan, S. K. Bashar, and M. I. H. Bhuiyan, "Automatic classification of sleep stages from single-channel electroencephalogram," in *2015 annual IEEE India conference (INDICON)*. IEEE, 2015, pp. 1–6.

[21] E. Oropesa, H. L. Cycon, and M. Jobert, "Sleep stage classification using wavelet transform and neural network," *International computer science institute*, 1999.

[22] 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, pp. 105–114, 2013.

[23] J. Zhang and Y. Wu, "Automatic sleep stage classification of single-channel eeg by using complex-valued convolutional neural network," *Biomedical Engineering/Biomedizinische Technik*, vol. 63, no. 2, pp. 177–190, 2018.

[24] N. Liu, Z. Lu, B. Xu, and Q. Liao, "Learning a convolutional neural network for sleep stage classification," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2017, pp. 1–6.

[25] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[26] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[27] D. Dijk, D. Beersma, S. Daan, and R. Van den Hoofdakker, "Effects of seganserin, a 5-ht2 antagonist, and temazepam on human sleepsstages and eeg power spectra," *European journal of pharmacology*, vol. 171, no. 2-3, pp. 207–218, 1989.

[28] F. Janssen, L. Beecher, P. Griep, and A. Declerck, "Short-term effects of temazepam in the eegs of healthy volunteers," *Neuropsychobiology*, vol. 22, no. 2, pp. 72–76, 1989.

[29] L. Fraiwan, N. Khaswaneh, and K. Lweesy, "Automatic sleep stage scoring with wavelet packets based on single eeg recording," *World Academy of Science, Engineering and Technology*, vol. 54, pp. 485–488, 2009.

[30] D. Acharya, A. Rani, S. Agarwal, and V. Singh, "Application of adaptive savitzky–golay filter for eeg signal processing," *Perspectives in science*, vol. 8, pp. 677–679, 2016.

[31] A. A. Gharbali, S. Najdi, and J. M. Fonseca, "Investigating the contribution of distance-based features to automatic sleep stage classification," *Computers in biology and medicine*, vol. 96, pp. 8–23, 2018.

[32] R. J. Davidson, "Affective neuroscience and psychophysiology: Toward a synthesis," *Psychophysiology*, vol. 40, no. 5, pp. 655–665, 2003.

[33] W. Klimesch, "Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain research reviews*, vol. 29, no. 2-3, pp. 169–195, 1999.

[34] F. S. Bao, X. Liu, and C. Zhang, "Pyeeg: an open source python module for eeg/meg feature extraction," *Computational intelligence and neuroscience*, vol. 2011, 2011.

[35] S. J. Kern, "Automatic sleep stage classification using convolutional neural networks with long short-term memory," 2017.

[36] Y. Yang, X. Zheng, and F. Yuan, "A study on automatic sleep stage classification based on cnn-lstm," in *Proceedings of the 3rd International Conference on Crowd Science and Engineering*. ACM, 2018, p. 4.

[37] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.