SUPERVISED LEARNING FOR SLEEP STAGE SCORING USING RANDOM FOREST: IS A "SIMPLER" MODEL ACCURATE ENOUGH ON UNSEEN INDIVIDUALS?

Eugenia Moris[♭], Cecilia Forcato[†] and Ignacio Larrabide[♭]

[♭] *PLADEMA, UNICEN-CONICET, Tandil, Argentina, emoris@plademe.exa.unicen.edu.ar, www.yatiris.github.io*
[†] *Laboratorio de Sueño y Memoria, Depto. de Ciencias de la Vida, Instituto Tecnológico de Buenos Aires , (ITBA)*

Abstract: Sleep scoring is a common method used by experts to monitor the quantity and quality of sleep in people, but it is a time-consuming and labour-intense task. As an alternative, automatic sleep scoring has been recently studied. In this work we used one multi-class Random Forest model and tree cascade Random Forest models to classify sleep in stages. The proposed models use discrete Wavelets to extract features. This models showed a general Fscore of $66,94\%$ (Multi-class), 71.58% (Model 1), 71.53% (Model 2) and 61.92% (Model 3). These results did not improve the general Fscore previously described in the literature, but it did for specific stages and models. With this, Random Forest, in combination with discrete Wavelets, provides a cost effective high performance classification tool. Model design and training is crucial to achieve a proper performance in subjects unseen by the model during training.

Keywords: *Sleep stages, Sleep scoring, Random Forest, Wavelets, Automatic classification.*
2000 AMS Subject Classification: 68T10

1 INTRODUCTION

Sleep scoring is a common method used by experts to monitor quality and quantity of sleep in people [2], but this work becomes time-consuming and labour-intense task, it is prone to errors due to fatigue and the inter scorer agreement among expert is around 83% [1]. Due to this, automatic sleep scoring has been recently studied using machine learning techniques.

This problem has been approached in different ways. Sors et al. [3] used CNN for sleep scoring classification. To improve the classification, the authors used 4 epochs as an input: the one being classified, the two previous epochs and the one after. They had good results in almost all classes except for the Stage 1 with a precision of 55% and a recall of 35%. Supratak et al. [4] used a parallel CNN with Bidirectional long short term memory (Bi-LSTM). In the data-set Sleep-EDF they got a Fscore of $82.0\% - 76.9\%$. The biggest limitation was observed during the classification of Stage 1, with a Fscore of 46.6%. Despite the power of CNNs, the use of Wavelets combined with statistics has proven to be a powerful tool to address this problem, as shown by the literature in the past few years. Hassan et al. used a tunable Q-factor wavelet transform and spectral features for sleep scoring, obtaining an accuracy of 90.1%, and like the previous works had low sensitivity of $38,74\%$ in detecting Stage 1 [5]. Silveira et al. [6] use a set of statistical features in the Wavelet domain with an accuracy of 91.5%. All the methods presented above include 5 classes, namely: Stage 1, Stage 2, Stage Slow Wave Sleep (SWS), REM and Awake. In this work, we use Wavelets-derived features, in combination with Random Forest, to go one step further in supervised sleep scoring. Further, only Stages 1, 2 and SWS were considered, since this is work under development.

2 MATERIALS AND METHODS

2.1 DATASET

The public dataset Sleep-EDF was used, which has record pings of two nights for 20 healthy subjects (mean age $28.7(\pm 2.9)$ years) [7]. Each record contains 2 EEG channels (Fpz-Cz and Pz-Cz) with a sampling frequency of 100 Hz. Each record is pre-classified by an expert in EEG sleep scoring following the R&K criteria [8].

All experiments were ran on Stage 1, Stage 2 and SWS, where SWS results from merging Stages 3 and 4 according to R&K criteria. REM and Awake stages were not considered so far, and therefore the epochs with this classification were removed.

A total of 2804 30 seconds epochs in Stage 1, 17799 epochs in Stage 2 and 5703 epochs in Stage SWS were obtained, with an average of 140 epochs in Stage 1, 990 epochs in Stage 2 and 285 epochs in Stage SWS for each patient.

2.2 WAVELETS & RANDOM FOREST

Wavelets are a powerful tool to extract features from 1D signals like EEG. Unlike Fourier transform, the wavelet has the ability to characterize time information in addition to frequency information. To process the data, Pywavelet library (V1.1.1) were used [9]. Wavelet transform was applied to each epoch and the derived coefficients were considered as features for each one. Resulting features were standardised ($\mu = 0$ and $\sigma = 1$).

We used the Random Forest implementation provided in Scikit-learn (V0.22.1), with a random seed ('1234'), ensuring that the same sequence of random numbers was generated in subsequent runs of the algorithm [11][10].

2.3 MODEL TRAINING AND META-PARAMETERS

For training and validation 14 subjects were randomly chosen and the remaining 6 subjects were used only at test time for model performance purposes, to assess it as accurately as possible. During training, a leave-one-out cross-entropy was used.

The number of epochs of Stage 2 largely exceed the number of epochs in the other two classes. Because of that, Random Under Sampling (RUS) was used to balance the number of epochs between classes. A total of 375 features were obtained after using Wavelet transform on the data. Having an average of 8832 epochs, we opted for dimensional reduction before further analysis. For this, Principal Component Analysis (PCA) was used. The input of the random forest was the standardised matrix of features, reduced by PCA, with the corresponding label for each epoch.

2.4 EXPERIMENTS AND VALIDATION

We first performed experiments with different Wavelet families to find the one that best captured the features that are relevant for our problem. Then, data needed to be balanced so that quality measures were not misguided by this natural imbalance in the number of samples for each epoch within a single subject. We trained different models while increasing RUS of Stage 2 by steps of 5% in each iteration. The value that resulted in the optimal classification performance was used for each model.

Next, we extracted features at different scales from the signal, by recursively using different levels of Wavelets. This allows capturing features that cannot be observed by a Wavelet in a single level. Afterward, a dimensional reduction was done using PCA and we defined the number of decision trees in the Random Forest.

To find the optimal meta-parameters for the models, 13 subjects were used for training leaving one subject out for validation. This process of excluding one subject was repeated for each subject. The final meta-parameters value was chosen using Fscore as quality measure, which is considered to produce the best classification results, in general [12].

2.5 MODEL CONSTRUCTION

General classification was computed as the mean of the performance measure over the full leave-one-out round. Four different Random Forest models were trained. The first model classifies all inputs into three different classes (Stage 1, Stage 2, and Stage SWS) in one shot.

For a second type of model, two Random Forest models were used in cascade. First, Stage 1 and 2 were merged, and the model was trained to distinguish between Stage 1/2 and Stage SWS. A second "cascaded" model was trained to distinguish between Stage 1 and Stage 2. Each model was trained separately with its own fine tuned meta-parameters. This process was repeated for all combinations (Stage 1 vs 2/SWS, and Stage 2 vs 1/SWS) and therefore, in addition to the simple multi-class model, we obtained three additional models. Model 1 is the one described before. Model 2 merges Stage 2 and SWS in the first pass. Finally,

	Acc	Fscore	Acc	Fscore	Acc	Fscore	Acc	Fscore	Fscore	Fscore
General	79,29%	66,94%	82,44%	*71,58%*	81,83%	71,53%	83,15%	61,92%	**76,9%**	76,23%
Stage 1	92,17%	38,41%	91,44%	46,23%	90,47%	*46,32%*	92,79%	19,35%	**46,6%**	44%
Stage 2	79,79%	82,63%	82,75%	85,66%	82,71%	85,94%	83,33%	***87,76%***	85,9%	85%
Stage SWS	86,62%	79,07%	90,70%	*82,84%*	90,47%	82,33%	90,17%	78,64%	84,8%	**86%**
	Multi-class		Model 1		Model 2		Model 3		Supratak et al. [4]	Koushik et al. [15]

Table 1: Average Fscore across test subjects by sleep stage and model during model test. In *italics* are highlighted the best results from our models, and in **bold** are highlighted the best results overall (proposed and literature).

Model 3 merges Stage 1 and SWS in the first pass. The merged Stages are classified in a second pass with a cascaded model. As with the Multi-class model, we use general Fscore to choose optimal meta-parameters. However, we use Stage 1 Fscore to choose the percentage of RUS used to classify Stage 1 and Stage 2 in Model 1.

3 Results

Different Wavelets families were tested. As opposed to previous works [13, 6, 14] that simply use Daubechies Wavelet, Discrete Meyer wavelet produced the best performance for Stage 1 and in general. In the multi-class model, Stage 2 samples were reduced by 30%, resulting in an average of 1979 epochs for Stage 1, 3439 epochs for Stage 2 and 3416 epochs for Stage SWS. Also, we used the approximation coefficient of the Level 3 Wavelet, resulting in 375 features, which were reduced to 250 with PCA. In the case of the cascade models the meta-parameter were different in each case. The decision was taken looking at the general Fscore.

Table 1 shows the quantitative result in the classification by class and model. Model 1 and 2 had similar results out of the classification, showing the higher Fscore of all classes. These results were only improved by Stage 2 classification of Model 3. But Model 3 showed inferior results in general.

The work presented by Supratak et al. [4] showed better results for Stage 1 classification. Meanwhile the work of Koushik et al. [15] was the best in the classification of Stage SWS. Finally the Model 3, presented here, had the best performance for Stage 2. Also, in Table 1, we can appreciate that the Fscore obtained by Supratak et al. are similar with the Fscore obtained in the Model 2.

4 Discussion and conclusions

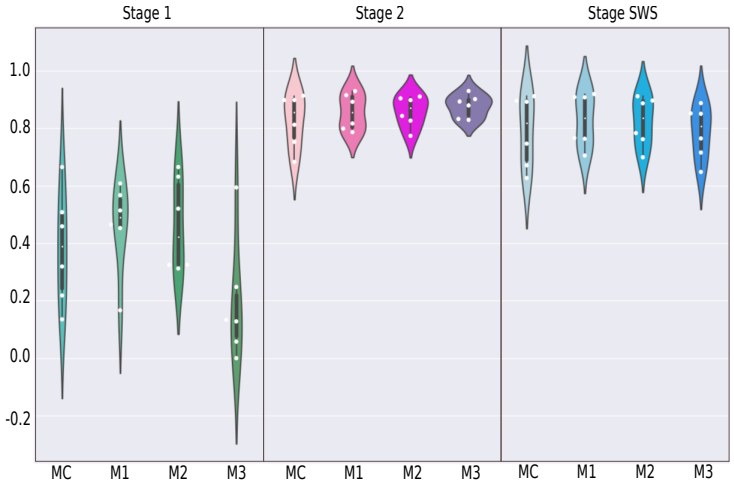

Figure 1: Classification performance of each model and for each sleep stage. The white dots represent subjects. The green violin shows Fscore for Stage 1. Violins in shades of red show results for Stage 2. Finally, violins in blue correspond to Stage SWS. We have the multi-class model (MC), the cascade model 1 (M1), the cascade model 2 (M2) and the cascade model 3 (M3)

Four different models, capable of classifying the Sleep Stages 1, 2 and SWS using Wavelet transform for features extraction, were developed (Figure 1). We compared the performance between a classic multi-class Random Forest classifier and a cascade model.

Wavelet transform has proven to be good for feature extraction when classifying sleep stages, mostly because they can extract time and frequency features, all at once. The characteristic of the oscillatory activity in each sleep stage are different, both in frequency and amplitude, and different Wavelet families perform differently depending on each case.

It is important to notice that with a simpler, low-memory requirements model produced similar results than more complex models, like CNN and LSTM, as observed in Table 1. The use of Random Forest was advantageous since for training a good model without the need of specific or high performance computing hardware, ultimately leads to lower costs. We also need to keep in mind that most models presented in the literature classify between 5 classes, adding Awake and REM classes. This is a possible reason for an improvement in classification performance and is currently under study. Stage 1 showed the worst performance, as seen previously in the literature [4, 3, 5, 6, 15]. Meanwhile Stage 2 showed good result, with a Fscore above 87%, followed by the Stage SWS that for some cases exceeded 82%. The low performance of Model 3 on Stage 1 samples might be because one of the two Random Forests, presented in the model, had trouble learning to identify the majority class (Stage 2).

The Table 1 shows a high accuracy for each class, mostly in Stage 1. This happens because of the Accuracy Paradox, where having a large numbers of true negatives, epochs that do not correspond to Stage 1, in contrast with the epochs that correspond to Stage 1. This proves the importance of looking at Fscore instead of just at the Accuracy.

ACKNOWLEDGMENTS

This work was partially funded by PICT 2016-0116 and by an NVIDIA hardware grant. First author is funded by a CONICET PhD Scholarship

REFERENCES

[1] R. S. ROSENBERG, AND S. VAN HOUT, *The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring*, Journal of clinical sleep medicine, 2013.

[2] S. MCKINLEY, AND OTHERS, *Sleep and psychological health during early recovery from critical illness: an observational study*, Journal of psychosomatic research (2013).

[3] A. SORS AND OTHERS , *A convolutional neural network for sleep stage scoring from raw single-channel EEG* Biomedical Signal Processing and Control (2018).

[4] A. SUPRATAK, H. DONG, C. WU AND Y. GUO, *DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG*, IEEE Transactions on Neural Systems and Rehabilitation Engineering(2017).

[5] A. R. Hassan and M. I. H. Bhuiyan,A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features, Journal of neuroscience methods (2016)

[6] T. LT. da Silveira, A. J. Kozakevicius and C. R. Rodrigues,Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain, Medical & biological engineering & computing (2017)

[7] A.L. Goldberger and others ,PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, circulation (2000)

[8] E. A. Wolpert,A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects., Archives of General Psychiatry (1969)

[9] G. R. Lee and others ,PyWavelets: A Python package for wavelet analysis, Journal of Open Source Software (2019).

[10] L. Breiman,Random forests, Machine learning (2001)

[11] F. Pedregosa and others ,Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research (2011)

[12] M, Sokolova, and G, Lapalme,A systematic analysis of performance measures for classification tasks, Information processing & management (2009)

[13] O. Ali, O. Aydoğan, and D. Tuncel,Automatic sleep stage classification using artificial neural network with wavelet transform, Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi

[14] M. Sharma and others,Automated detection of sleep stages using energy-localized orthogonal wavelet filter banks, Arabian Journal for Science and Engineering (2020)

[15] A. Koushik, J. Amores, and P. Maes,Real-time Smartphone-based Sleep Staging using 1-Channel EEG, 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (2019).