EAI Endorsed Transactions
on Bioengineering and Bioinformatics

Research Article EAI.EU

Towards exploration and evaluation of sleep staging classification schemes for healthy and patient subjects

Christos Timplalexis[1,*] and Dimitrios Chasanidis[2] and Ioanna Chouvarda[3] and Konstantinos Diamantaras[2]

[1]Information Technologies Institute/ Centre for Research and Technology Hellas, 6th Km Charilaou-Thermis 57001, Thermi-Thessaloniki, Greece
[2]Department of Information and Electronic Engineering, International Hellenic University, Sindos, 57400, Greece
[3]Lab of Computing, Medical Informatics and Biomedical Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece

Abstract

INTRODUCTION: Sleep stage classification is an important task for the timely diagnosis of sleep-related disorders, which are one the most common indicator of illness.

OBJECTIVE: An automated sleep scoring implementation with promising generalization capabilities is presented, aiding towards eliminating the tedious procedure of manual sleep scoring.

METHODS: Two Electroencephalogram (EEG) channels and the Electrooculogram (EOG) channel are utilized as inputs for feature extraction both in the time and frequency domain, while temporal feature changes are utilized in order to capture contextual information of the signals. An ensemble tree-based and a neural network approach are presented at the classification process.

RESULTS: A total of 66 subjects belonging to three different groups (healthy, placebo, drug intake) were included in the study. The tree-based classification method outperforms the neural network at all cases.

CONCLUSION: State of the art results are achieved, while it is highlighted that using jointly the healthy and patient subjects dataset, boosts the model's accuracy and generalization capability.

Received on 30 June 2020; accepted on 01 October 2020; published on 19 October 2020

Keywords: sleep staging, sleep classification, EEG, PSG, AASM

doi:10.4108/eai.19-10-2020.166665

1. Introduction

Sleep is a state characterized by loss of consciousness and greatly reduced responsiveness to external stimuli. It is distinguished from coma or anaesthesia by its rapid reversibility [1, 2]. The maintenance of a human's well-being and cognitive performance as well as mood and behavior are affected by the quality of sleep. Sufficient sleep is an essential ingredient for good health, that may help significantly avoiding disorders such as insomnia, sleep apnea, hypersomnia, circadian rhythm sleep-wake disorders etc. The main tool for monitoring sleep quality and diagnosing sleep disorders is PSG recordings. PSG is comprised of physiological signals such as EEG, EOG and Electromyogram (EMG) that are utilized to assess an individual's sleep, by assigning a specific sleep stage to a sleep epoch of predefined time duration. This assessment, namely sleep staging, is usually conducted by a trained human expert, consequently it is a subjective process. It is a demanding

[*]Corresponding author. Email: ctimplalexis@iti.gr

task that requires considerable work and it is affected by the scorer's experience and fatigue. The percentage of agreement between two human scorers is often below 90% highlighting the need for automated scoring [3, 4]. Sleep scoring is following specific rules, based on some predefined standards. Two of the most widely used standards are Rechtschaffen and Kales' (R&K) rules [5] published in 1968 and a more recent guideline revised in 2012 by the American Academy of Sleep Medicine (AASM) [6]. The current study is based on the AASM standard which includes the following sleep stages that alternate cyclically every 90 to 110 minutes, during a human's sleep [7, 8]:

- Awake stage (W) stage is characterized by alpha frequency bands as well as frequent eye movements.
- N1 stage 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 Rapid Eye Movement (REM) stage the breathing rate increases and the eyes move rapidly.

According to the standard, the scorer must assign a sleep stage every 30 seconds for the whole duration of a subject's sleep (7-8 hours). The process cannot be easily repeated due to its high cost, the subject's inconvenience and the scorer's demanding task. Moreover, PSG recordings may severely differ from person to person depending on age, health condition, sleep condition etc. It is common that even the same person has different PSG recordings on different days, mainly due to different health and sleeping condition. Machine learning algorithms have been applied to the sleep scoring problem for many years, offering automated sleep scoring services. Thanks to the great progress made during the recent years in the field of computational power capabilities, more complex deep learning approaches have been employed, delivering promising results. Despite the above, the vast majority of sleep physicians still shows disbelief on sleep scoring algorithms, picking the tedious task of manual scoring over the automated process [9]. From the machine learning point of view, sleep staging is an unbalanced classification problem. Class imbalance is merely the cause for different predictive accuracy at each class.

The current study extends the work presented in [10] in terms of the number of subjects used in the experiments, the classification methods, as well as the comparison of two different evaluation methods.

The rest of the paper is organized as follows: The related work is presented in Section 2, while Section 3 describes the data and methods that were implemented. Section 4 presents and analyzes the experimental results. Section 5 discusses what has been presented in the paper and proposes potential future work. Finally, conclusions are drawn in Section 6.

2. Related work

Automated sleep stage scoring is a topic that is gradually getting more attention in the literature, as computational power increasing capabilities give the opportunity for more complex and computationally expensive models. Although access to sleep lab data is restricted, a number of recordings have been made available in online repositories [11–13] serving as research datasets for a number of studies. Data sampling frequency usually ranges between 100-256 Hz while a whole night's sleep recording lasts for about 8 hours, resulting into a massive amount of data . As a result, in most studies a relatively small number of sleep subjects is being used, even less than 10.

Reviewing the literature it is deduced that benchmarking and comparing different studies is a challenging task, since there are a number of parameters determined by the authors when defining the problem. First of all, the available datasets are comprised of different subjects in terms of age, sex, sleep disorders etc. Some of the most well-known datasets are available online, like Sleep-EDFx [11], CAP sleep dataset [12], St. Vincent's University Hospital / University College Dublin Sleep Apnea Database [14]. There are also datasets provided by sleep clinics or universities that do not have free access like [15]. All of the datasets contain multiple subjects and each study usually selects a subset of those subjects ignoring the rest. Sleep is affected by many factors such as age or health condition, so different subjects may provide results that vary significantly even if the same algorithm is used. Some datasets may also contain both healthy and patient subjects, while patients, depending on the dataset, may suffer from sleep apnea, insomnia, mild disorders, REM sleep behavior disorder etc. The results that derive from patients with different disorders are obviously not comparable with each other. Another issue that arises is that a PSG may include different EEG channels, and potentially EOG and EMG. Even if the most common data source is used, namely EEG, data may derive from different channels in terms of the electrodes positioning on the subjects' head. In some cases, even the data sampling rate varies as well. The ground truth of each dataset may follow a specific standard, but as it has already been pointed out in the introduction, there is a big percentage of disagreement between experts.

Taking a closer look at the approaches followed by sleep staging studies, the feature extraction process seems to include three main type of features: i) time-domain, ii) frequency-domain, iii) non-linear features. Time-domain features include statistical features such as first, second and higher order statistics of the time-series raw signal [16, 17]. Fourier transform (FT) has been extensively used for feature extraction in the frequency domain [18, 19]. Recent studies tend to use wavelet transform (WT) more often as it has a strong advantage compared to FT, since it is able to localize the frequency components into the time-domain [19, 20]. Non-linear features are commonly used with EEG data since they are able to portray the the non-linear dependencies of different parameters associated with EEG [21].

Another key point that can be used to make a fair discrimination of the studies found in the literature, is the method that is used for the evaluation of the suggested model. Two methods are predominantly used. For the context of this study, the first method will be named Internal Subjects Evaluation and the second, External Subjects Evaluation. In internal evaluation a part of each subject's sleep is used for training and the rest sleep is kept for testing. In external evaluation, the model's testing is done on totally unseen subjects. The type of testing is not explicitly declared by all studies, as cross-validation may imply the first or the second method, depending on whether the tested subjects are completely kept out of the training process (leave-one-out cross validation).

Regarding the classification methods that are used to classify sleep stages, earlier approaches were predominantly based on Support Vector Machine (SVMs) [16, 22, 23], Decision Trees (DTs) [24–26] and Hidden Markov Models (HMMs) [15, 27] while most recent approaches are based on Neural Networks. Convolutional (CNN) [28, 29] are the most commonly found architectures, achieving the highest accuracy.

Starting with shallow machine learning approaches, in [27] HMM is used trying to correlate the transitions between sleep stages. It is deduced than predictive accuracy varies among sleep stages. Stage 1 is underrepresented, as it represents a small part of a whole night's sleep. Its accuracy is usually lower compared to the other stages. In the current study it is below 50% while the rest stages are close to 90%. The same issue regarding the accuracy of sleep stage 1 is also noticed at [16], where a set of statistical features of the Pz-Oz electrode are fed into an SVM classifier. In [18], frequency domain features such as spectral energy band, central frequency, bandwidth and Itakura Distance are evaluated in the context of the sleep stage classification task. The extraction of multiple features in the time and frequency domain may lead to models that have better generalization capabilities. In [30], EEG, EOG and EMG signals are converted into the frequency domain and band features were extracted. MLP classifier outperformed other approaches, but the accuracy obtained remains relatively low, as the study was conducted on patients with sleep apnea. A comparison between FT and WT is attempted in [19]. It is found that WT is more efficient mainly due to the fact that EEG signals are non-stationary, so small changes may not be realized by FT and the analysis may change depending on the length of data. A method totally based on WT performed on the EOG signal is introduced in [31]. Db4 is selected as the mother wavelet and features are extracted from the WT's detail and approximation coefficients. SVM and tree-based under-sampling boosting classifiers were used while internal validation was carried out. A comparison of probabilistic classifiers, on healthy and patients subjects, using external validation is analyzed in [15]. Conditional Random Fields (CRF) classifiers are proved to be superior, providing moderate sleep stage classification results for patients with apnea, outperforming earlier work. However, stage 1 low predictive accuracy is also highlighted by the authors. The authors of [25] put more emphasis on the feature selection process, performing feature selection using mRMR, after extracting feature importances with the Kruskal-Wallis statistical test. Very high accuracy is obtained for the 6-class and 5-class sleep stage classification problem, as a result of the internal validation. Moreover, the wake stage is not discarded from the EEG recordings, boosting the model's accuracy. An alternative approach is presented in [23], as graph domain features are extracted from an EEG channel. The mapping of the signal segments into visibility graphs ends up into an SVM classifier performing internal validation.

Deep learning approaches have been extensively tried out in the field of sleep stage classification, as the majority of the published studies over the last few years are based on ANNs. The first study that utilized WT for feature extraction relied on a feedforward NN for the classification task [20]. Computational power capabilities at that time only allowed the existence of one hidden layer with 10 neurons, while the input layer comprised of 13 neurons. The method provided relatively poor results but paved the way for future deep learning application in the sleep staging field. Coming to more recent researches, in [28] the implementation of a complex-valued CNN is examined. The selection of CNN is backed by the claim that the construction of handcrafted features able to reveal information about sleep patterns, is a process that requires an experienced domain expert. Complex

CNN outperforms classical CNN approaches but poor accuracy for stage N1 remains a problem. The authors of [29] are also presenting a CNN approach utilizing Fractional Discrete Fourier Transform (F-DFT) in order 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. State-of-the-art performance is achieved but the model is tested only with internal validation. A comparison of three different NN classifiers is introduced in [32]. A recurrent classifier, a feedforward NN and a probabilistic NN are compared. As expected, temporal information enclosed in the PSG recording time-series data, boosts the recurrent model's accuracy making it by far more accurate than the others. The same logic also applies to [33], where LSTM is utilized for classification. The most encouraging finding after the internal validation process, is the improvement of stage 1 predictive accuracy, compared to other methods.

3. Experimental Method

3.1. Data Acquisition and Preprocessing

The dataset used in this paper is Sleep-EDF [Expanded] Database which is publicly available online from PhysioNet [14]. The database contains 197 PSGs with accompanying hypnograms, while the data are acquired from two studies. The first one named SC* is the study of different age effects while sleeping. 153 recordings from healthy Caucasian males aged 25-101 are found in this study. The second one named ST* is the study of temazepam effects while sleeping in 22 Caucasian subjects without other medication, and in that study, subjects had mild difficulty falling asleep but were otherwise healthy. Each subject was recorded for 2 nights, one of which was after temazepam intake and the other after placebo intake.

The recordings contain two EEG channels (from Fpz-Cz location and Pz-Oz locations), EOG (horizontal), submental chin EMG, and an event marker. EEG and EOG signals were sampled at 100Hz. Each recording was scored by well-trained experts according to the R&K manual, but based on Fpz-Cz/Pz-Oz EEGs instead of the suggested C4-A1/C3-A2 (Fig. 1) [5]. Annotations of every 30-second epoch contain 1, 2, 3, 4, W, R, M and ?, which represent stages S1, S2, S3, S4, Awake, REM, 'Movement' and 'not scored' respectively. The hypnograms are converted to the AASM scoring standard for the needs of the current study. Movement data and not scored epochs were completely disregarded from the dataset.

The subjects were separated into 4 diferrent groups (Table 1). The first group contains totally healthy subjects, obtained from the SC* study of Sleep-EDFx dataset. The second and third group contain subjects from ST* study, that have difficulty falling asleep. The second group contains the placebo intake nights, while the third group corresponds to the temazepam intake nights. Finally, in group 4 all of the three groups described above are joined into one group containing both healthy and patients subjects.

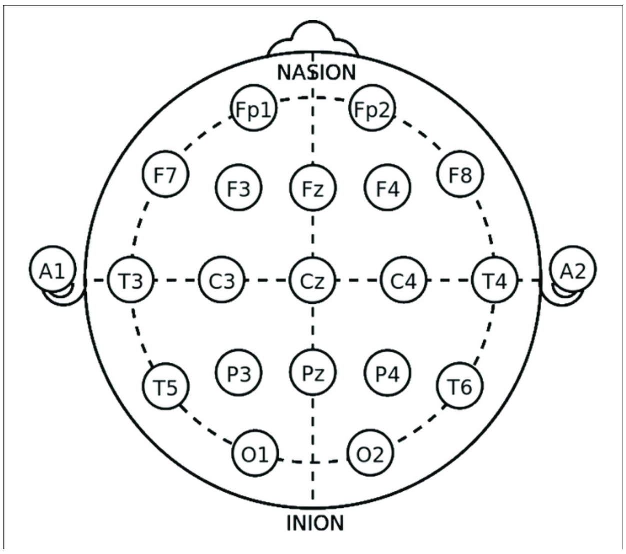


Figure 1. The 10–20 electrodes positioning system

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 [34], twenty healthy males aged 21-26 years were recorded 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 [35], 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 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 [6], as this is the standard that is followed by most of the recent studies. An inconsistency between AASM

Table 1. Subject Groups Used In The Study

Group	Subjects description	Number of subjects
Group A	Healthy Subjects	30
Group B	Mild difficulty falling asleep - no medication	18
Group C	Mild difficulty falling asleep - temazepam intake	18
Group D	Group A + Group B + Group C	66

and R&K scoring 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 night's sleep. Moreover, the new rules suggest sampling frequency of 500Hz while most datasets (including the one used in this study) include signals sampled at 100Hz. Furthermore, 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. However, most of the studies ignore the above inconsistencies, following AASM standard 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).

3.2. Feature Extraction

The fact that raw PSG signals are non-stationary and their statistics change over time is taken into account during the feature extraction process. Time-domain analysis is not sufficient, so most studies also use frequency, time-frequency and non-linear features [36]. For the current study, the features were extracted at epochs of 30 seconds and they are briefly presented below.

Time domain features are extracted in order to capture information regarding how the signal changes over time. The first and second moment statistics, namely mean value (eq. 1) and variance (eq. 2) are measuring the epoch's average value and the spread of the data points from this value. Moving into higher order statistics, skewness (eq. 3) defines the extent to which a distribution differs from a normal distribution. The skewness of a normal distribution is zero, while positive and negative skewness indicates that data are skewed right and left respectively. Kurtosis (eq. 4) is a statistical measure that describes the degree to which the data points are concentrated around the peak or the tails compared to the normal distribution. All of the statistical features described above are extracted both for the EEG and EOG channels.

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

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

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

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

Entropy (eq. 5) is a measure of randomness describing the lack of order or predictability. High entropy denotes a stochastic process that does not form a specific pattern.

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

In terms of spectral content, EEG waveforms are characterised by components belonging to five different frequency bands (fig. 2). The EEG signal energy in each of these frequency bands is calculated via FT. Signal energy has been successfully used as a feature to many machine learning problems related to EEG analysis, from classification of sleep stages [18] to epilepsy detection, human emotion recognition [37] and cognitive performance [38].

Power spectral Density is also a frequency domain feature describing 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 FT.

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 (PFD) is a feature extracted from PyEEG library [39] which is an open-source python module for EEG/MEG feature extraction.

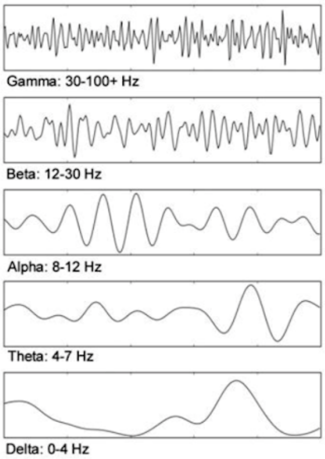
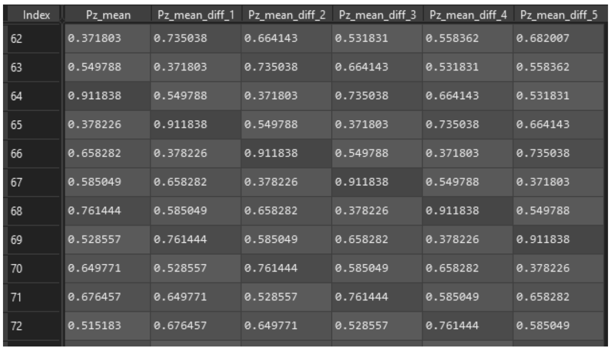


Figure 2. Comparison of EEG bands.



Index	Pz_mean	Pz_mean_diff_1	Pz_mean_diff_2	Pz_mean_diff_3	Pz_mean_diff_4	Pz_mean_diff_5
62	0.371803	0.735038	0.664143	0.531831	0.558362	0.682007
63	0.549788	0.371803	0.735038	0.664143	0.531831	0.558362
64	0.911838	0.549788	0.371803	0.735038	0.664143	0.531831
65	0.378226	0.911838	0.549788	0.371803	0.735038	0.664143
66	0.658282	0.378226	0.911838	0.549788	0.371803	0.735038
67	0.585049	0.658282	0.378226	0.911838	0.549788	0.371803
68	0.761444	0.585049	0.658282	0.378226	0.911838	0.549788
69	0.528557	0.761444	0.585049	0.658282	0.378226	0.911838
70	0.649771	0.528557	0.761444	0.585049	0.658282	0.378226
71	0.676457	0.649771	0.528557	0.761444	0.585049	0.658282
72	0.515183	0.676457	0.649771	0.528557	0.761444	0.585049

Figure 3. Time-delay handcrafted features

Lag features of PSG recordings. PSG recordings are essentially time-series data, so extraction of temporal information is expected to boost our models' accuracy. This concept has been successfully implemented in the literature, mostly at deep learning models that used LSTM layers [40, 41]. The integration of that information at a static model is feasible with the generation of time-delay features from the original ones. Those, so called, lagged features, are feature vectors containing data from previous time steps. The batches of lag features are finally concatenated with the original features, eventually shaping the dataset used for training (fig. 3).

Feature selection. In Fig. 4 the feature importance of the 252 extracted features is presented, utilizing Extreme Gradient Boosting (XGBoost's) built-in method. Feature importance represents how much each feature contributes to decreasing node impurity, weighted by the probability of reaching that node. The more that a feature is used for decision making in a tree, the higher its relative importance. The final value for each feature is calculated by averaging importances across all the decision trees within the model. The sum of all the features importances is equal to 1. Multiple configurations were tried, utilizing subsets of the total features as inputs. More specifically, starting from the ten most significant features, a new model was created every time by iteratively adding the next most significant features, at groups of ten. The highest accuracy was achieved when all of the features were fed into the model. However, for computational complexity reasons a trade-off could have been made between the accuracy and the number of features, as adding the least significant features resulted to a marginal improvement of the proposed model.

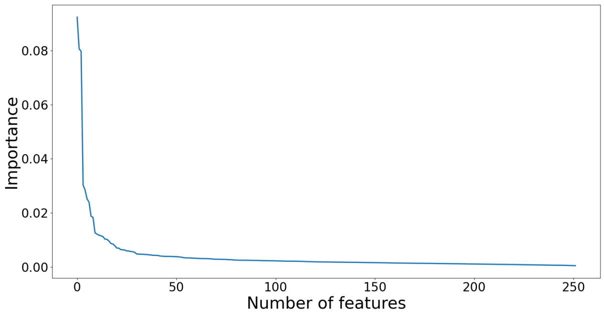


Figure 4. Feature importances of the models' extracted features

3.3. Classification Algorithms

Two classification approaches with different characteristics are compared on the results. The first approach is a static model based on decision trees and the weak learners boosting technique, called XGBoost. High speed and performance make XGBoost stand out among other ensemble methods. Since this is a static model, temporal information is incorporated into the model, using the lag features method described above. The second approach is based on NNs and more specifically Recurrent NNs. LSTM network is a dynamic model able to process input sequences of variable length. It is a model widely used with time-series data since it is able to learn from important events that occurred on some past time steps.

XGBoost. XGBoost is an optimized implementation of Gradient Boosted Trees (GBT) designed to be highly efficient, flexible and portable. GBT is a specific type of gradient boosting model, a technique usually used for regression and classification problems, producing a prediction model in the form of an ensemble of weak prediction models (called *weak learners*). Weak learners are trained sequentially, each one correcting the errors made by its predecessor. In the case of GBT, the weak learners are decision trees. GBT aims to minimise an objective function that combines a convex loss function and a penalty term for model complexity.

The training process proceeds iteratively, adding new trees that predict the residuals of errors of prior trees that are then combined with previous trees to make the final prediction. The simplified form of the objective function for the new tree f_t is [42]:

$$\sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t), \tag{6}$$

where g_i and h_i are first and second order gradient statistics on the loss function. They are defined as follows:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), \tag{7}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}). \tag{8}$$

The second term of the objective function $\Omega(f_t)$, represents a regularisation term in charge of seeking the appropriate final weights to avoid over-fitting.

For the implementation presented in the current study, XGBoost parameters (estimators, learning rate, max depth) were optimized, utilizing grid search. Following the same logic, the optimal number of lag features that was fed into the model, was found to be 5 time steps for each feature.

LSTM. Classification of sequential data is a problem commonly tackled using recurrent neural networks. The idea behind RNNs is that given a sequence of states an RNN will find patterns and optimize itself. Most RNNs suffer from the vanishing and explosive gradient problem. Long Short-Term Memory network, is a more complex variation of a typical recurrent neural network and overcomes the vanishing gradient but not the exploding gradient problem. LSTMs use “forget gate” units in order to decide whether previous information should be kept or forgotten. The most basic hyperparameters that should be taken into consideration to train an LSTM is the batch size of each iteration, the time steps and the number of hidden units of the LSTM itself. A time step is the number of previous inputs that are fed into the network. The number of hidden units refers to the dimensionality of the hidden state and dimensionality of the output state. The network of the current study consists of an LSTM layer followed by a fully connected layer, followed by a softmax activation layer. In order to decide on the hyperparameters 900 different configurations were tried, resulting into an optimal configuration of 50 hidden units and 3 time steps.

4. Results

4.1. Internal Validation

The first set of experiments is using the internal validation process for testing. 10-fold cross-validation with stratified splits was implemented separately to all 4 groups of subjects. The average results of the 10-fold CV, for every group and every sleep stage are presented in Tables 3 and 4 where XGBoost and LSTM approaches are examined.

It is obvious that XGBoost algorithm outperforms LSTM at all cases, achieving higher classification accuracy across all groups and all of the sleep stages. As expected, better results are obtained on group A which is comprised of healthy subjects. More specifically, classification accuracy reaches 91%, while on groups B and C it drops almost by 10%, reaching at 80% and 82% respectively. An interesting finding is that group D which contains all of the subjects, has an accuracy of 87%, consequently it seems that adding healthy and patient subjects to the training set may help the model generalize better, but that is a claim that will be investigated more thoroughly on the external validation. As commonly referred in the literature, stage N1 has the lowest accuracy among all sleep stages. This is also confirmed from the current study and it also seems that N1 is the sleep stage that LSTM mostly struggles to predict. Comparing groups B and C, it seems that they do not have any significant differences regarding the accuracy scores of each sleep stage, however they both have low scores for the Awake stage compared to the healthy subjects.

4.2. External Validation

The second set of experiments was tested with external validation. This means that the subjects that were used for testing were completely kept out of the training process. Initial groups were split into training set and test set, as seen in Table 2. In group D, a model was trained, using subjects from groups A, B and C. Then, the external validation is done separately on subjects of those three groups (Tables 7, 8).

The accuracy seems to drop dramatically at all cases, comparing to the internal validation. This happens because during internal validation one part of each subject’s sleep was used at the training phase, so using the rest of the same subject’s sleep for testing results in higher accuracy. XGBoost surpasses LSTM again in this case, while the problem with N1 stage’s accuracy becomes even worse. The accuracy of group A falls at 82%, while groups B and C drop at 57% and 55% respectively. A significant improvement to the results was observed when a single model was trained with subjects that belonged to group D. The predictive accuracy improved to each group separately (A, B and C), comparing to the case that each group was trained and tested only with subjects that belonged to this specific group. That confirms the initial claim that models are able to generalize better when they are trained jointly with healthy and

Table 2. Train/Test sets for external validation

Group	Train subjects	Test subjects
Group A	20	10
Group B	10	8
Group C	10	8
Group D	40	26

Table 3. XGBoost internal validation (P: precision, R: recall, F1: f1-score).

	Healthy			**Placebo**			**Temazepam**			**All**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Sleep Stage N1	**0.73**	**0.57**	**0.64**	0.66	0.47	0.55	0.65	0.36	0.46	0.72	0.52	0.61
Sleep Stage N2	**0.91**	**0.94**	**0.93**	0.79	0.91	0.85	0.82	0.93	0.87	0.86	0.94	0.90
Sleep Stage N3	**0.92**	**0.91**	**0.91**	0.85	0.77	0.81	0.84	0.79	0.82	0.90	0.83	0.87
REM	**0.91**	**0.91**	**0.91**	0.83	0.75	0.79	0.88	0.83	0.86	0.90	0.88	0.89
Awake	**0.93**	**0.94**	**0.94**	0.75	0.69	0.72	0.74	0.62	0.67	0.89	0.85	0.87
Accuracy	**0.91**			0.80			0.82			0.87		
Average F1-score	**0.87**			0.74			0.74			0.83		

Table 4. LSTM internal validation

	Healthy			**Placebo**			**Temazepam**			**All**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Sleep Stage N1	0.45	0.52	0.48	0.36	0.38	0.37	0.31	0.32	0.31	0.37	0.38	0.38
Sleep Stage N2	0.87	0.87	0.87	0.73	0.70	0.72	0.66	0.70	0.68	0.77	0.78	0.77
Sleep Stage N3	0.87	0.84	0.86	0.61	0.66	0.64	0.56	0.53	0.55	0.72	0.68	0.70
REM	0.79	0.74	0.77	0.61	0.60	0.61	0.47	0.43	0.45	0.68	0.64	0.66
Awake	0.85	0.89	0.87	0.46	0.47	0.47	0.50	0.45	0.47	0.70	0.76	0.73
Accuracy	0.82			0.63			0.57			0.70		
Average F1-score	0.77			0.56			0.49			0.65		

patient subjects. Comparing to the state-of-the-art, the proposed approach of the current study for healthy subjects ranks second among six studies (ranging between 71% and 87%) that implemented external validation utilizing the Sleep-EDFx dataset [43]. The same study presents results for patient subjects which cannot be compared with this study, since the subjects were suffering from a different sleep-related disease, however the reported accuracy ranges between 49% and 69%.

5. Discussion

The current study attempts to propose a reliable sleep stage classification algorithm, trying to contribute towards the replacement of manual sleep scoring with automated solutions. The contribution of frequency domain features extracted from non-stationary PSG signals in combination with the temporal information incorporated from time delay features, results into a robust model, achieving high accuracy, with the appropriate configuration of a tree boosting algorithm. XGBoost surprisingly outperforms LSTM at all cases. This could probably happen because the feature extraction process was not based on the concept of long and short range correlations. A different LSTM setup, like separating each epoch at more narrow time windows, or utilizing LSTM as an auto-encoder from raw data could probably improve its poor accuracy.

Most studies are based only on healthy subjects, nonetheless the results obtained highlight that adding patient subjects may improve the model's generalization capability. The need for efficient sleep scoring is anyway bigger for subjects that suffer from sleep-related problems. Moreover, internal validation may lead to somehow biased results, that can make someone overestimate a model's capabilities. That is why the authors suggest that research on sleep staging should be targeted on studies that focus on the external validation of patient subjects. State-of-the-art results suggest that there is a large margin of improvement at this

Table 5. XGBoost external validation

	Healthy			Placebo			Temazepam		
	P	R	F1	P	R	F1	P	R	F1
Sleep Stage N1	0.61	0.42	0.49	0.31	0.40	0.35	0.21	0.17	0.19
Sleep Stage N2	0.82	0.93	0.87	0.72	0.67	0.69	0.68	0.73	0.70
Sleep Stage N3	0.85	0.90	0.87	0.48	0.58	0.53	0.46	0.44	0.45
REM	0.88	0.64	0.74	0.46	0.45	0.45	0.38	0.29	0.33
Awake	0.88	0.93	0.90	0.54	0.46	0.50	0.39	0.52	0.45
Accuracy	0.82			0.57			0.55		
Average F1-score	0.77			0.50			0.424		

Table 6. LSTM external validation

	Healthy			Placebo			Temazepam		
	P	R	F1	P	R	F1	P	R	F1
Sleep Stage N1	0.31	0.27	0.29	0.36	0.38	0.37	0.25	0.17	0.21
Sleep Stage N2	0.80	0.81	0.81	0.73	0.70	0.72	0.63	0.72	0.67
Sleep Stage N3	0.78	0.88	0.83	0.61	0.66	0.64	0.42	0.54	0.47
REM	0.63	0.59	0.61	0.61	0.60	0.61	0.41	0.30	0.34
Awake	0.78	0.79	0.79	0.46	0.47	0.47	0.41	0.19	0.26
Accuracy	0.74			0.63			0.53		
Average F1-score	0.666			0.562			0.39		

Table 7. XGBoost external validation, group D training

	Healthy			Placebo			Temazepam		
	P	R	F1	P	R	F1	P	R	F1
Sleep Stage N1	0.62	0.42	0.50	0.37	0.37	0.37	0.22	0.19	0.20
Sleep Stage N2	0.83	0.93	0.88	0.73	0.68	0.70	0.69	0.74	0.72
Sleep Stage N3	0.86	0.87	0.97	0.51	0.69	0.59	0.45	0.45	0.45
REM	0.88	0.69	0.77	0.49	0.49	0.49	0.40	0.30	0.34
Awake	0.87	0.94	0.90	0.60	0.53	0.56	0.41	0.52	0.46
Accuracy	0.83			0.60			0.56		
Average F1-score	0.80			0.54			0.43		

Table 8. LSTM external validation, group D training

	Healthy			Placebo			Temazepam		
	P	R	F1	P	R	F1	P	R	F1
Sleep Stage N1	0.39	0.37	0.38	0.30	0.37	0.33	0.20	0.24	0.22
Sleep Stage N2	0.86	0.79	0.83	0.76	0.65	0.70	0.70	0.58	0.64
Sleep Stage N3	0.68	0.91	0.78	0.59	0.70	0.64	0.45	0.55	0.49
REM	0.70	0.74	0.72	0.51	0.35	0.42	0.24	0.22	0.23
Awake	0.83	0.76	0.79	0.34	0.71	0.46	0.32	0.61	0.42
Accuracy	0.75			0.57			0.48		
Average F1-score	0.70			0.51			0.40		

field. A first step towards this goal is the utilization of more data, attempting diversity in terms of sleep health condition. This could lead to an automated scoring algorithm that could be easily implemented on subjects of different age and health condition, substituting human scorers. Using more data would especially benefit ANNs, since the limited number of samples in the N1 stage of sleep is reflected in the limited ability of the network to identify it correctly. Another way to avoid the problem of the limited N1 stage samples is to use class weights thus forcing indirectly the network to focus more on the underrepresented class. Including data from additional inputs besides the PSG recording (EEG, EOG, EMG, HR) could also be a promising approach. In addition to actigraphy and heart rate, based on consumer electronics, Electrodermal Activity (EDA) sensors, which measure the changes in skin conductance resulting from the sympathetic nervous system activity, have been proposed for sleep monitoring [44–46]. Studies utilizing EDA data are mainly targeted on sleep/wake discrimination and sleep quality characterization mainly applicable in environments out of sleep lab. In combination with PSG recordings it is assumed that this extra piece of information on the autonomic function could slightly improve sleep staging predictive accuracy, while at the same time increasing the problem's complexity due to higher dimensionality of the feature space. However, EDA information is not included in the dataset used in the current study. Even though there are numerous 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 real-life 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.

There are also a number of issues that were identified, that make comparing and benchmarking different solutions a demanding task. 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. Finally, as shown in the related work section and in accordance with the current study's results, the models' evaluation method plays an important role, as internal evaluation yields far more accurate results.

6. Conclusion

In this work, a sleep stage classification study for healthy and patient subjects is analyzed. The proposed approach utilizes a mixture time-domain and frequency domain features extracted from 2 EEG and the EOG signals. Two different classification approaches are presented, the first based on tree boosting and the second on LSTM NNs. Sleep staging results are presented for two different evaluation methods and the differences between those methods are highlighted. The suggested tree boosting model achieves results that rank among the state-of-the art models found in the literature. It is also deduced that predictive accuracy is improved when healthy and patient subjects are used jointly at the training process. One possible limitation of the current study is that it does not seem totally applicable for implementation on a real-time wearable sleep scoring machine as it would be rather complex due to the fact that it utilizes 3 different channels as inputs.

References

[1] Siegel, J.M. (2005) Clues to the functions of mammalian sleep. *Nature* **437**(7063): 1264–1271.

[2] Arzi, A., Shedlesky, L., Ben-Shaul, M., Nasser, K., Oksenberg, A., Hairston, I.S. and Sobel, N. (2012) Humans can learn new information during sleep. *Nature neuroscience* **15**(10): 1460.

[3] Stepnowsky, C., Levendowski, D., Popovic, D., Ayappa, I. and Rapoport, D.M. (2013) Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters. *Sleep medicine* **14**(11): 1199–1207.

[4] Wang, Y., Loparo, K.A., Kelly, M.R. and Kaplan, R.F. (2015) Evaluation of an automated single-channel sleep staging algorithm. *Nature and science of sleep* **7**: 101.

[5] Kales, Anthony Rechtschaffen, A. (1968) *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).

[6] Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C., Vaughn, B. *et al.* (2012) The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* **176**.

[7] Tian, P., Hu, J., Qi, J., Ye, X., Che, D., Ding, Y. and Peng, Y. (2017) A hierarchical classification method for automatic sleep scoring using multiscale entropy features and proportion information of sleep architecture. *Biocybernetics and Biomedical Engineering* **37**(2): 263–271.

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

[9] Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.L., Favaro, P., Roth, C., Bargiotas, P. *et al.* (2019) Automated sleep scoring: A review of the latest approaches. *Sleep medicine reviews* .

[10] Timplalexis, C., Diamantaras, K. and Chouvarda, I. (2019) Classification of sleep stages for healthy subjects and patients with minor sleep disorders. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (IEEE): 344–351.

[11] Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A.C. and Oberye, J.J.L. (2000) Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering* **47**(9): 1185–1194.

[12] Terzano, M.G., Parrino, L., Sherieri, A., Chervin, R., Chokroverty, S., Guilleminault, C., Hirshkowitz, M. *et al.* (2001) Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep. *Sleep medicine* **2**(6): 537–553.

[13] Khalighi, S., Sousa, T., Santos, J.M. and Nunes, U. (2016) Isruc-sleep: a comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine* **124**: 180–192.

[14] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E. *et al.* (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* **101**(23): e215–e220.

[15] Fonseca, P., Den Teuling, N., Long, X. and Aarts, R.M. (2018) A comparison of probabilistic classifiers for sleep stage classification. *Physiological measurement* **39**(5): 055001.

[16] Alickovic, E. and Subasi, A. (2018) Ensemble svm method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement* **67**(6): 1258–1265.

[17] Geethanjali, P., Mohan, Y.K. and Sen, J. (2012) Time domain feature extraction and classification of eeg data for brain computer interface. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*: 1136–1139.

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

[19] Akin, M. (2002) Comparison of wavelet transform and fft methods in the analysis of eeg signals. *Journal of medical systems* **26**(3): 241–247.

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

[21] Gopan, K.G., Sinha, N. and Babu, J.D. (2015) Eeg signal classification in non-linear framework with filtered training data. In *2015 23rd European Signal Processing Conference (EUSIPCO)*: 624–628.

[22] Lajnef, T., Chaibi, S., Ruby, P., Aguera, P.E., Eichenlaub, J.B., Samet, M., Kachouri, A. *et al.* (2015) Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of neuroscience methods* **250**: 94–105.

[23] Zhu, G., Li, Y. and Wen, P.P. (2014) Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal. *IEEE journal of biomedical and health informatics* **18**(6): 1813–1821.

[24] Imtiaz, S.A. and Rodriguez-Villegas, E. (2015) Automatic sleep staging using state machine-controlled decision trees. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE): 378–381.

[25] Memar, P. and Faradji, F. (2017) A novel multi-class eeg-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **26**(1): 84–95.

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

[27] Doroshenkov, L., Konyshev, V. and Selishchev, S. (2007) Classification of human sleep stages based on eeg processing using hidden markov models. *Biomedical Engineering* **41**(1): 25–28.

[28] Zhang, J. and Wu, Y. (2018) Automatic sleep stage classification of single-channel eeg by using complex-valued convolutional neural network. *Biomedical Engineering/Biomedizinische Technik* **63**(2): 177–190.

[29] Liu, N., Lu, Z., Xu, B. and Liao, Q. (2017) 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): 1–6.

[30] Yulita, I.N., Rosadi, R., Purwani, S. and Suryani, M. (2018) Multi-layer perceptron for sleep stage classification. *Journal of Physics: Conference Series* **1028**: 012212. doi:10.1088/1742-6596/1028/1/012212.

[31] Rahman, M.M., Bhuiyan, M.I.H. and Hassan, A.R. (2018) Sleep stage classification using single-channel eog. *Computers in biology and medicine* **102**: 211–220.

[32] Hsu, Y.L., Yang, Y.T., Wang, J.S. and Hsu, C.Y. (2013) Automatic sleep stage recurrent neural classifier using energy features of eeg signals. *Neurocomputing* **104**: 105–114.

[33] Michielli, N., Acharya, U.R. and Molinari, F. (2019) Cascaded lstm recurrent neural network for automated sleep stage classification using single-channel eeg signals. *Computers in biology and medicine* **106**: 71–81.

[34] Dijk, D., Beersma, D., Daan, S. and Van den Hoofdakker, R. (1989) Effects of seganserin, a 5-ht2 antagonist, and temazepam on human sleepsstages and eeg power spectra. *European journal of pharmacology* **171**(2-3): 207–218.

[35] Janssen, F., Beecher, L., Griep, P. and Declerck, A. (1989) Short-term effects of temazepam in the eegs of healthy volunteers. *Neuropsychobiology* **22**(2): 72–76.

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

[37] Davidson, R.J. (2003) Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology* **40**(5): 655–665.

[38] Klimesch, W. (1999) Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews* **29**(2-3): 169–195.

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

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

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

[42] Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*: 785–794.

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

[44] Sano, A. and Picard, R.W. (2011) Toward a taxonomy of autonomic sleep patterns with electrodermal activity. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE): 777–780.

[45] Sano, A., Picard, R.W. and Stickgold, R. (2014) Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology* **94**(3): 382–389.

[46] Romine, W., Banerjee, T. and Goodman, G. (2019) Toward sensor-based sleep monitoring with electrodermal activity measures. *Sensors* **19**(6): 1417.