Automatic Sleep Stage Detection: A Study on the Influence of Various PSG Input Signals

Alexandra-Maria Tăuțan[1,2,*], Alessandro C. Rossi[2], Ruben de Francisco[2], and Bogdan Ionescu[1]

Abstract— **Automatic sleep stage detection can be performed using a variety of input signals from a polysomnographic (PSG) recording. In this study, we investigate the effect of different input signals on the performance of feature-based automatic sleep stage classification algorithms with both a Random Forest (RF) and Multilayer Perceptron (MLP) classifier. Combinations of the EEG (electroencephalographic) signal and ECG (electrocardiographic), EMG (electromyographic) and respiratory signals as input are investigated as input with respect to using single channel and multi-channel EEG as input. The Physionet "You Snooze, You Win" dataset is used for the study. The RF classifier consistently outperforms our MLP implementation in all cases and is positively affected by specific signal combinations. The overall classification performance using a single channel EEG is high (an accuracy, precision and recall of 86.91 %, 89.52%, 86.91% respectively) using RF. The results are comparable to the performance obtained using six EEG channels as input. Adding respiratory signals to the inputs processed by RF increases the N2 stage detection performance with 20%, while adding the EMG signal improves the accuracy of the REM stage detection with 5%. Our analysis shows that adding specific signals as input to RF improves the accuracy of specific sleep stages and increases the overall performance. Using a combination of EEG and respiratory signals we achieved an accuracy of 93% for the RF classifier.**

I. Introduction

Automatic sleep stage detection algorithms are helpful in speeding up the process of analyzing PSG (polysomnographic) data. When a sleep disorder is suspected, a PSG study involving the acquisition of several physiological signals related to sleeping patterns is prescribed to the patient. These can include EEG (electroencephalographic) channels, electromyograms (EMG), electrocardiograms (ECG), photoplethysmograms (PPG), signals for tracking respiratory effort through belts placed on the chest and abdomen, nassal cannulas, etc. These biomedical signals are commonly used by medical professionals to label sleep stages and detect anomalies during sleep.

Some combinations of signals are more frequently used than others when detecting specific sleep stages or conditions. The American Academy of Sleep Medicine (AASM) guideline defines 5 sleep stages: Wake, N1 and N2 (light sleep), N3 (deep sleep) and REM (Rapid Eye Movement) [1]. When labeling REM sleep, the EEG activity presented during this stage can easily be misclassified with wakefulness. EMG can be used to clearly differentiate between the two states. In conjunction with EEG, it can also help diagnose motor related sleep disorders such as periodic limb movements or REM sleep behavior disorder [2], [3].

As the sleep cycle progresses from light sleep (N1, N2) into deep sleep (N3), several physiological changes occur: the heart rate (HR) decreases, the heart rate variability (HRV) becomes more stable and respiration becomes slower and more regular. Non-REM sleep stages can be distinguished using cardiac and respiratory signals. Disorders such as sleep apnea can be more clearly identified using these types of signals rather than using only using EEG signals [4].

Typically, clinical professionals use all signals available to label sleep stages and provide a better diagnostic. To perform a PSG diagnosis, targeted signals can be acquired to reduce the discomfort and the time required for analysis. An overview of the signals and signal subsets used for sleep analysis is available in [5]. Automatic sleep staging can also be performed using a subset of the PSG recordings. Sleep stage labeling can be performed using multiple EEG channels [6], single EEG channels [7] or cardiac signals [8]. In [6] all six EEG channels are used with several classification methods resulting in an accuracy of 85.76%. In [7] a single channel EEG is fed into a convolutional neural network resulting in a mean accuracy of 82%. In [8] cardiac signals are used with a linear discriminant analysis for classifying all sleep stages with an overall accuracy of 69%. An overview of the recently developed methods is available in [9].

In this context, this study proposes the following contributions: (i) identify the benefits of performing automatic sleep staging using different signal subsets from PSG recordings, (ii) compare the performance and the final available information with single channel EEG sleep stage detection, (iii) study the effect of data set class balancing. An ideal classification algorithm should provide a similar performance across all sleep stages. If less signals can offer an equivalent performance to using multiple input signals, ambulatory monitoring could be a good alternative to regular in-clinic PSG measurements. We aim to study this possibility with the RF and MLP classifiers.

The paper is organized as follows. Section II provides a description of the methods used for the comparison: an overview of the automatic detection algorithms, feature extraction, data pre-processing methods and the classification algorithms used. Section III presents and discusses the results. In section IV, we provide conclusive remarks and highlight possible directions for future development.

* This work was supported by Onera Health

[1] Alexandra-Maria Tăuțan and Bogdan Ionescu are with Research Center CAMPUS University Politehnica of Bucharest, Research Center CAMPUS, 6 Iuliu Maniu Bd., 061344, Bucharest, Romania alexandra.tautan@upb.ro

[2] Alexandra-Maria Tăuțan, Alessandro C. Rossi and Ruben de Francisco are with Onera Health, 5617AB Eindhoven, The Netherlands

II. Methods

We investigated two classification methods for automatic sleep stage label generation with different inputs: Random Forests and Multilayer Perceptron. Prior to classification, the same processing steps are applied. Features are extracted from the raw signals. These are fed to the classification algorithms selected (see section II-C). The obtained labels are compared to the ground truth using the validation procedure. We consider six raw signals and raw signal combinations as input: (i) EEG - single frontal EEG; (ii) EEGs - six EEG channels (for reference purposes); (iii) ECG - ECG signal; (iv) EEG+ECG - single frontal EEG and ECG signal; (v) EEG+EMG - single frontal EEG and EMG signal; (vi) EEG+Resp - single frontal EEG and three respiratory signals (from the chest,abdomen and airflow).

A. Feature Extraction

Different features are extracted from each type of raw signal taking into account their specific characteristics and the information content required for sleep staging. All features are computed on 30 second epochs. The time window is selected according to the AASM guideline [1]. If combinations of signals are used, the features are pooled together as input for classification:

EEG - a total of 28 features are extracted from each EEG channel. These include time and frequency domain features such as mean and maximum amplitudes, kurtosis, skew and standard deviation of the signals, statistics of the power spectral densities and ratios between specific EEG power bands (delta/theta, theta/alpha, delta/alpha). More details on the extracted features are presented in our previous work [10]. When all six EEG channels are used as input, features extracted from each EEG channel are considered.

ECG - many features have been proposed in literature to characterize sleep using cardiac signals in the context of automatic sleep staging [8]. The cardiac signal is most interesting for obtaining HR and HRV features [8]. As a subject goes from the wake state into light sleep and into deep sleep, HR values drop significantly while the HRV values decrease as well. We selected some of the most commonly used features extracted from HR and HRV. These are detailed in Table I. After detecting the R peaks of the ECG signal [11], [12], the proposed features are computed.

EMG - muscle activity as recorded by EMG can be used to better distinguish sleep stages that in some cases might have similar characteristics. Less activity is present on the EMG when the subject is asleep as compared to awake. Distinguishing between wake and REM stages based solely on EEG signals can be difficult, EMG is often used to differentiate the two. During REM sleep, muscles present atonia (lack of movement) and therefore the EMG is more orderly compared to wakefulness. Movement on the EMG can be detected based on features reflecting amplitude and frequency changes [3], [2]. The proposed features are elaborated in Table I.

Respiration - similar to the cardiac signals, respiratory signals (respiratory efforts measured from the chest and abdomen and airflow) change during the different sleep stages. While going deeper into sleep, the breathing pattern becomes more regular and decreases its rate. This in turn can be described with frequency and time domain features. Typical respiration based features are extracted from all of the three respiratory signals [4]. These are detailed in Table I.

TABLE I: Features extracted from ECG, EMG and Respiratory signals.

Signal	Features	Description
ECG	RR interval	Mean interval between detected R peaks of ECG signal
	BPM	Beats per minute
	TF	Mean power spectrum of HRV <0.4Hz
	VLF	Mean power spectrum of HRV <0.04Hz (very low frequencies)
	LF	Mean power spectrum of HRV between 0.04 and 0.15Hz (low frequencies)
	HF	Mean power spectrum of HRV between 0.15Hz and 0.4Hz (high frequencies)
	LFHF	Ratio between LF and HF
	RMSSD	Root mean square of the HRV signal
	SDNN	Standard Deviation of the HRV signal
	min, max, skew, kurtosis	Statistical measures of the HRV signal
	entropy	Spectral entropy of the HRV signal
EMG	mean, max, min, skew, kurtosis, variance	Statistical measures of the EMG signal
	RMS	root mean square of the EMG
	Entropy	spectral entropy of the EMG signal
	Max Freq	Frequency at which the power spectrum is maximum
	max, mean PSD	maximum and mean values of the PSD
Resp	mean, skew, kurtosis, variance, standard deviation	Statistical measures for respiration based signals
	Max Freq	Frequency at which the power spectrum is maximum
	max, mean PSD	maximum and mean values of the PSD
	NPeaks	Number of peaks detect
	mean, stand dev, skew of distance	Statistical measures of the distance between the detected peaks

B. Data Set Balancing

Most PSG data sets available for training and testing are not balanced. When a data set is unbalanced, more instances belong to one class as compared to the others. This might impact the classification performance. PSG data is naturally unbalanced as a normal sleep pattern contains more non-REM sleep than REM sleep, more light sleep than deep sleep [1].

Balancing the PSG data set for training might be beneficial as the classifier would be presented with an equal number of instances from each class. To overcome this issue, we chose to select the smallest class and randomly sample instances from the other classes such that in the balanced data set all classes have the same size. More information on the content of the data set used in this study is available in section III-A.

C. *Classification*

Many methods have been proposed in literature for the automatic classification of sleep stages [9]. Given our focus on the effect of the input signals, we have selected two common representatives: a tree based approach representative of shallow learning and a neural classifier representative of deep learning.

Random Forests (RF) - an ensemble learning method that provides an output as a combination of several decision trees [13]. In our experiments, a total number of 10 decision trees were used.

Multilayer Perceptron (MLP) - a feed forward artificial neural network that has a minimum number of three layers: input, output and hidden layers. The input layer contains the features while the output layer will contain the information for the predicted sleep stages. By increasing the number of hidden layers a deep neural network is created [14]. The hidden layers contain several perceptrons with a tanh activation function. Experiments were performed with 1, 3, 5 and 7 hidden layers and changing the number of perceptrons per hidden layer. Experiments were performed with 500 and 1,000 perceptron per layer.

III. Experimental Results

A. *Data Set and Metrics*

The data set used in this study is the *"You Snooze, You Win: PhysioNet/Computing in Cardiology Challenge from 2018"* (MGH dataset) available on Physionet [15], [16]. The training set is AASM annotated and it was used for the development of this study as it provides a great wealth of annotated data (994 subjects from the training set) [10].

The MGH dataset contains PSG recordings that include 6 EEG channels (F3M2, F4M1, C3M2, C4M1, O1M2, O2M1), a submentalis EMG, ECG, SaO2 and signals monitoring respiratory effort from the chest, abdomen and the airflow signal. All subjects were included in the analysis. In this work, we considered the frontal F3M2 EEG signal, the EMG, the ECG and the three respiratory signals (chest and abdomen effort signals and airflow).

The proposed methods were tested using a 10-fold cross validation experiment. The performance was assessed by means of accuracy, precision and recall as defined in the following:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \qquad (1),$$

$$Recall = TP/(TP+FN) \qquad (2),$$

$$Precision = TP/(TP+FP) \qquad (3).$$

where TP - true positive, TN - true negative, FP - false positive, FN - false negative

The normalized confusion matrix is also used for specific sleep stage performance assessment. The evaluation protocol is the one used in [10].

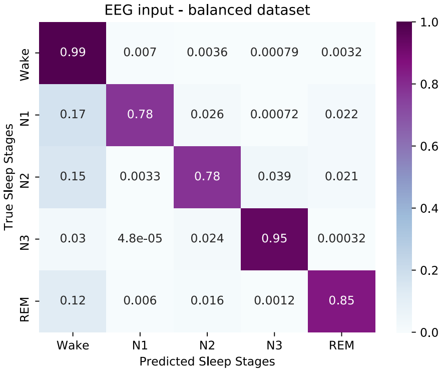


Fig. 1: Normalized confusion matrix for RF classification using the class balanced data set of features extracted from the frontal EEG channel.

B. *Effects of Data Balancing*

The use of data set class balancing was tested on all signal combinations and with both classifiers. An improvement in both overall performance and individual class prediction performance was observed in all cases. As an example, when using the RF classifier with the features extracted only from the frontal EEG in the unbalanced case, the accuracy, precision and recall were of 72.96%, 79.69% and 71.82%, respectively. By balancing the dataset, the overall performance increased to 86.91%, 89.52%, and 86.91%, respectively which is an improvement of 13.95, 9.83 and 19.09 percentage points. For the MLP classifier using the same input, an accuracy of 75.5% for the unbalanced data set is obtained, while for the balance data set it increases to 82.55%.

When using an unbalanced data set with a single channel frontal EEG as input, sleep stages N1, N2 and REM have a lower prediction performance [10]. When using a balanced data set, the three sleep stages have an improved prediction performance of about 10 percentage points. Figure 1 shows the improved performance in the corresponding normalized confusion matrix. The balanced data set was used in our experiments.

C. *Classification Algorithms*

The best RF performance was obtained with a minimum number of samples per leaf of 10 [10]. For the MLP classifier, the best performance was obtained when using 3 hidden layers with 500 units per layer. Table II presents the results obtained for all input signal combinations using the classifiers with the optimized parameters.

The RF classifier systematically outperforms the MLP for all signal combinations used as input. The performance obtained with a single channel EEG is comparable to that obtained when using 6 EEG channels. The wake versus REM sleep stage discrimination is improved when adding the EMG signal as input for both classifiers. This is reflected in the overall performance when using the RF classifier, but it is not valid for the MLP classifier. When using a single channel EEG and respiratory signals, the performance for the RF classification is higher than for the other combinations.

This does not hold for the MLP classifier, when the reported performance for this input combination is the lowest. The MLP classification might be further improved by slightly changing the network architecture. For signal combinations that increase the number of input features, the performance decreases. This might be a sign of overfitting.

D. Effect of Different Input Signals

Table II provides a comparison of the overall performance for each signal combination. The classification results obtained from the single channel frontal EEG are comparable to the results obtained on all 6 EEG channels from our study (see Table II) and from literature [6] where an accuracy of 86.91 % was reported. The RF classification performance using a single channel EEG with feature extraction and data set balancing reaches an accuracy, precision and recall of approximately 86%, 89% and 87%, respectively. Adding both the EMG and respiration signals improves the performance (up to approximately 6 percentage points increase). The highest accuracy of 93% is obtained when adding the the respiration signals as input to the RF classifier. When using the MLP classifier, the highest performance is obtained using the EEG as input. The lowest performance was obtained when using only the ECG signal as input. This might be due to the combination of ECG features chosen and does not imply that the ECG might not be useful should another feature set e considered. When combining the ECG and EEG signals, a small improvement is obtained.

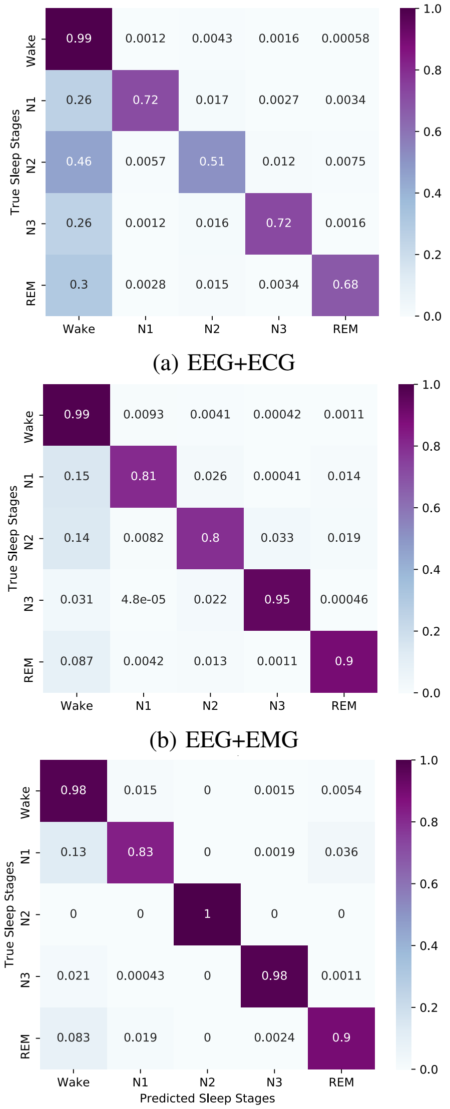
The sleep stage prediction performance obtained when combining a single channel EEG with other PSG signals is presented in Figure 2 as normalized confusion matrices. The presented results are from the RF classifier. As expected, when adding the EMG signal the REM sleep stage classification is slightly improved (see Fig. 2b). The addition of the respiratory signals increases the prediction abilities for the N2 sleep stage by approximately 20 percentage points (see Fig. 2c). It also improves the discrimination of REM sleep which implies respiration signals might also be a viable alternative to using the EMG as input. The features obtained from the ECG signal do not improve the classification in our analysis, but have a rather negative impact on performance (see Fig. 2a). Although the overall performance of the automatic sleep scoring algorithm using only one channel EEG data is sufficiently high, adding EMG and respiratory signals provides more information for specific sleep stages such as N2 and REM.

IV. Conclusions

In this study we investigated the effects of using different signal combinations as input to feature based automated sleep stage detection algorithms. A Random Forest and Multilayer Perceptron network were used as classifiers for automated sleep stage detection using as input a single channel EEG, 6 EEG channels and combinations of a single channel EEG and ECG, EMG and chest, abdomen and airflow respiratory signals. The best accuracy of 93% was obtained when adding respiratory signals to EEG with the Random Forest classifier.

TABLE II: Comparison of performance for Automated Sleep Stage Scoring using different input signals and the two classifiers. *A - Accuracy, P - Precision, R - Recall*

Input Signals	**Random Forests**			**Multilayer Perceptron**		
	A[%]	**P[%]**	**R[%]**	**A[%]**	**P[%]**	**R[%]**
EEG	86.89	88.80	86.91	82.55	82.56	82.54
EEGs	86.65	89.31	86.68	73.93	73.67	73.93
ECG	72.45	85.56	72.45	59.23	60.25	59.23
EEG+ECG	72.52	85.72	72.50	60.28	55.00	60.27
EEG+EMG	88.65	90.62	88.63	66.70	66.84	66.69
EEG+Resp	93.72	94.34	93.71	52.27	53.5	52.26



(a) EEG+ECG

(b) EEG+EMG

(c) EEG+Resp

Fig. 2: Normalized confusion matrices for the RF classifier, using a single channel frontal EEG combined with ECG, EMG and respiratory signals, respectively. The input data set was balanced.

REM sleep stage detection was improved when adding the chin EMG signal. N2 stage prediction was improved when using the respiratory signals. The accuracy obtained using a single channel EEG and six EEG channels was in the same range of approximately 86%. When considering the used performance metrics, the same information can be obtained from a single channel frontal EEG as from the full six channel EEG montage recommended for PSG recordings.

Data set class balancing for training purposes improved the per class performance. The RF classifier outperformed the implemented version of the MLP. Using different classification models might improve the general performance and the per class predictions specifically when using single channel EEG as input. For future work, the MLP classifier network can be improved by selecting a different architecture. Adding dropout layers might prevent overfitting. Making use of temporal information can capture the natural progression through sleep stages and might also positively impact the performance. The RF classifier might perform better also due to its inherent feature relevance selection. Implementing a method that incorporates better feature selection might also improve class predictions.

References

[1] R. B. Berry;, R. Brooks;, C. Gamaldo;, S. Harding;, R. Lloyd;, S. F. Quan;, M. Troester;, and B. V. Vaughnh, “The AASM Manual for the Scoring of Sleep and Associated Events. Version 2.4.” *American Academy of Sleep Medicine*, 2017.

[2] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. Hu, and M. De Vos, “Detection of REM sleep behaviour disorder by automated polysomnography analysis,” *Clinical Neurophysiology*, vol. 130, no. 4, pp. 505–514, 2019. [Online]. Available: https://doi.org/10.1016/j.clinph.2019.01.011

[3] J. Kempfner, H. B. D. Sorensen, M. Nikolic, and P. Jennum, “Early Automatic Detection of Parkinson ’ s Disease Based on Sleep Recordings,” *American Clinical Neurophisiology Society*, vol. 31, no. 5, pp. 409–415, 2014.

[4] T. Van Steenkiste, W. Groenendaal, J. Ruyssinck, P. Dreesen, S. Klerkx, C. Smeets, R. De Francisco, D. Deschrijver, and T. Dhaene, “Systematic Comparison of Respiratory Signals for the Automated Detection of Sleep Apnea,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, 2018, pp. 449–452.

[5] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. D. Clifford, “A review of signals used in sleep analysis,” *Physiological Measurement*, vol. 35, no. 1, 2014.

[6] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. B. Westover, M. T. Bianchi, and J. Sun, “SLEEPNET: Automated Sleep Staging System via Deep Learning,” *arXiv*, pp. 1–17, 2017. [Online]. Available: http://arxiv.org/abs/1707.08262

[7] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, “Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks,” *arXiv*, 2016. [Online]. Available: https://arxiv.org/pdf/1610.01683.pdf

[8] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, “Sleep stage classification with ECG and respiratory effort,” *Physiological Measurement*, vol. 36, no. 10, pp. 2027–2040, 2015.

[9] L. Fiorillo, A. Puiatti, M. Papandrea, P.-l. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, and F. D. Faraci, “Automated sleep scoring : A review of the latest approaches,” *Sleep Medicine Reviews*, vol. 48, p. 101204, 2019. [Online]. Available: https://doi.org/10.1016/j.smrv.2019.07.007

[10] A.-M. Tautan, A. C. Rossi, R. De Franciso, and B. Ionescu, “Automatic Sleep Stage Detection using a Single Channel Frontal EEG,” in *The 7th IEEE International Conference on E-Health and Bioengineering - EHB 2019*, 2019.

[11] H. Sedghamiz, “Matlab Implementation of Pan Tompkins ECG QRS detector,” Tech. Rep., 2014.

[12] J. Pan and W. J. Tompkins, “Pan Tomkins 1985 - QRS detection.pdf,” *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.

[13] F. Schroff, A. Criminisi, and A. Zisserman, “Object class segmentation using random forests,” *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*, 2008.

[14] S. O. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Ontario Canada, 1999.

[15] M. M. Ghassemi, B. E. Moody, L. W. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford, “You Snooze, You Win: The PhysioNet/Computing in Cardiology Challenge 2018,” *Computing in Cardiology*, vol. 2018-Septe, pp. 20–23, 2018.

[16] A. L. Goldberger, L. A. 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, Journal of the American Heart Association*, vol. 101, no. 23, 2000.